Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayessches Lernen

Christoph Sawade/Niels Landwehr/Paul Prasse
Silvia Makowski
Tobias Scheffer

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayessche Lineare Regression, Naive Bayes

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayessche Lineare Regression, Naive Bayes

Statistik & Maschinelles Lernen

- Maschinelles Lernen: eng verwandt mit (induktiver) Statistik
- Zwei Gebiete in der Statistik:
  - *Deskriptive Statistik*: Beschreibung, Untersuchung von Eigenschaften von Daten.

Mittelwerte

Varianzen

Unterschiede zwischen Populationen

  - *Induktive Statistik*: Welche Schlussfolgerungen über die Realität lassen sich aus Daten ziehen?

Modellbildung

Erklärungen für Beobachtungen

Zusammenhänge, Muster in Daten

Thomas Bayes

- 1702-1761
- „An essay towards solving a problem in the doctrine of chances“, 1764 veröffentlicht.

- Arbeiten von Bayes grundlegend für induktive Statistik.

- „Bayessche Wahrscheinlichkeiten“ wichtige Sichtweise auf Unsicherheit & Wahrscheinlichkeit

Frequentistische / Bayessche Wahrscheinlichkeit

- Frequentistische Wahrscheinlichkeiten
  - Beschreiben die Möglichkeit des Eintretens intrinsisch stochastischer Ereignisse (z.B. Münzwurf).
  - Definition über *relative Häufigkeiten* möglicher Ergebnisse eines *wiederholbaren Versuches*

„Wenn man eine faire Münze 1000 Mal wirft, wird etwa 500 Mal Kopf fallen“

„In 1 Gramm Potassium-40 zerfallen pro Sekunde ca. 260.000 Atomkerne“

Frequentistische / Bayessche Wahrscheinlichkeit

- Bayessche, „subjektive“ Wahrscheinlichkeiten
  - Grund der Unsicherheit ein Mangel an Informationen
    - Wie wahrscheinlich ist es, dass der Verdächtige X das Opfer umgebracht hat?
    - Neue Informationen (z.B. Fingerabdrücke) können diese subjektiven Wahrscheinlichkeiten verändern.

- Bayessche Sichtweise im maschinellen Lernen wichtiger
- Frequentistische Sichtweise auch manchmal verwendet, mathematisch äquivalent

Bayessche Wahrscheinlichkeiten im Maschinellen Lernen

- Modellbildung: Erklärungen für Beobachtungen finden
- Was ist das „wahrscheinlichste“ Modell? Abwägen zwischen
  - Vorwissen (Prior über Modelle)
  - Evidenz (Daten, Beobachtungen)
- Bayessche Sichtweise:
  - Evidenz (Daten) verändert „subjektive“ Wahrscheinlichkeiten für Modelle (Erklärungen)
  - A-posteriori Modellwahrscheinlichkeit, MAP Hypothese

Wahrscheinlichkeitstheorie, Zufallsvariablen

- Zufallsexperiment: definierter Prozess, in dem ein Elementarereignis ω erzeugt wird.
- Ereignisraum Ω: Menge aller Elementarereignisse.
- Ereignis A: Teilmenge des Ereignisraums.
- Wahrscheinlichkeitsfunktion p: Funktion, die Ereignissen $A \subseteq \Omega$ Wahrscheinlichkeiten zuweist.

Wahrscheinlichkeitstheorie

- Gültige Wahrscheinlichkeitsfunktion p (Kolmogorow-Axiome)
  - Wahrscheinlichkeit von Ereignis $A \subseteq \Omega$: $0 \leq p(A) \leq 1$
  - Sicheres Ereignis: $p(\Omega) = 1$, und $p(\varnothing) = 0$
  - Für die Wahrscheinlichkeit zweier inkompatibler Ereignisse $A \subseteq \Omega, B \subseteq \Omega$ (d.h. $A \cap B = \varnothing$) gilt:

$$p(A \cup B) = p(A) + p(B)$$

Wahrscheinlichkeitstheorie: Beispiel

- Würfeln
  - Ereignisraum $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Elementarereignisse haben Wsk $p(\{\omega\}) = 1/6$
  - Ereignis gerade Zahl: $A = \{2, 4, 6\}$
  - Wahrscheinlichkeit des Ereignisses: $p(A) = 1/2$

Wahrscheinlichkeitstheorie, Zufallsvariablen

- Zufallsvariable X: Abbildung von Elementarereignissen auf numerische Werte

$$X : \Omega \to \mathbb{R}$$
$$\omega \mapsto x$$

Experiment weisst Zufallsvariable X den Wert $x = X(\omega)$ zu

- Wahrscheinlichkeit dafür, dass Ereignis $X=x$ eintritt (Zufallsvariable X wird mit Wert x belegt).
  - $p(X = x) = p(\{\omega \in \Omega \mid X(\omega) = x\})$

- Zusammenfassen in Wahrscheinlichkeitsverteilung, der Variable X unterliegt

$p(X)$ Verteilung gibt an, wie Wahrscheinlichkeiten über Werte x verteilt sind

$X \sim p(X)$ „X ist verteilt nach $p(X)$“

Zufallsvariable: Beispiel

- Würfeln mit 2 Würfeln
  - Ereignisraum $\Omega = \{(\omega_1, \omega_2) \mid \omega_i \in \{1,2,3,4,5,6\}\}$
  - Elementarereignisse haben Wsk $p(\{(\omega_1, \omega_2)\}) = 1/36$
  - Zufallsvariable: Summe der beide Augenzahlen

$$X((\omega_1, \omega_2)) = \omega_1 + \omega_2$$

  - Wahrscheinlichkeit für Wert der ZV:

$$p(X = 5) = ?$$

Zufallsvariable: Beispiel

- Würfeln mit 2 Würfeln
  - Ereignisraum $\Omega = \{(\omega_1, \omega_2) \mid \omega_i \in \{1,2,3,4,5,6\}\}$
  - Elementarereignisse haben Wsk $p(\{(\omega_1, \omega_2)\}) = 1/36$
  - Zufallsvariable: Summe der beide Würfel

$$X((\omega_1, \omega_2)) = \omega_1 + \omega_2$$

  - Wahrscheinlichkeit für Wert der ZV:

$$p(X = 5) = p(\{(1,4),(2,3),(3,2),(4,1)\})$$

$$= 4/36$$

Diskrete/kontinuierliche Zufallsvariablen

- Diskrete Zufallsvariablen: *D=X(Ω)* diskret
- Kontinuierliche Zufallsvariablen: *D=X(Ω)* kontinuierlich

- Für diskrete Zufallsvariablen gilt:

$$\sum_{x \in D} p(X = x) = 1 \qquad D \text{ diskreter Wertebereich}$$

- Beispiel: *N* Münzwürfe
  - Zufallsvariablen $X_1, ..., X_N \in \{0,1\}$
  - Münzparameter μ gibt Wahrscheinlichkeit für „Kopf“ an

$$p(X_i = 1) = \mu \qquad \text{Wahrscheinlichkeit für „Kopf“}$$

$$p(X_i = 0) = 1 - \mu \qquad \text{Wahrscheinlichkeit für „Zahl“}$$

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i} \qquad \text{Bernoulli-Verteilung}$$

Diskrete Zufallsvariablen

- Beispiel: Anzahl „Köpfe“ bei *N* Münzwürfen
  - ZV „Anzahl Köpfe“: $X = \sum_{i=1}^{N} X_i, \qquad X \in \{0, \ldots, N\}$
  - Binomial-Verteilung

$$X \sim \mathrm{Bin}(X \mid N, \mu)$$

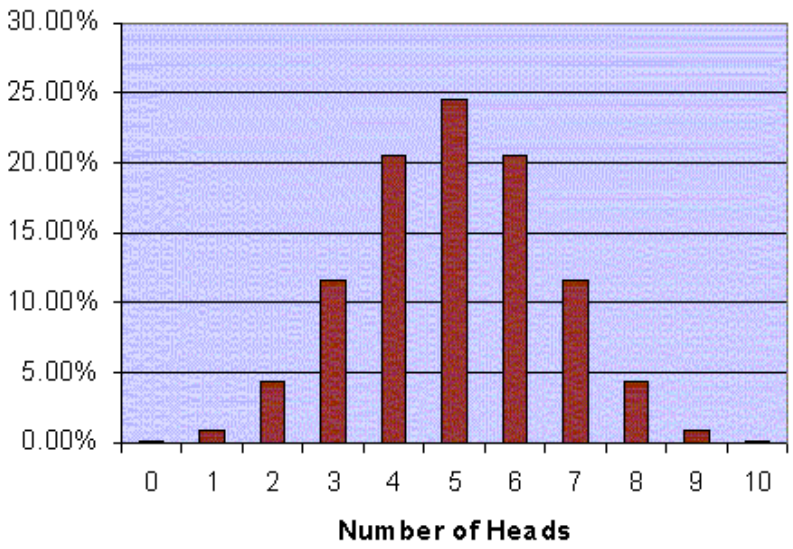
$$\mathrm{Bin}(X \mid N, \mu) = ?$$

Diskrete Zufallsvariablen

- Beispiel: Anzahl „Köpfe“ bei *N* Münzwürfen
  - ZV „Anzahl Köpfe“: $X = \sum_{i=1}^{N} X_i, \qquad X \in \{0, \ldots, N\}$
  - Binomial-Verteilung

$$X \sim \mathrm{Bin}(X \mid N, \mu)$$

$$\mathrm{Bin}(X \mid N, \mu) = \binom{N}{X} \mu^X (1-\mu)^{N-X}$$



$N = 10, \quad \mu = 0.5$

Kontinuierliche Zufallsvariablen

- Kontinuierliche Zufallsvariablen
  - Unendlich (überabzählbar) viele Werte möglich
  - Typischerweise Wahrscheinlichkeit $p(X = x) = 0$

- Statt Wahrscheinlichkeiten für einzelne Werte: Dichtefunktion

$f_X : \mathbb{R} \to \mathbb{R}$ „Dichte“ der ZV X

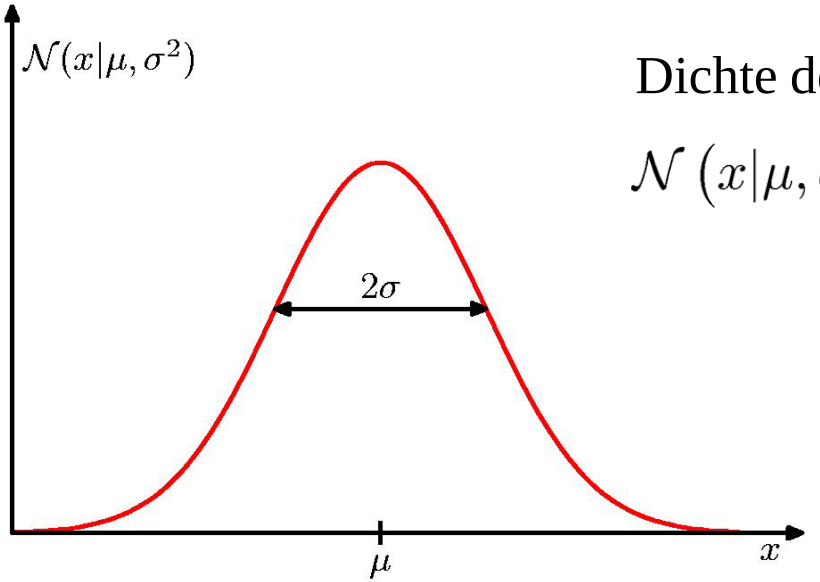
$\forall x : f_X(x) \geq 0, \qquad \int_{-\infty}^{\infty} f_X(x) = 1 \qquad f_X(x) > 1$ möglich

- Wahrscheinlichkeit, dass ZV *X* Wert zwischen *a* und *b* annimmt

$$p(X \in [a,b]) = \int_a^b f_X(x)dx,$$

Kontinuierliche Zufallsvariablen

- Beispiel: Körpergröße X
  - X annähernd Gaußverteilt („Normalverteilt“)
  - $X \sim \mathcal{N}(x \mid \mu, \sigma^2)$



Dichte der Normalverteilung

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

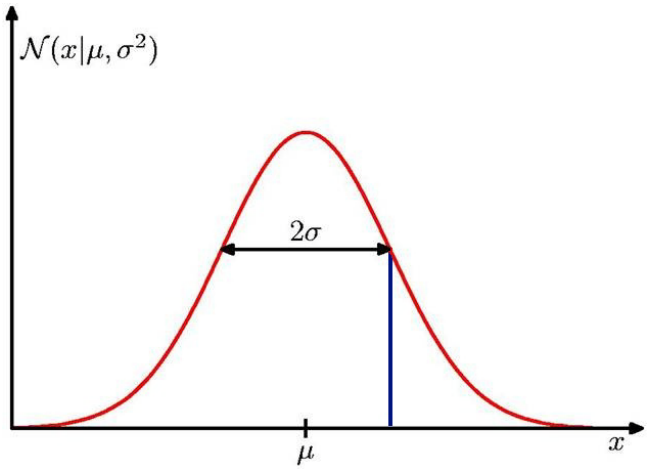
z.B. $\mu = 170, \sigma = 10$

Kontinuierliche Zufallsvariablen

- Beispiel: Körpergröße
  - Wie groß ist die Wahrscheinlichkeit, dass ein Mensch genau 180cm groß ist?

$$p(X = 180) = 0$$

  - Wie groß ist die Wahrscheinlichkeit, dass ein Mensch zwischen 180cm und 181cm groß ist?



$$p(X \in [180,181]) = \int_{180}^{181} \mathcal{N}(x \mid 170, 10^2)\,dx$$

Kontinuierliche Zufallsvariablen

- Verteilungsfunktion

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f_X(z)dz,$$

$$p(X \in [a,b]) = F(b) - F(a)$$

- Dichte ist Ableitung der Verteilungsfunktion

$$f_X(x) = \frac{dF(x)}{dx}$$

- Veranschaulichung Dichte:

$$f_X(x) = \lim_{\varepsilon \to 0} \frac{p(X \in [x-\varepsilon, x+\varepsilon])}{2\varepsilon}$$

Notation

- Notation: wenn der Zusammenhang klar ist, schreiben wir auch manchmal

$p(x)$ statt $p(X = x)$ (diskrete Wahrscheinlichkeit)

$p(x)$ statt $f_X(x)$ (kontinuierliche Dichte)

Konjunktion von Ereignissen

- Wahrscheinlichkeit für Eintreten mehrerer Ereignisse:

$p(X = x, Y = y)$ gemeinsame Wahrscheinlichkeit

$f_{X,Y}(x, y)$ gemeinsame Dichte

- Gemeinsame Verteilung (diskret/kontinuierlich)

$p(X, Y)$

Bedingte Wahrscheinlichkeiten

- Wie beeinflusst zusätzliche Information die Wahrscheinlichkeitsverteilung?
  - $p(X \mid \text{zusätzliche Information})$
- Bedingte Wahrscheinlichkeit eines Ereignisses:
  - $$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \quad \text{diskret}$$
- Bedingte Dichte:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{kontinuierlich}$$

- Bedingte Verteilung (diskret/kontinuierlich):
  - $$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

Bedingte Wahrscheinlichkeiten

- Produktregel

$$p(X,Y) = p(X \mid Y)\,p(Y)$$ diskret/kontinuierlich

- Summenregel

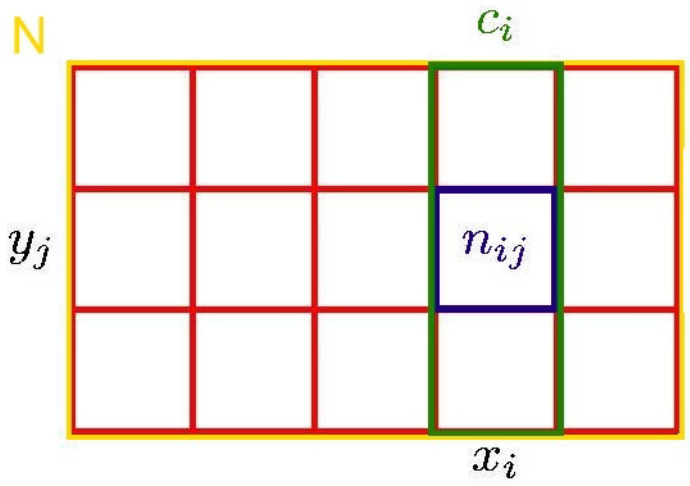
$$p(X = x) = \sum_{y \in D} p(X = x, Y = y)$$ diskret

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$ kontinuierlich

$p(X = x)$ heisst auch "Randwahrscheinlichkeit"

Gemeinsame und bedingte Wahrscheinlichkeiten

„Zufälligen Punkt auf Fläche auswählen“



Gemeinsame Wahrscheinlichkeit

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Randwahrscheinlichkeit

$$p(X = x_i) = \frac{c_i}{N}.$$

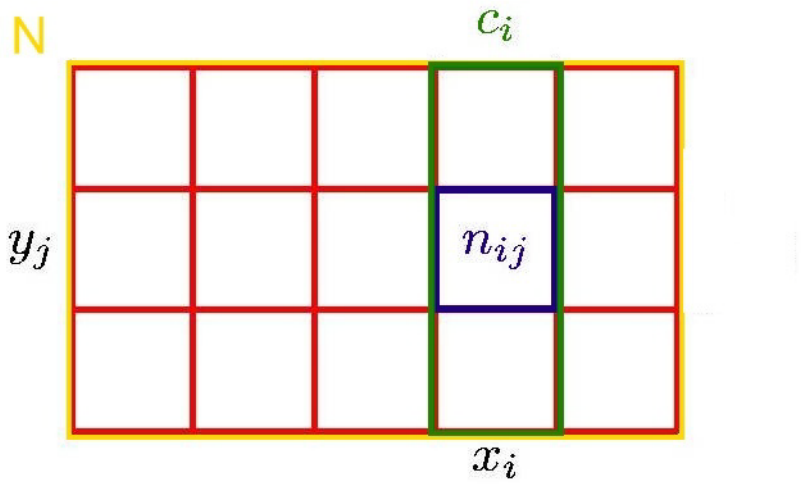
Bedingte Wahrscheinlichkeit

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

$$= \frac{p(Y = y_j, X = x_i)}{p(X = x_i)}$$

Wichtige Rechenregeln für Wahrscheinlichkeiten

„Zufälligen Punkt auf Fläche auswählen“



Summenregel

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_j n_{ij}$$

$$= \sum_j p(X = x_i, Y = y_j)$$

Produktregel

$$p(Y = y_j, X = x_i) = p(Y = y_j \mid X = x_i)\, p(X = x_i)$$

$$p(Y = y_j, X = x_i) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p(Y = y_j \mid X = x_i)\, p(X = x_i)$$

Unabhängigkeit

- Zwei Zufallsvariablen sind unabhängig, wenn:
  - $p(X,Y) = p(X)\,p(Y)$
- Äquivalent dazu
  - $p(X \mid Y) = p(X)$ und $p(Y \mid X) = p(Y)$

- Beispiel: wir würfeln zweimal mit fairem Würfel, bekommen Augenzahlen x_1, x_2
  - ZV X_1, X_2 sind unabhängig
  - ZV $X_+ = X_1 + X_2$ und $X_- = X_1 - X_2$ sind abhängig

Erwartungswert

- Erwartungswert einer Zufallsvariable:

$$E(X) = \sum_x xp(X = x)$$ X diskrete ZV

$$E(X) = \int xp(x)dx$$ X kontinuierliche ZV mit Dichte $p(x)$

- Veranschaulichung: gewichtetes Mittel, Schwerpunkt eines Stabes mit Dichte *p(x)*

- Rechenregeln Erwartungswert

$$E(aX + b) = aE(X) + b$$

$$E(X + Y) = E(X) + E(Y)$$

Erwartungswert

- Erwartungswert additiv

$$E(X+Y) = \sum_{x,y} (x+y)\, p(X=x, Y=y)$$

$$= \sum_{x,y} x p(X=x, Y=y) + \sum_{x,y} y p(X=x, Y=y)$$

$$= \sum_{x} x \sum_{y} p(X=x, Y=y) + \sum_{y} y \sum_{x} p(X=x, Y=y)$$

Summenregel

$$= \sum_{x} x p(X=x) + \sum_{y} y p(Y=y)$$

$$= E(X) + E(Y)$$

Varianz, Standardabweichung

- Varianz:
  - Erwartete quadrierte Abweichung von *X* von *E(X)*
  - Mass für die Stärke der Streuung

$$Var(X) = E((X - E(X))^2) = \sum_x (x - E(X))^2 \, p(X = x)$$

$$Var(X) = E((X - E(X))^2) = \int (x - E(X))^2 \, p(x) dx$$

- Standardabweichung

$$\sigma_X = \sqrt{Var(X)}$$

- Verschiebungssatz

$$Var(X) = E(X^2) - E(X)^2$$

Varianz, Standardabweichung

- Verschiebungssatz

$$Var(X) = E((X - E(X))^2)$$

$$= E(X^2 - 2E(X)X + E(X)^2)$$

$$= E(X^2) - 2E(X)E(X) + E(X)^2$$

$$= E(X^2) - E(X)^2$$

Rechenregeln Varianz

- Rechenregeln Varianz/Standardabweichung

$$Var(aX+b) = a^2 Var(X), \qquad \sigma_{aX+b} = a\sigma_X$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Cov(X,Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

- Covarianz misst „gemeinsame Schwankung“ der Variablen
  - Falls Variablen unabhängig:

$$Cov(X,Y) = 0, \qquad Var(X+Y) = Var(X) + Var(Y)$$

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i}$$

$$E(X_i) = ?$$

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i}$$

$$E(X_i) = \sum_{x \in \{0,1\}} x p(X_i = x)$$

$$= 1\mu + 0(1-\mu) = \mu$$

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i}$$

$$E(X_i) = \sum_{x \in \{0,1\}} x p(X_i = x)$$

$$= 1\mu + 0(1-\mu) = \mu$$

- Erwartungswert Binomialverteilung

$$X \sim \text{Bin}(X \mid N, \mu) \qquad X = \sum_{i=1}^{N} X_i$$

$$E(X) = \sum_{x=0}^{N} x p(X = x)$$

$$= \sum_{x=0}^{N} x \binom{N}{x} \mu^x (1-\mu)^{N-x}$$

$$= ?$$

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i}$$

$$E(X_i) = \sum_{x \in \{0,1\}} x p(X_i = x)$$

$$= 1\mu + 0(1-\mu) = \mu$$

- Erwartungswert Binomialverteilung

$$X \sim \text{Bin}(X \mid N, \mu) \qquad X = \sum_{i=1}^{N} X_i$$

$$E(X) = \sum_{x=0}^{N} x p(X = x)$$

$$= \sum_{x=0}^{N} x \binom{N}{x} \mu^x (1-\mu)^{N-x}$$

$$= N\mu$$

Summe der Erwartungswerte der Bernoulli-Variablen

Erwartungswert, Varianz Binomialverteilung

- Varianz Bernoulliverteilung?

$$X_i \sim \text{Bern}(X_i \mid \mu)$$

$$Var(X_i) = ?$$

Erwartungswert, Varianz Binomialverteilung

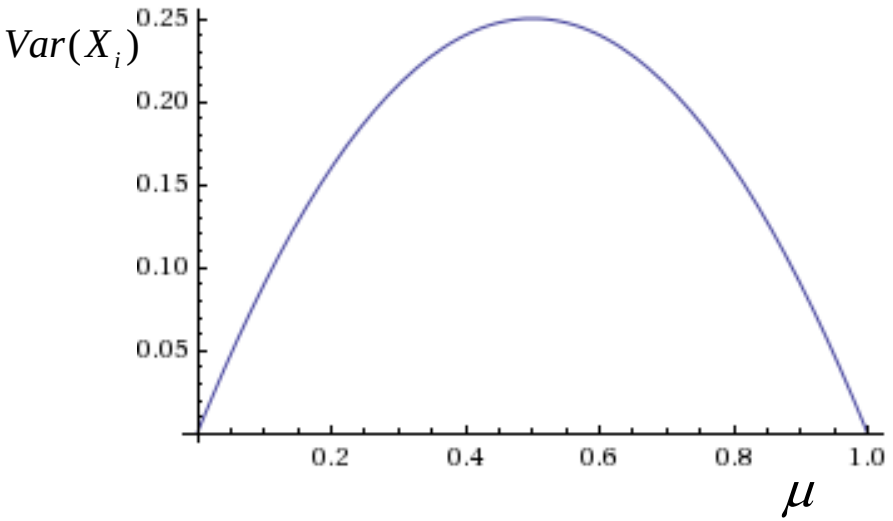
- Varianz Bernoulliverteilung?

$$X_i \sim \text{Bern}(X_i \mid \mu)$$

$$Var(X_i) = ?$$

Verschiebungssatz:

$$Var(X_i) = E(X_i^2) - E(X_i)^2$$
$$= \mu - \mu^2 = \mu(1-\mu)$$



Erwartungswert, Varianz Binomialverteilung

- Varianz Binomialverteilung

$$X \sim \text{Bin}(X \mid N, \mu)$$

$$Var(X) = ?$$

$$X = \sum_{i=1}^{N} X_i$$

$$X_i \sim \text{Bern}(X_i \mid \mu) = \mu^{X_i}(1-\mu)^{1-X_i}$$

$$Var(X_i) = \mu(1-\mu) \Rightarrow Var(X) = N\mu(1-\mu) \qquad X_i \text{ unabhängig}$$

Erwartungswert, Varianz Normalverteilung

- Erwartungswert Normalverteilung

$$X \sim \mathcal{N}(x \mid \mu, \sigma^2)$$

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$E(X) = \int_{-\infty}^{\infty} x\mathcal{N}(x \mid \mu, \sigma^2)dx$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$$= \int_{-\infty}^{\infty} x\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx$$

$z = x - \mu$

$$= \int_{-\infty}^{\infty} (z+\mu)\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz$$

$$= \mu\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz + \int_{-\infty}^{\infty} z\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz = \mu$$

Erwartungswert, Varianz Normalverteilung

- Erwartungswert Normalverteilung

$$X \sim \mathcal{N}(x \mid \mu, \sigma^2)$$

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$E(X) = \int_{-\infty}^{\infty} x\mathcal{N}(x \mid \mu, \sigma^2)dx$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$$= \int_{-\infty}^{\infty} x\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)dx$$

$z = x - \mu$

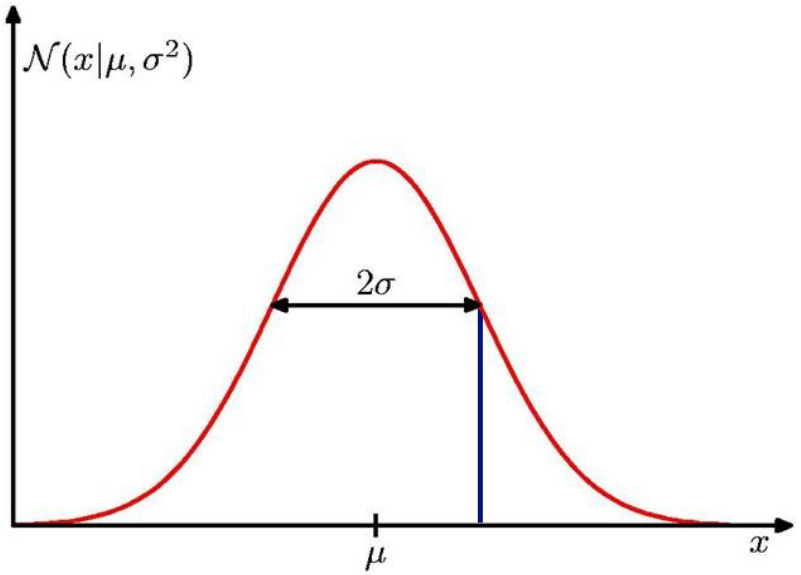
$$= \int_{-\infty}^{\infty} (z+\mu)\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz$$

$$= \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz}_{=1} + \underbrace{\int_{-\infty}^{\infty} z\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left(-\frac{1}{2\sigma^2}z^2\right)dz}_{=0} = \mu$$

Erwartungswert, Varianz Normalverteilung

- Varianz Normalverteilung
  - Man kann zeigen dass

$$X \sim \mathcal{N}(x \mid \mu, \sigma^2) \quad \Rightarrow \quad Var(X) = \sigma^2$$



Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen

Erinnerung: Problemstellung Lernen

- Eingabe Lernproblem: Trainingsdaten.
  - $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$
  - $\mathbf{x}_i \in \mathbb{R}^k$ Merkmalsvektoren

    $y_i \in \mathcal{Y}$ Labels

spam

ok

spam

- Ausgabe: Klassifikator (auch als Modell bezeichnet).

$$f : X \to Y$$

$$f_{\mathbf{w}}(\mathbf{x}) = \begin{cases} \text{spam}: \ \textit{wenn } \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \geq 0 \\ \text{ok}: \ \textit{sonst} \end{cases}$$

Linearer Klassifikator mit Parametervektor $\mathbf{w}$.

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{m} w_i x_i$$

Modellvorstellung Bayes‘sches Lernen

- Viele Verfahren des maschinellen Lernens basieren auf probabilistischen Überlegungen
- Modellvorstellung beim Lernen:
  - Jemand hat echtes Modell f^* nach A-Priori Verteilung („Prior“) $p(f)$ gezogen
  - f^* ist nicht bekannt, aber $p(f)$ reflektiert Vorwissen (was sind wahrscheinliche Modelle?)
  - Trainingseingaben $\mathbf{x}_i$ werden gezogen.
  - Klassenlabels y_i werden nach $p(y_i \mid \mathbf{x}_i, f^*)$ gezogen.

    Intuition: $y_i = f^*(\mathbf{x}_i)$

    Um beispielsweise Datenrauschen abzubilden, $y_i \sim p(y_i \mid \mathbf{x}_i, f^*)$
  - Fragestellung Lernen: Gegeben L und $p(f)$, was ist wahrscheinlichstes „echte“ Modell ?
    → Versuche, f^* (ungefähr) zu rekonstruieren

Bayessche Regel

- Wichtigstes Werkzeug im Bayes‘schen Lernen: Bayes‘sche Regel

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

- Beweis einfach: $p(X \mid Y) = \frac{p(X, Y)}{p(Y)} = \frac{p(Y \mid X)\, p(X)}{p(Y)}$

Definition bedingte Verteilung

Produktregel

- Wichtige Grundeinsicht für das maschinelle Lernen: Erlaubt den Rückschluss auf *Modellwahrscheinlichkeiten* gegeben *Wahrscheinlichkeiten von Beobachtungen*

Bayessche Regel

- Anwendung Bayes‘sche Regel: Modellwahrscheinlichkeit gegeben Daten und Vorwissen

Relativ einfach anzugeben: wie hoch ist die Wahrscheinlichkeit, bestimmte Daten zu sehen, unter der Annahme dass *Modell* das korrekte Modell ist?

$$p(Modell \mid Daten) = \frac{p(Daten \mid Modell)\, p(Modell)}{p(Daten)}$$

A-priori Verteilung über Modelle: Vorwissen

Interresanter Term: wie ist die Wahrscheinlichkeit für Modelle, gegeben Evidenz der Trainingsdaten?

Wahrscheinlichkeit der Daten, unabhängig von Modell

- Erlaubt die Berechnung des *maximum a-posteriori (MAP)* Modells

$$Modell_{MAP} = \arg\max_{Modell} p(Modell \mid Daten)$$

Bayessche Regel

- Wahrscheinlichkeit der Daten uninteressant, weil unabhängig von Modell

$$p(Modell \mid Daten) = \frac{p(Daten \mid Modell)\, p(Modell)}{p(Daten)}$$

$$= \frac{1}{Z}\, p(Daten \mid Modell)\, p(Modell)$$

$$\propto p(Daten \mid Modell)\, p(Modell)$$

Notation „Prop-To“: gleich bis auf multiplikative Konstante

***Likelihood**: wie gut erklärt Modell die Daten?*

***Prior**: wie wahrscheinlich ist Modell a priori?*

Maximum-A-Posteriori-Hypothese

- Wahrscheinlichstes Modell gegeben die Daten

$$f_{MAP} = \arg\max_{f_{\mathbf{w}}} p(f_{\mathbf{w}}|L)$$

$$= \arg\max_{f_{\mathbf{w}}} \frac{p(L|f_{\mathbf{w}})\,p(f_{\mathbf{w}})}{p(L)}$$

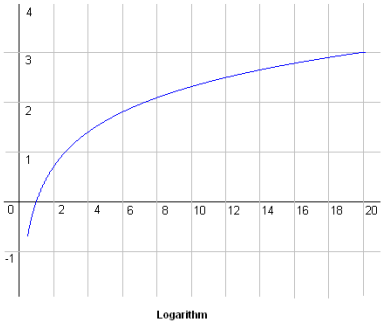
Anwendung Bayes'sche Regel

$$= \arg\max_{f_{\mathbf{w}}} p(L|f_{\mathbf{w}})\,p(f_{\mathbf{w}})$$

- Für nicht-negative reellwertige Funktionen gilt:

$$\arg\max_{z} G(\mathbf{z}) = \arg\max_{\mathbf{z}} \log G(\mathbf{z})$$

Weil Logarithmus monoton:



Maximum-A-Posteriori-Hypothese

- Wahrscheinlichstes Modell gegeben die Daten

$$
\begin{aligned}
f_{MAP} &= \arg\max_{f_{\mathbf{w}}} \; p(L|f_{\mathbf{w}})\, p(f_{\mathbf{w}}) \\
&= \arg\max_{f_{\mathbf{w}}} \; \log(p(L|f_{\mathbf{w}})\, p(f_{\mathbf{w}})) \\
&= \arg\min_{f_{\mathbf{w}}} \; -\log P(L|f_{\mathbf{w}}) - \log p(f_{\mathbf{w}})
\end{aligned}
$$

Log-Likelihood

Log-Prior

$\Rightarrow$ Optimierungskriterium bestehend aus log-likelihood und log-prior

$\Rightarrow$ Erinnerung: Lernen als Optimierungsproblem,
Summe aus Verlustfunktion und Regularisierer

Log-Likelihood

- Wie wahrscheinlich sind die Daten gegeben das Modell?

$$-\log p(L \mid f_{\mathbf{w}}) = -\log p(y_1,...,y_N,\mathbf{x}_1,...,\mathbf{x}_N \mid f_{\mathbf{w}})$$
$$= -\log(p(y_1,...,y_N \mid \mathbf{x}_1,...,\mathbf{x}_N, f_{\mathbf{w}})\, p(\mathbf{x}_1,...,\mathbf{x}_N))$$
$$= -\log p(y_1,...,y_N \mid \mathbf{x}_1,...,\mathbf{x}_N, f_{\mathbf{w}}) - \log p(\mathbf{x}_1,...,\mathbf{x}_N)$$
$$= -\log p(y_1,...,y_N \mid \mathbf{x}_1,...,\mathbf{x}_N, f_{\mathbf{w}}) + const$$

Produktregel

Konstanter Faktor, unabhängig von f

- Annahme: Datenpunkte unabhängig
  - Beispiel: Label einer Email hängt nur von Merkmalsvektor und Modell ab, nicht von anderen Merkmalsvektoren oder Labels

$$p(y_1,...,y_N \mid \mathbf{x}_1,...,\mathbf{x}_N, f_{\mathbf{w}}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, f_{\mathbf{w}})$$

- Einsetzen ergibt

$$-\log p(L \mid f_{\mathbf{w}}) = -\log \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, f_{\mathbf{w}}) + const$$
$$= -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, f_{\mathbf{w}}) + const$$

Log-Likelihood

- Was ist $p(y_i \mid f_{\mathbf{w}}, \mathbf{x}_i)$?
- Definition mit Verlustfunktion, beispielsweise

$$\ell(f(\mathbf{x}_i), y_i) = \begin{cases} 0 : f(\mathbf{x}_i) = y_i \\ c : f(\mathbf{x}_i) \neq y_i \end{cases}$$

- Je höher der Verlust (Differenz zwischen Vorhersage und beobachtetem Label), desto geringer die Wahrscheinlichkeit dieser Beobachtung

$$p(y_i \mid f_{\mathbf{w}}, \mathbf{x}_i) = \frac{1}{Z} \exp\left(-\ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\right)$$

Normalisierer

Annahme: spezielle Exponentialverteilung

$$\log p(y_i \mid f_{\mathbf{w}}, \mathbf{x}_i) = -\ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + const$$

- Einsetzen ergibt

$$-\log p(L \mid f_{\mathbf{w}}) = \sum_i \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + const$$

- **Negative Log-Likelihood entspricht Verlustterm!**

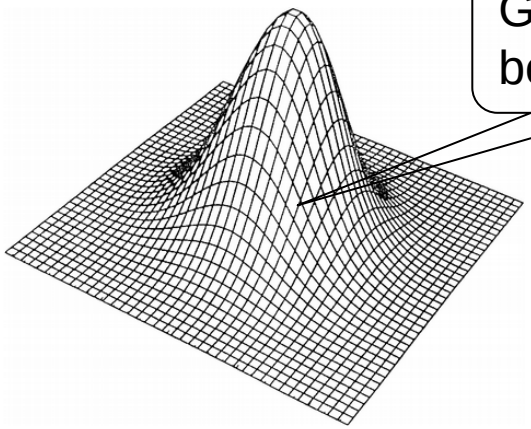
A-Priori-Wahrscheinlichkeit (Prior)

- Was ist a-priori Verteilung $p(f_{\mathbf{w}})$?
- Erinnerung an Diskussion der Regularisierer:
  - Vorwissen: Modelle mit wenig Attributen sind wahrscheinlicher
  - Vorwissen: $\|\mathbf{w}\|^2$ eher klein

- Multivariate Normalverteilung

$$p(f_{\mathbf{w}}) = \mathcal{N}(\mathbf{w} \mid 0, \sigma^2 I)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\|\mathbf{w}\|^2}$$



- Negativer Log-Prior:

$$-\log p(f_{\mathbf{w}}) = \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 + const$$

Konstante Terme, unabhängig von f

- **Negativer Log-Prior = Regularisierer!**

A-Posteriori-Wahrscheinlichkeit (Posterior)

- Wahrscheinlichstes Modell gegeben Vorwissen und Daten.
  - $$f_{MAP} = \arg\max_{f_{\mathbf{w}}} p(f_{\mathbf{w}}|L)$$
    $$= \arg\min_{f_{\mathbf{w}}} -\log p(L|f_{\mathbf{w}}) - \log p(f_{\mathbf{w}})$$
    $$= \arg\min_{f_{\mathbf{w}}} \sum_i l(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \lambda \|\mathbf{w}\|^2$$
    $$\lambda = \frac{1}{2\sigma^2}$$

- ArgMin über regularisierte Verlustfunktion!
- Rechtfertigung für Optimierungskriterium?
  - **Wahrscheinlichste Hypothese (*MAP-Hypothese*).**

Lernen und Vorhersage

- Bisher haben wir mit Hilfe der Bayes‘schen Regel das wahrscheinlichste Modell gegeben die Daten bestimmt:

$$f_{MAP} = \arg\max_{f_w} p(f_{\mathbf{w}}|L)$$

- Löst Lernproblemstellung:
  - Gegeben: Daten L, Vorwissen $p(f)$
  - Gesucht: Modell $f : \mathcal{X} \rightarrow \mathcal{Y}$

- Vorhersagen werden mit Hilfe des gelernten Modells getroffen:

$y = f_{MAP}(\mathbf{x})$ $\mathbf{x}$ neue Testinstanz

- Zweistufiger Prozess:
  - Erst Modell Lernen
  - Dann Vorhersage mit gelerntem Modell

Lernen und Vorhersage

- Wenn wir uns auf ein Modell festlegen müssen, ist MAP Modell sinnvoll
- Aber eigentliches Ziel ist *Vorhersage* einer Klasse!
- Besser, sich nicht auf ein Modell festzulegen, solange noch Unsicherheit über bestes Modell besteht
- Stattdessen *Bayessche Vorhersage*: direkt optimale Vorhersage ausrechnen, ohne sich auf Modell festzulegen

Lernen und Vorhersage: Beispiel

- Modellraum mit 4 Modellen: $H = \{ f_1, f_2, f_3, f_4 \}$
- Binäres Klassifikationsproblem, $\mathcal{Y} = \{0,1\}$
- Trainingdaten *L*
- Wir haben a-posteriori-Wahrscheinlichkeiten berechnet

$p(f_1 \mid L) = 0.3$ $\quad$ $p(f_3 \mid L) = 0.25$

$p(f_2 \mid L) = 0.25$ $\quad$ $p(f_4 \mid L) = 0.2$

- MAP Modell ist $f_1 = \arg\max_{f_i} p(f_i \mid L)$

Lernen und Vorhersage: Beispiel

- Modelle f_i probabilistische Klassifikatoren:
  - Modell liefert Wahrscheinlichkeit für positive Klasse

$$p(y=1 \mid \mathbf{x}, f_i) \in [0,1] \qquad \text{("80\% Sicherheit für Klasse Spam")}$$

  - Vorhersage:

$$f_i(\mathbf{x}) = \begin{cases} 1 : p(y=1 \mid \mathbf{x}, f_i) > 0.5 \\ 0 : sonst \end{cases}$$

Lernen und Vorhersage: Beispiel

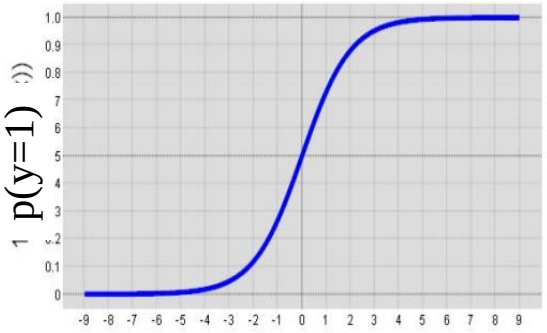
- Beispiel für probabilistischen Klassifikator: Logistische Regression
  - Lineares Modell:

Entscheidungsfunktionswert $\mathbf{w}^T\mathbf{x}$

$$p(y=1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$

$\mathbf{w}$ Parametervektor

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$



Entscheidungsfunktionswert wx

„logistische Regression“

Lernen und Vorhersage: Beispiel

- Wir wollen neues Testbeispiel $\mathbf{x}$ klassifizieren

$$p(y=1|\mathbf{x}, f_1) = 0.6 \qquad p(y=1|\mathbf{x}, f_3) = 0.2$$
$$p(y=1|\mathbf{x}, f_2) = 0.1 \qquad p(y=1|\mathbf{x}, f_4) = 0.3$$

- Klassifikation mit MAP Modell f_1 : $y=1$
- Idee: nicht auf Modell festlegen, solange noch Unsicherheit über Modelle besteht
- Alternativer Ansatz: Bestimme direkt die Wahrscheinlichkeit für die optimale Vorhersage, gegeben die beobachteten Daten

$$p(y=1|\mathbf{x}, L) = ?$$

Beobachtungen:
Trainingsdaten *L*,
Merkmalsvektor der Testinstanz **x**

Lernen und Vorhersage: Beispiel

- Ausrechnen:

$$p(y=1|\mathbf{x},L) = \sum_{i=1}^{4} p(y=1,\ f_i\ |\mathbf{x},L) \qquad \text{Summenregel}$$

$$= \sum_{i=1}^{4} p(y=1|\ f_i,\mathbf{x},L)\, p(\ f_i\ |\mathbf{x},L) \qquad \text{Produktregel}$$

$$= \sum_{i=1}^{4} p(y=1|\mathbf{x},\ f_i)\, p(\ f_i\ |L)$$

$$= 0.6*0.3+0.1*0.25+0.2*0.25+0.3*0.2 = 0.315$$

- Vorhersage $y=0$, ungleich MAP-Modell!

Lernen und Vorhersage: Beispiel

- Wenn Ziel Vorhersage ist, sollten wir $p(y=1|\mathbf{x},L)$ verwenden
  - Nicht auf ein Modell festlegen, solange noch Unsicherheit über Modelle besteht
  - Grundidee der Bayesschen Vorhersage

Bayessches Lernen und Vorhersage

- Problemstellung Bayes‘sche Vorhersage
- Gegeben:
  - Trainingsdaten L,
  - neue Testinstanz $\mathbf{x}$.
- Gesucht:
  - Verteilung über Labels y für gegebenes $\mathbf{x}$: $p(y \mid \mathbf{x}, L)$
  - Bayessche Vorhersage: $y_* = \arg\max_y p(y \mid \mathbf{x}, L)$

- Minimiert Risiko einer falschen Vorhersage.
- Heißt auch Bayes-optimale Entscheidung oder Bayes-Hypothese.

Bayessches Lernen und Vorhersage

- Berechnung Bayessche Vorhersage

$$y_* = \arg\max_y p(y \mid \mathbf{x}, L)$$

Summenregel

$$= \arg\max_y \int p(y, \theta \mid \mathbf{x}, L) d\theta$$

θ Modell

Produktregel

$$= \arg\max_y \int p(y \mid \theta, \mathbf{x}) p(\theta \mid L) d\theta$$

Bayesian Model Averaging

Vorhersage, gegeben Modell

Modell gegeben Trainingsdaten

- Bayes‘sche Vorhersage:
  - Mitteln der Vorhersage über alle Modelle.
  - Gewichtung: wie gut passt Modell zu Trainingsdaten.

Bayessches Lernen und Vorhersage

- Bayessche Vorhersage praktikabel?
  - $y_* = \arg\max_y \; p(y \mid \mathbf{x}, L)$

    $= \arg\max_y \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid L) d\theta$

    - Bayesian Model Averaging: Mitteln über i.A. unendlich viele Modelle
    - Wie berechnen? Nur manchmal praktikabel, geschlossene Lösung.
- Kontrast zu Entscheidungsbaumlernen:
  - Finde **ein** Modell, das gut zu den Daten passt.
  - Triff Vorhersagen für neue Instanzen basierend auf diesem Modell.
  - Trennt zwischen Lernen eines Modells und Vorhersage.

Bayessches Lernen und Vorhersage

- Wie Bayes-Hypothese ausrechnen?

$$y_* = \arg\max_y p(y \mid \mathbf{x}, L)$$

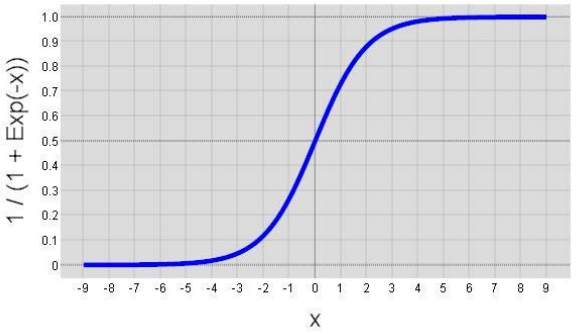
$$= \arg\max_y \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid L)\, d\theta$$

- Wir brauchen:
  - 1) Wsk für Klassenlabel gegeben Modell, $p(y \mid \mathbf{x}, \theta)$

z.B. linearer probabilistischer Klassifikator (logistische Regression)

$$p(y = 1 \mid \mathbf{x}, \theta) = \sigma(\mathbf{w}^T \mathbf{x})$$

$$p(y = 0 \mid \mathbf{x}, \theta) = \sigma(-\mathbf{w}^T \mathbf{x})$$



Bayessches Lernen und Vorhersage

- Wie Bayes-Hypothese ausrechnen?

$$y_* = \arg\max_y \; p(y \mid \mathbf{x}, L)$$

$$= \arg\max_y \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid L) d\theta$$

- Wir brauchen:
  - 2) Wsk für Modell gegeben Daten, a-posteriori-Wahrscheinlichkeit $p(\theta \mid L)$

→ Ausrechnen mit Bayes Regel

Bayessches Lernen und Vorhersage

- Berechnung der a-posteriori Verteilung über Modelle
  - Bayes‘ Gleichung

Likelihood,
Wie gut passt
Modell zu Daten?

Posterior,
A-Posteriori-
Verteilung

$$p(\theta \mid L) = \frac{p(L \mid \theta)\, p(\theta)}{p(L)}$$

Prior,
A-Priori-
Verteilung

$$= \frac{1}{Z}\, p(L \mid \theta)\, p(\theta)$$

Normierungskonstante

Bayessche Regel:
„Posterior = Likelihood x Prior“

Bayessche Regel

- Brauchen: Likelihood $p(L \mid \theta)$.
  - Wie wahrscheinlich wären die Trainingsdaten, wenn θ das richtige Modell wäre.
  - Wie gut passt Modell zu den Daten.
  - Typischerweise Unabhängigkeitsannahme:

$$L = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$$

$$p(L \mid \theta) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \theta)$$

Wahrscheinlichkeit des in L beobachteten Klassenlabels gegeben Modell θ

Bayessche Regel

- Brauchen: Prior $p(\theta)$.
  - Wie wahrscheinlich ist Modell θ bevor wir irgendwelche Trainingsdaten gesehen haben.
- Annahmen über $p(\theta)$ drücken datenunabhängiges Vorwissen über Problem aus.
- Beispiel lineare Modelle:
  -

Bayessche Regel

- Brauchen: Prior $p(\theta)$.
  - Wie wahrscheinlich ist Modell θ bevor wir irgendwelche Trainingsdaten gesehen haben.
- Annahmen über $p(\theta)$ drücken datenunabhängiges Vorwissen über Problem aus.
- Beispiel lineare Modelle:
  - $|\mathbf{w}|^2$ möglichst niedrig ($\mathbf{w} = \theta$)

Bayessche Regel

- Brauchen: Prior $p(\theta)$.
  - Wie wahrscheinlich ist Modell θ bevor wir irgendwelche Trainingsdaten gesehen haben.
- Annahmen über $p(\theta)$ drücken datenunabhängiges Vorwissen über Problem aus.
- Beispiel Entscheidungsbaumlernen:
  -

Bayessche Regel

- Brauchen: Prior $p(\theta)$.
  - Wie wahrscheinlich ist Modell θ bevor wir irgendwelche Trainingsdaten gesehen haben.
- Annahmen über $p(\theta)$ drücken datenunabhängiges Vorwissen über Problem aus.
- Beispiel Entscheidungsbaumlernen:
  - Kleine Bäume sind in vielen Fällen besser als komplexe Bäume.
  - Algorithmen bevorzugen deshalb kleine Bäume.

Zusammenfassung Bayessche Vorhersage

- Um Risiko einer Fehlentscheidung zu minimieren: wähle Bayessche Vorhersage

$$y_* = \arg\max_y \; p(y \mid \mathbf{x}, L)$$

$$= \arg\max_y \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid L)\, d\theta$$

- Problem: In vielen Fällen gibt es keine geschlossene Lösung, Integration über alle Modelle unpraktikabel.
- Maximum-A-Posteriori- (MAP-)Hypothese: wähle

$$\theta_* = \arg\max_\theta \; p(\theta \mid L)$$

$$y_* = \arg\max_y \; p(y \mid \mathbf{x}, \theta_*)$$

- Entspricht Entscheidungsbaumlernen.
  - Finde bestes Modell aus Daten,
  - Klassifiziere nur mit diesem Modell.

Zusammenfassung Bayessche Vorhersage

- Um MAP-Hypothese zu bestimmen müssen wir Posterior (Likelihood x Prior) kennen.
- Unmöglich, wenn kein Vorwissen (Prior) existiert.
- Maximum-Likelihood- (ML-)Hypothese:
  - $$\theta_* = \arg\max_\theta \; p(L \mid \theta)$$
    $$y_* = \arg\max_y \; p(y \mid \mathbf{x}, \theta_*)$$
  - Berücksichtigt nur Beobachtungen in L, kein Vorwissen.
  - Problem der Überanpassung an Daten

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayes‘sche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayes‘sche Lineare Regression, Naive Bayes

Parameter von Verteilungen schätzen

- Oft können wir annehmen, dass Daten einer bestimmten Verteilung folgen
  - Z.B. Binomialverteilung für *N* Münzwürfe
  - Z.B. Gaußverteilung für Körpergröße, IQ, …
- Diese Verteilungen sind parametrisiert
  - Binomialverteilung: Parameter μ ist Wahrscheinlichkeit für „Kopf“
  - Gaußverteilung: Parameter μ, σ für Mittelwert und Standardabweichung
- „Echte“ Wahrscheinlichkeiten/Parameter kennen wir nie.
- Welche Aussagen über echte Wahrscheinlichkeiten können wir machen, gegeben Daten?

Parameter von Verteilungen schätzen

- Problemstellung Parameter von Verteilungen schätzen:
  - Gegeben parametrisierte Familie von Verteilungen (z.B. Binomial, Gauß) mit Parametervektor θ
  - Gegeben Daten L: Ausprägungen der Zufallsvariable
  - Gesucht: a-posteriori Verteilung $P(\theta \mid L)$ bzw. maximum a-posteriori Schätzung

$$\theta^* = \arg\max_\theta P(\theta \mid L)$$

- Verwende Bayessche Regel:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

Binomialverteilte Daten Schätzen

- Beispiel: Münzwurf, schätze Parameter $\mu = \theta$
  - N Mal Münze werfen.
  - Daten L: N_k mal Kopf, N_z mal Zahl.

- Beste Schätzung θ gegeben L? Bayes‘ Gleichung:

Likelihood der Daten gegeben Parameter, wie gut erklärt Parameter die Beobachtungen?

A-priori Verteilung über Parameter, repräsentiert Vorwissen

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

A-posteriori Verteilung über Parameter, charakterisiert wahrscheinliche Parameterwerte und verbleibende Ungewissheit

Wahrscheinlichkeit der Daten, nur Normalisierer

Binomialverteilte Daten Schätzen

- Likelihood der Daten:

$$P(L \mid \theta)$$

($\theta = \mu$ Wahrscheinlichkeit für „Kopf“)

- Likelihood ist binomialverteilt:

$$P(L \mid \theta) = P(N_k, N_z \mid \theta) = \mathrm{Bin}(N_k \mid N, \theta) \qquad N = N_k + N_z$$

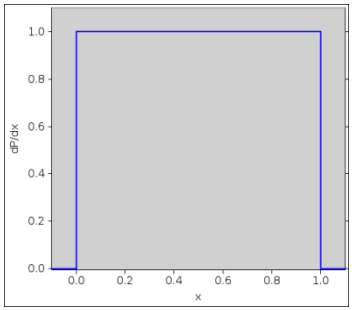
$$= \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z}$$

Wahrscheinlichkeit, bei N Münzwürfen N_k-mal Kopf und N_Z-mal Zahl zu sehen, für Münzparameter θ

Binomialverteilte Daten Schätzen

- Was ist der Prior $P(\theta)$ im Münzwurfbeispiel?
- 1) Versuch: Kein Vorwissen

$$P(\theta) = \begin{cases} 1 : 0 \le \theta \le 1 \\ 0 : sonst \end{cases} \qquad \text{Dichte}$$



- Beispiel:
  - Daten L = {Zahl,Zahl,Zahl}
  - MAP Modell:

$$\theta^* = \arg\max_{\theta\in[0,1]} P(\theta \mid L) = \arg\max_{\theta\in[0,1]} \frac{P(L \mid \theta)P(\theta)}{P(L)}$$

$$= \arg\max_{\theta\in[0,1]} P(L \mid \theta) = \arg\max_{\theta\in[0,1]} \binom{3}{0} \theta^0 (1-\theta)^3 = 0$$

- Schlussfolgerung: Münze wird **niemals** „Kopf“ zeigen
  - Schlecht, Überanpassung an Daten („Overfitting“)

Binomialverteilte Daten Schätzen

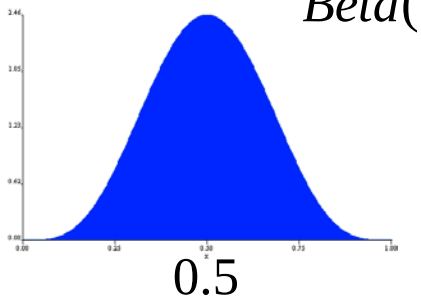
- Was ist der Prior $P(\theta)$ im Münzwurfbeispiel?
- Besser mit Vorwissen: Unwahrscheinlich, dass Münze immer Kopf oder immer Zahl zeigt
- Gutes Modell für Vorwissen über θ: Beta-Verteilung.

$$P(\theta) = Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

$(\theta \in [0,1])$

$Beta(\theta \mid 5,5)$



0.5

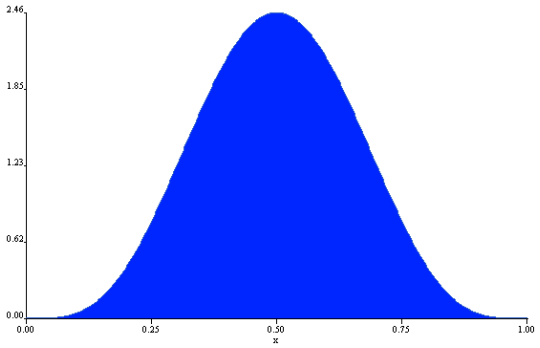
- Gamma-Funktion $\Gamma(\alpha)$ kontinuierliche Fortsetzung der Fakultätsfunktion

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \qquad \forall n \in \mathbb{N} : \Gamma(n) = (n-1)!$$

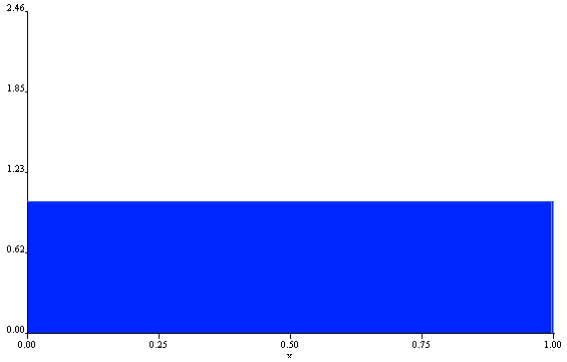
Binomialverteilte Daten Schätzen

- α_k und α_z sind Parameter der Beta-Verteilung („Hyperparameter“)
- Beta-Verteilung ist Verteilung über Verteilungen

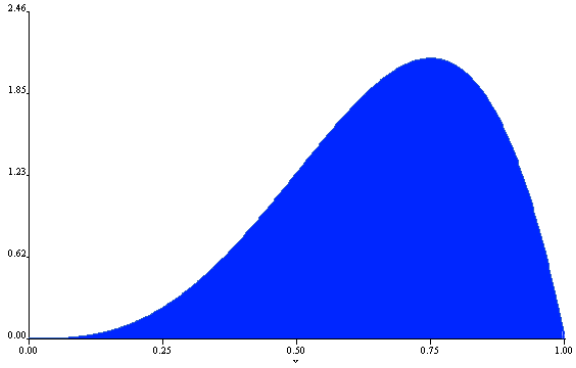
$\alpha_K = 5, \quad \alpha_Z = 5$



$\alpha_K = 1, \quad \alpha_Z = 1$



$\alpha_K = 4, \quad \alpha_Z = 2$



- Normalisierte Dichte $\int_0^1 Beta(\theta \mid \alpha_K, \alpha_Z) d\theta = 1$

Binomialverteilte Daten Schätzen

- Warum gerade diese a-priori-Verteilung?
- Strukturelle Ähnlichkeit mit Likelihood:

Prior $$P(\theta) = Beta(\theta \mid \alpha_k, \alpha_z) = \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1}(1-\theta)^{\alpha_z - 1}$$

Likelihood $$P(L \mid \theta) = Bin(N_k \mid N, \theta) = \binom{N_k + N_Z}{N_k} \theta^{N_k}(1-\theta)^{N_z}$$

- Einfach, Beobachtungen zu berücksichtigen: Produkt aus Likelihood und Prior hat wieder dieselbe Form wie Prior

$$P(\theta \mid L) \propto P(L \mid \theta) P(\theta)$$

Binomialverteilte Daten Schätzen

- Wenn wir den Beta-Prior in Bayes‘ Gleichung einsetzen, dann:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

$$= \frac{1}{Z} \mathrm{Bin}(N_K \mid N, \theta) Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{1}{Z} \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z} \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1} (1-\theta)^{\alpha_z - 1}$$

$$= \frac{1}{Z'} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= ?$$

Binomialverteilte Daten Schätzen

- Wenn wir den Beta-Prior in Bayes‘ Gleichung einsetzen, dann:

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

$$= \frac{1}{Z} \mathrm{Bin}(N_K \mid N, \theta) Beta(\theta \mid \alpha_k, \alpha_z)$$

$$= \frac{1}{Z} \binom{N_k + N_z}{N_k} \theta^{N_k} (1-\theta)^{N_z} \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \theta^{\alpha_k - 1} (1-\theta)^{\alpha_z - 1}$$

$$= \frac{1}{Z'} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= \frac{\Gamma(\alpha_k + N_k + \alpha_z + N_z)}{\Gamma(\alpha_k + N_k)\Gamma(\alpha_z + N_z)} \theta^{\alpha_k + N_k - 1} (1-\theta)^{\alpha_z + N_z - 1}$$

$$= Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$

- **Beta-Verteilung ist „konjugierter“ Prior: Posterior ist wieder Beta-verteilt**

Zusammenfassung Bayessche Parameterschätzung Binomialverteilung

- Zusammenfassung Berechnung der a-posteriori Verteilung:
- Bayessche Regel

$$P(\theta \mid L) = \frac{P(L \mid \theta) P(\theta)}{P(L)}$$

- Posterior $P(\theta \mid L)$: Wie wahrscheinlich ist Modell θ, nachdem wir Daten L gesehen haben?

- Vorwissen $P(\theta)$ und Evidenz der Trainingsdaten L werden zu neuem Gesamtwissen $P(\theta \mid L)$ integriert.

- Beispiel Münzwurf: Vorwissen Beta($\theta \mid \alpha_k, \alpha_z$) und Beobachtungen N_k, N_z werden zu Posterior Beta($\theta \mid \alpha_k + N_k, \alpha_z + N_z$).

Münzwurf: Wahrscheinlichste Wahrscheinlichkeit

- Wahrscheinlichster Parameter θ.

$$\arg\max_\theta P(\theta \mid L) = \arg\max_\theta Beta(\theta \mid \alpha_k + N_k, \alpha_z + N_z)$$

$$= \arg\max_\theta \frac{\Gamma(\alpha_k + \alpha_z + N_k + N_z)}{\Gamma(\alpha_k + N_k)\Gamma(\alpha_z + N_z)} \theta^{\alpha_k + N_k - 1}(1-\theta)^{\alpha_z + N_z - 1}$$

Ableiten, Ableitung null setzen ($\alpha_z \geq 1,\ \alpha_k \geq 1$)

$$= \frac{N_k + \alpha_k - 1}{N_k + N_z + \alpha_k + \alpha_z - 2}$$

Normalisierer, unabhängig von θ

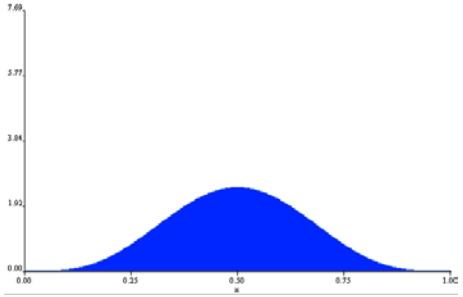
  - Für $\alpha_z = \alpha_k = 1$ ergibt sich ML Schätzung

- Interpretation der Hyperparameter $\alpha_z - 1 / \alpha_k - 1$:
  - $\alpha_z - 1 / \alpha_k - 1$ „Pseudocounts“ , die auf beobachtete „Counts“ N_z / N_k aufgeschlagen werden
  - wie oft im Leben Münzwurf mit „Kopf“/“Zahl“ gesehen?

Münzwurf: Wahrscheinlichste Wahrscheinlichkeit

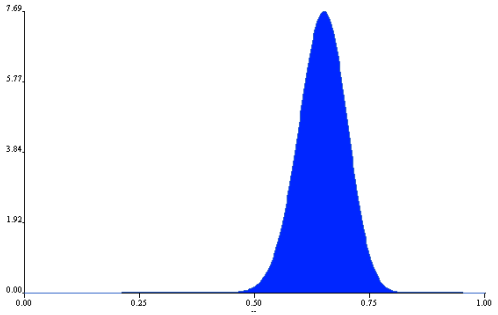
- Beispiel MAP Schätzung Parameter

Prior $P(\theta) = Beta(\theta \mid 5,5)$



Posterior nach L={50x Kopf, 25x Zahl}:

$P(\theta \mid L) = Beta(\theta \mid 55,30)$



$N_k = 50, N_z = 25, \alpha_k = 5, \alpha_z = 5$

MAP Schätzung: $\theta^* = \arg\max_\theta P(\theta \mid L) = \frac{54}{54+29} \approx 0.65$

Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayessches Lernen (3)

Christoph Sawade/Niels Landwehr
Tobias Scheffer

Überblick

- Wahrscheinlichkeiten, Erwartungswerte, Varianz
- Grundkonzepte des Bayesschen Lernens
- (Bayessche) Parameterschätzung für Wahrscheinlichkeitsverteilungen
- Bayessche Lineare Regression, Naive Bayes

Überblick

- Bayessche Lineare Regression
- Modellbasiertes Klassifikationslernen: Naive Bayes

Wiederholung: Regression

- Regressionsproblem:
  - Trainingsdaten

$$L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$ Merkmalsvektoren

$y_i \in \mathbb{R}$ reelles Zielattribut

  - Matrixschreibweise

Merkmalsvektoren

$$X = (\mathbf{x}_1 \quad ... \quad \mathbf{x}_N) = \begin{pmatrix} x_{11} & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1m} & x_{Nm} \end{pmatrix}$$

Zugehörige Labels (Werte Zielattribut)

$$\mathbf{y} = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix}$$

- Problemstellung Vorhersage:
  - Gegeben L, neues Testbeispiel $\mathbf{x}$
  - Finde optimale Vorhersage y für $\mathbf{x}$

Exkurs: Multivariate Normalverteilung

- Zufallsvariable **x** mit *d* Dimensionen.

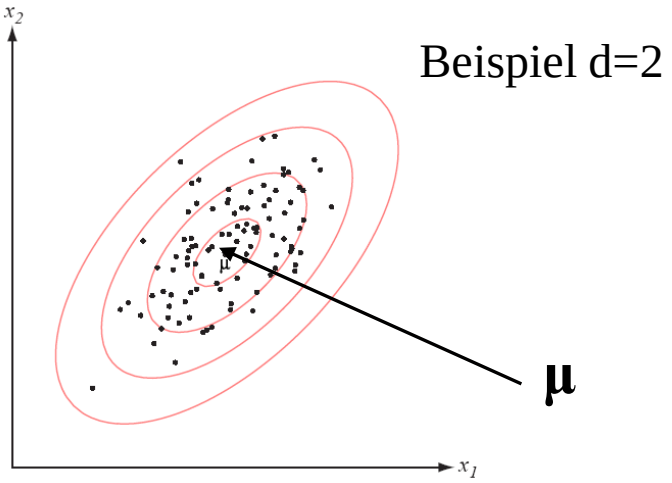
$\mathbf{x} \in \mathbb{R}^d$ normalverteilt, wenn Verteilung beschrieben wird durch Dichte

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Determinante

- Mittelwertvektor $\boldsymbol{\mu} \in \mathbb{R}^d$

- Kovarianzmatrix Σ



- Koarianzmatrix entscheidet, wie Punkte streuen

Wiederholung: Lineare Regression

- Modellraum lineare Regression:

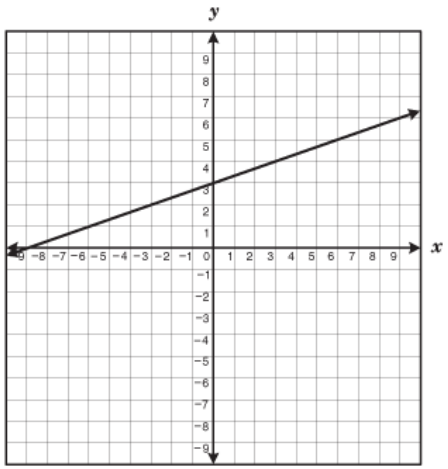
$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

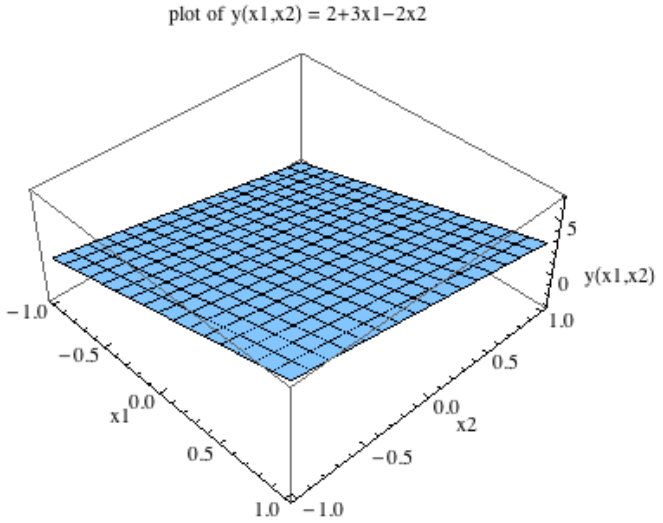
$$= w_0 + \sum_{i=1}^{m} w_i x_i$$

$\mathbf{w}$ „Parametervektor“, „Gewichtsvektor“

Zusätzliches konstantes Attribut $x_0 = 1$

- Lineare Abhängigkeit von $f_{\mathbf{w}}(\mathbf{x})$ von Parametern $\mathbf{w}$
- Lineare Abhängigkeit von $f_{\mathbf{w}}(\mathbf{x})$ von Eingaben $\mathbf{x}$





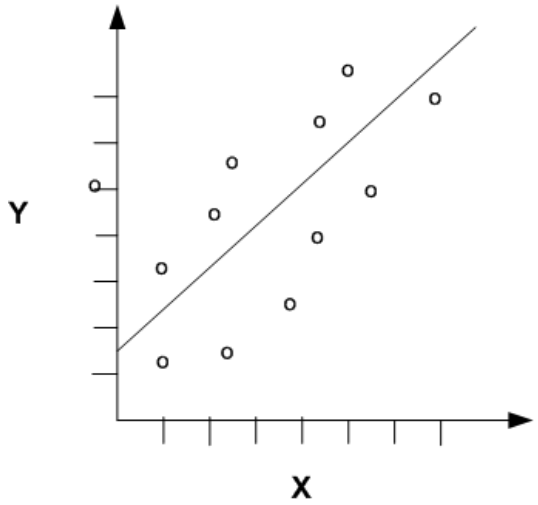
Bayessche Regression: Daten

- Modellvorstellung beim Bayesschen Lernen: Prozess der Datengenerierung
  - „Echtes“ Modell f_* wird aus Prior-Verteilung $P(f)$ gezogen
  - Merkmalsvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_N$ werden unabhängig voneinander gezogen (nicht modelliert)
  - Für jedes $\mathbf{x}_i$ wird das Label y_i gezogen nach Verteilung $P(y_i \mid \mathbf{x}_i, f_*)$ (Anschauung: $y_i \approx f_*(\mathbf{x}_i)$)
  - Daten L fertig generiert

- Wie sieht $P(y_i \mid \mathbf{x}_i, f_*)$ für Regressionsprobleme aus?

Bayessche Regression: Daten

- Annahme, dass es „echtes“ Modell $f_*(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w}_*$ gibt, dass die Daten perfekt erklärt, unrealistisch
  - Daten folgen nie genau einer Regressions-Geraden/Ebene

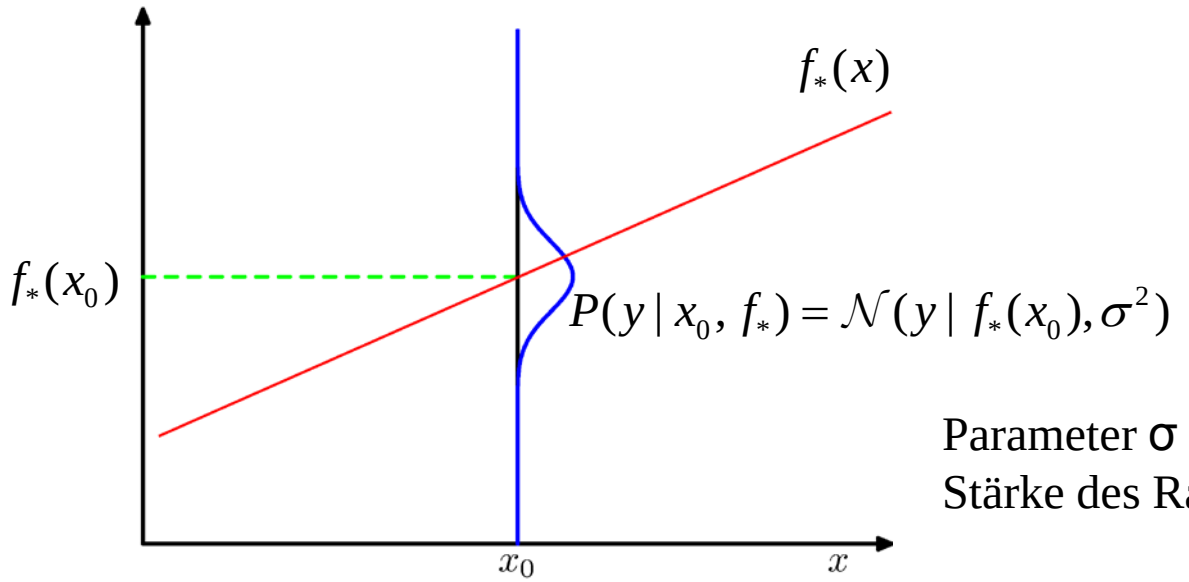


- Alternative Annahme: Daten folgen $f_*(\mathbf{x})$ bis auf kleine, zufällige Abweichungen (Rauschen)

Bayessche Regression: Daten

- Alternative Annahme: Daten folgen $f_*(\mathbf{x})$ bis auf kleine, zufällige Abweichungen (Rauschen)
- Modellvorstellung:
  - Zielattribut y generiert aus $f_*(\mathbf{x})$ plus normalverteiltes Rauschen

$$y = f_*(\mathbf{x}) + \varepsilon \quad \text{mit} \quad \varepsilon \sim \mathcal{N}(\varepsilon \mid 0, \sigma^2)$$



Parameter σ modelliert Stärke des Rauschens

Bayessche Regression: Vorhersageverteilung

- Ziel: Bayessche Vorhersage
  - $y_* = \arg\max_y P(y \mid \mathbf{x}, L)$
- Erinnerung: Berechnung mit Bayesian Model Averaging

Vorhersage, gegeben Modell

Modell gegeben Trainingsdaten

$$P(y \mid \mathbf{x}, L) = \int P(y \mid \mathbf{x}, \theta) P(\theta \mid L) d\theta$$

$$P(\theta \mid L) = \frac{1}{Z} P(L \mid \theta) P(\theta)$$

Likelihood:Trainingsdaten gegeben Modell

Prior über Modelle

Bayessche Regression: Likelihood

- Likelihood der Daten L:

Ziehen der $\boldsymbol{x}_i$ nicht modelliert

$$P(\mathbf{y} \mid X, \mathbf{w}) = P(y_1, ..., y_N \mid X, \mathbf{w})$$

Beispiele unabhängig

$$= \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$f_{\mathbf{w}}(\mathbf{x}_i) = \mathbf{x}_i^T \mathbf{w}$$

Nachrechnen: Multidimensionale Normalverteilung mit Kovarianzmatrix $\sigma^2 \mathbf{I}$

$$= \prod_{i=1}^{N} \mathcal{N}(y_i \mid \mathbf{x}_i^T \mathbf{w}, \sigma^2)$$

$$= \mathcal{N}\left(\mathbf{y} \mid X^{\mathrm{T}} \mathbf{w}, \sigma^2 \mathbf{I}\right)$$

Einheitsmatrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix}$$

$$X^{\mathrm{T}} \mathbf{w} = \begin{pmatrix} \mathbf{x}_1^T \mathbf{w} \\ ... \\ \mathbf{x}_N^T \mathbf{w} \end{pmatrix}$$

Vektor der Vorhersagen

Bayessche Regression: Prior

- Bayessches Lernen: Prior über Modelle f
  - Modelle parametrisiert durch Gewichtsvektor $\mathbf{w}$
  - Prior $P(\mathbf{w})$ über Gewichtsvektoren
- Geeignete Prior-Verteilung: Normalverteilung
  - Normalverteilung ist konjugiert zu sich selbst,
  - normalverteilter Prior und normalverteilte Likelihood ergeben wieder normalverteilten Posterior
  - Deshalb

    $\mathbf{w} \sim \mathcal{N}(\mathbf{w} \mid 0, \Sigma_p)$ „erwarten kleine Attributgewichte, $|\mathbf{w}|^2$ klein“

    Σ_p Kovarianzmatrix, oft $\Sigma_p = \sigma_p^2 \mathbf{I}$

    $\sigma_p^2 \in \mathbb{R}$ steuert Stärke des Priors

Bayessche Regression: Posterior

- Posterior-Verteilung über Modelle gegeben Daten

$$P(\mathbf{w} \mid L) = \frac{1}{Z} P(L \mid \mathbf{w}) P(\mathbf{w})$$ Bayessche Regel

$$= \frac{1}{Z} \mathcal{N}(\mathbf{y} \mid X^{\mathrm{T}}\mathbf{w}, \sigma^2 I) \cdot \mathcal{N}(\mathbf{w} \mid 0, \Sigma_p)$$

Ohne Beweis

$$= \mathcal{N}(\mathbf{w} \mid \bar{\mathbf{w}}, A^{-1})$$

mit $\bar{\mathbf{w}} = \sigma^{-2} A^{-1} X\mathbf{y}$ $\quad A = \sigma^{-2} XX^{\mathrm{T}} + \Sigma_p^{-1}$

- Posterior ist wieder normalverteilt, mit neuem Mittelwert $\bar{\mathbf{w}}$ und Kovarianzmatrix A^{-1}

Bayessche Regression: Posterior

- Posterior:

$$p(\mathbf{w} \mid L) = \mathcal{N}(\mathbf{w} \mid \bar{\mathbf{w}}, A^{-1})$$

- MAP-Hypothese:
  - $\mathbf{w}_{MAP} = ?$

Bayessche Regression: Posterior

- Posterior:

$$p(\mathbf{w} \mid L) = \mathcal{N}(\mathbf{w} \mid \bar{\mathbf{w}}, A^{-1})$$

- MAP-Hypothese:
  - $\mathbf{w}_{MAP} = \bar{\mathbf{w}}$

$$= \sigma^{-2} A^{-1} X \mathbf{y}$$

Sequentielles Update des Posteriors

- Berechnung des Posterior als sequentielles Update: Aufmultiplizieren der Likelihood einzelner Instanzen

$$P(\mathbf{w} \mid L) \propto P(\mathbf{w}) P(L \mid \mathbf{w})$$

Instanzen unabhängig

$$= P(\mathbf{w}) \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i, \mathbf{w})$$

„Likelihood für y_i einzeln an Prior multiplizieren“

- Sei $P_0(\mathbf{w}) = P(\mathbf{w})$, $P_n(\mathbf{w})$ der Posterior, wenn wir nur die ersten n Instanzen in L verwenden:

$$P(\mathbf{w} \mid L) \propto \underbrace{\underbrace{\underbrace{\underbrace{P(\mathbf{w}) P(y_1 \mid \mathbf{x}_1, \mathbf{w})}_{P_1(\mathbf{w})} P(y_2 \mid \mathbf{x}_2, \mathbf{w})}_{P_2(\mathbf{w})} P(y_3 \mid \mathbf{x}_3, \mathbf{w})}_{P_3(\mathbf{w})} \cdot \ldots \cdot P(y_N \mid \mathbf{x}_N, \mathbf{w})}_{P_N(\mathbf{w})}$$

Sequentielles Update des Posteriors

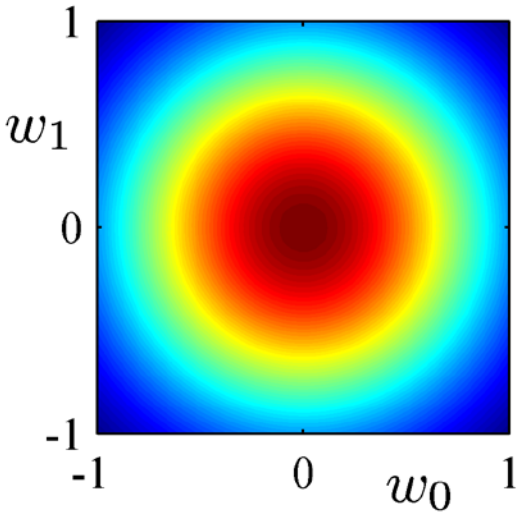
- Sequentielles Update:
  - Datenpunkte nacheinander anschauen
  - Neue Informationen (Datenpunkte) verändern Stück für Stück die Verteilung über $\mathbf{w}$

Beispiel Bayessche Regression

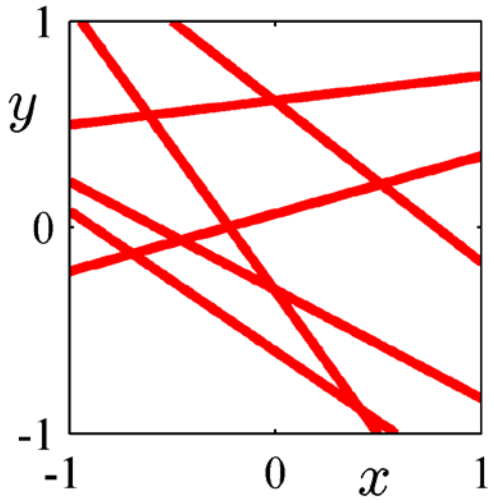
$f(x) = w_0 + w_1 x$ (eindimensionale Regression)

Sequentielles Update: $P_0(\mathbf{w}) = P(\mathbf{w})$

$P_0(\mathbf{w}) = P(\mathbf{w})$



Sample aus $P_0(\mathbf{w})$

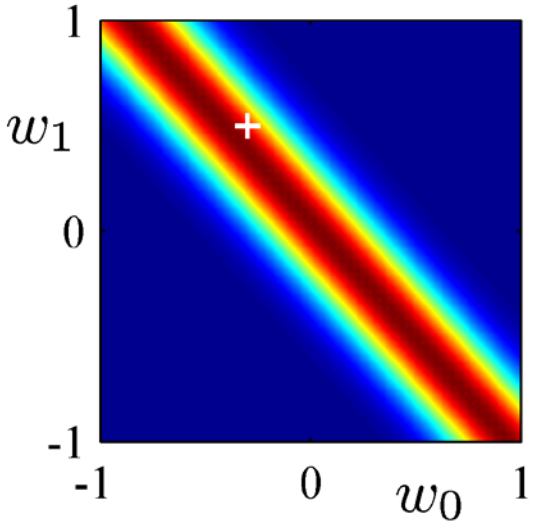


Beispiel Bayessche Regression

$f(x) = w_0 + w_1 x$ (eindimensionale Regression)

Sequentielles Update: $P_1(\mathbf{w}) \propto P_0(\mathbf{w}) P(y_1 \mid x_1, \mathbf{w})$

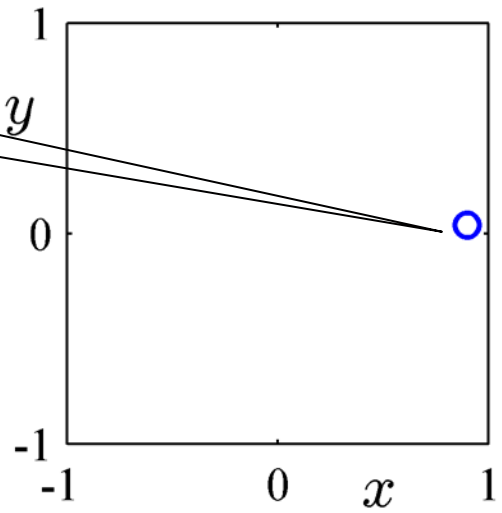
Likelihood $P(y_1 \mid x_1, \mathbf{w})$



Datenpunkt x_1, y_1

$$y_1 = f(x_1) + \varepsilon$$
$$= w_0 + w_1 x_1 + \varepsilon$$

$$\Rightarrow w_0 = -w_1 x_1 + y_1 - \varepsilon$$

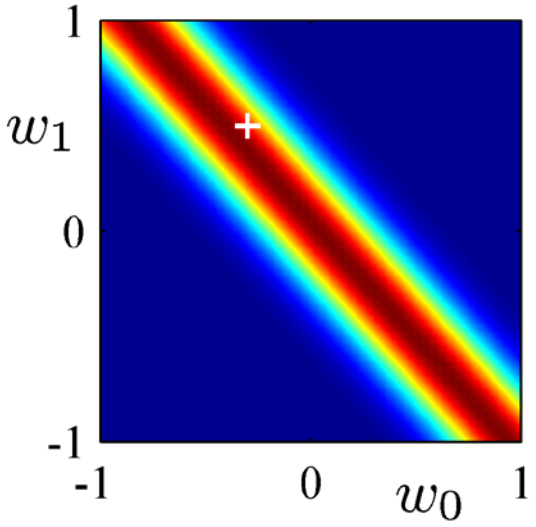


Beispiel Bayessche Regression

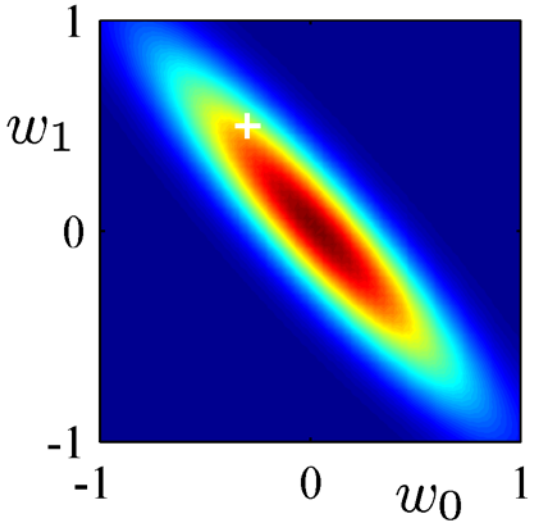
$f(x) = w_0 + w_1 x$ (eindimensionale Regression)

Sequentielles Update: $P_1(\mathbf{w}) \propto P_0(\mathbf{w}) P(y_1 \mid x_1, \mathbf{w})$

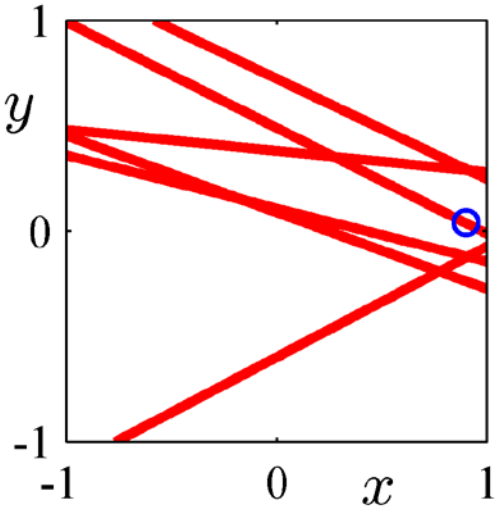
Likelihood $P(y_1 \mid x_1, \mathbf{w})$



Posterior $P_1(\mathbf{w})$



Sample aus $P_1(\mathbf{w})$

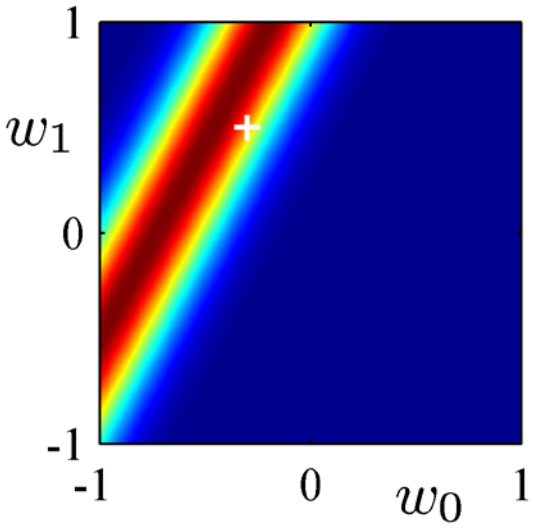


Beispiel Bayessche Regression

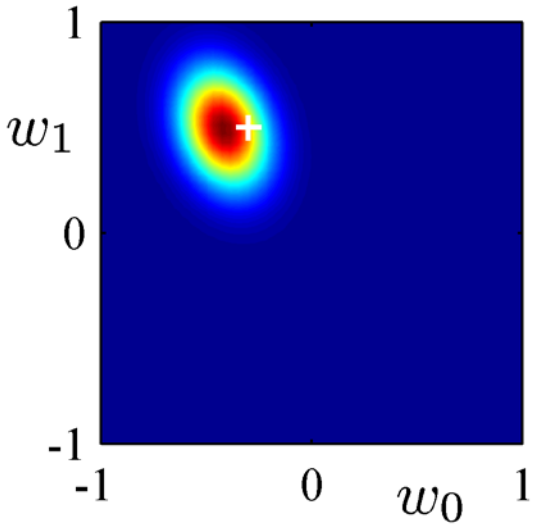
$f(x) = w_0 + w_1 x$ (eindimensionale Regression)

Sequentielles Update: $P_2(\mathbf{w}) \propto P_1(\mathbf{w}) P(y_2 \mid x_2, \mathbf{w})$

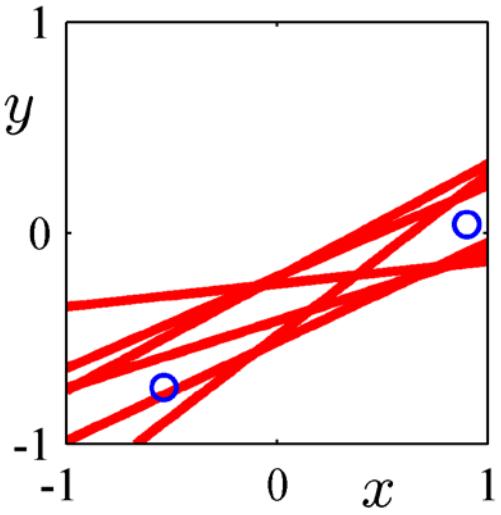
$P(y_2 \mid x_2, \mathbf{w})$



$P_2(\mathbf{w})$



Sample aus $P_2(\mathbf{w})$

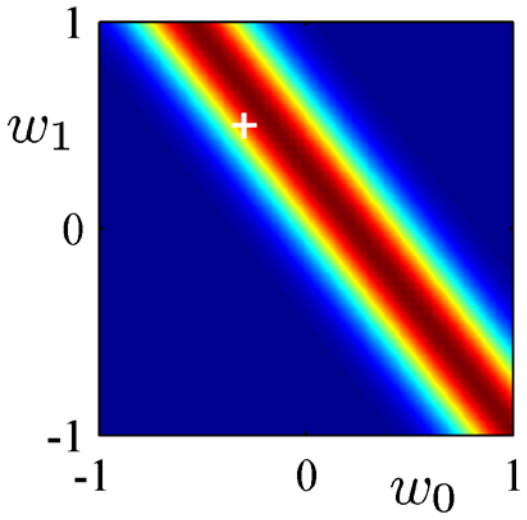


Beispiel Bayessche Regression

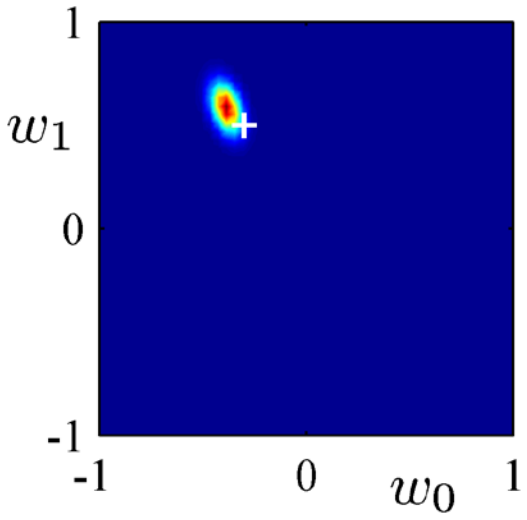
$f(x) = w_0 + w_1 x$ (eindimensionale Regression)

Sequentielles Update: $P_N(\mathbf{w}) \propto P_{N-1}(\mathbf{w}) P(y_N \mid x_N, \mathbf{w})$

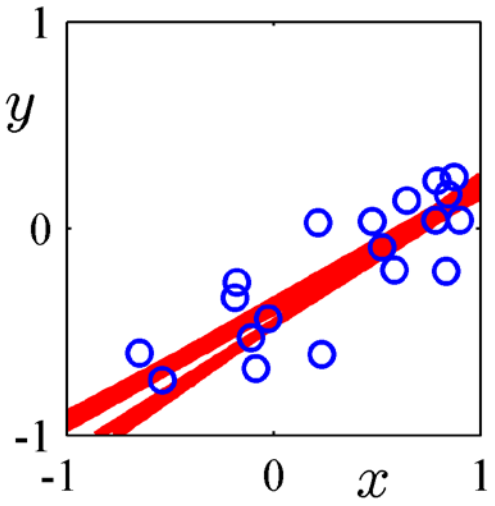
$P(y_N \mid x_N, \mathbf{w})$



$P_N(\mathbf{w})$



Sample aus $P_N(\mathbf{w})$



Bayessche Regression: Vorhersageverteilung

- Bayes‘sche Vorhersage: wahrscheinlichstes y.
  - $y_* = \arg\max_y P(y \mid \mathbf{x}, L)$
- Erinnerung: Berechnung mit Bayesian Model Averaging
  - $$P(y \mid \mathbf{x}, L) = \int P(y \mid \mathbf{x}, \theta) P(\theta \mid L) d\theta$$

Bayesian Model Averaging

Vorhersage, gegeben Modell

Modell gegeben Trainingsdaten

- Bayessche Vorhersage:
  - Mitteln der Vorhersage über alle Modelle.
  - Gewichtung: wie wahrscheinlich ist Modell a posteriori.

Bayessche Regression: Vorhersageverteilung

- Vorhersageverteilung wieder normalverteilt:

$$P(y \mid \mathbf{x}, L) = \int P(y \mid \mathbf{x}, \mathbf{w}) P(\mathbf{w} \mid L) d\mathbf{w}$$

$$= \int \mathcal{N}(y \mid \mathbf{x}^T \mathbf{w}, \sigma^2) \mathcal{N}(\mathbf{w} \mid \bar{\mathbf{w}}, A^{-1}) d\mathbf{w}$$

Ohne Beweis

$$= \mathcal{N}\left(y \mid \mathbf{x}^T \bar{\mathbf{w}}, \mathbf{x}^T A^{-1} \mathbf{x}\right)$$

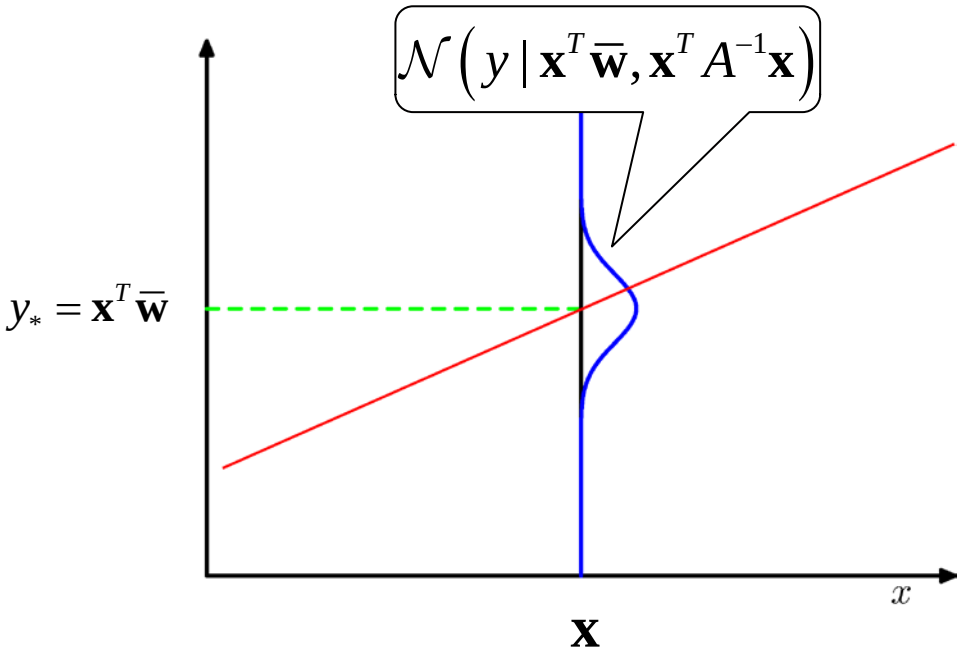
mit $\bar{\mathbf{w}} = \sigma^{-2} A^{-1} X \mathbf{y}$ $\qquad A = \sigma^{-2} X X^{\mathrm{T}} + \Sigma_p^{-1}$

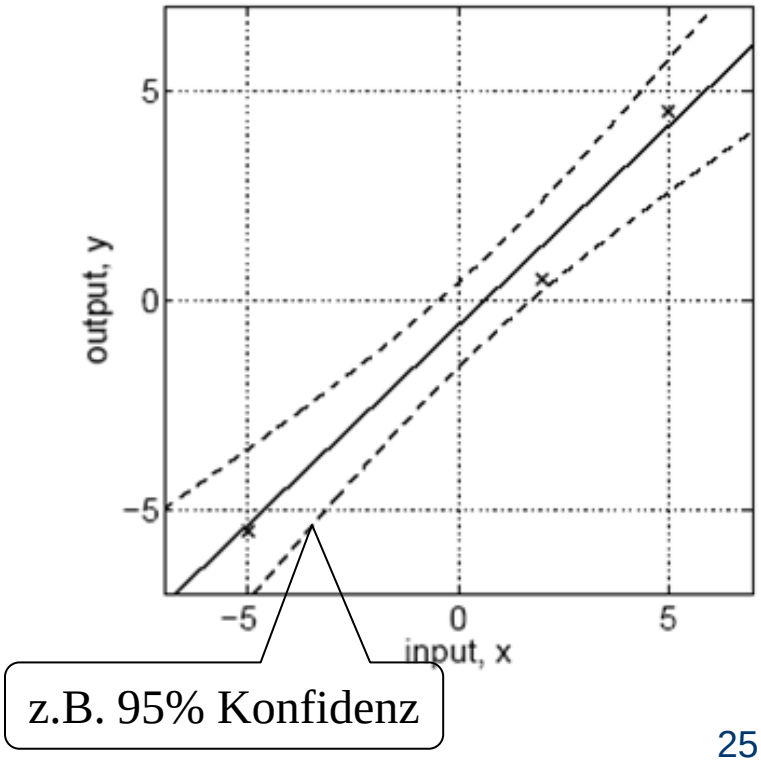
  - Optimale Vorhersage: Eingabevektor $\mathbf{x}$ wird mit $\bar{\mathbf{w}}$ multipliziert:

$$y_* = \mathbf{x}^T \bar{\mathbf{w}}$$

Bayessche Regression: Vorhersageverteilung

- Bayessche Regression liefert nicht nur optimale Vorhersage $y_* = \mathbf{x}^T \bar{\mathbf{w}}$ sondern Dichte von y und damit auch einen Konfidenzkorridor.

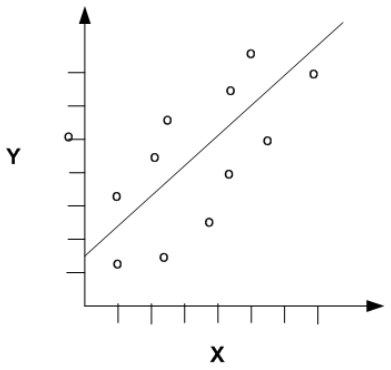




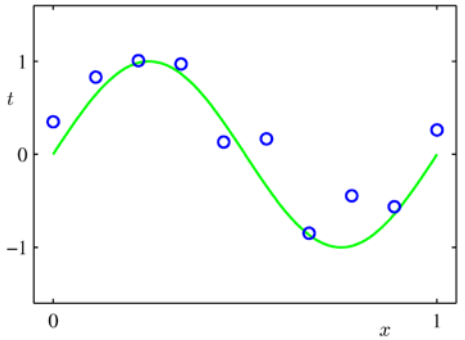
Nichtlineare Basisfunktionen

- Einschränkung der bisherigen Modelle: nur lineare Abhängigkeiten zwischen $\mathbf{x}$ und $f(\mathbf{x})$.

Lineare Daten



Nicht-lineare Daten



- In vielen Fällen nützlich: nicht-lineare Abhängigkeit

Nichtlineare Basisfunktionen

- Einfachster Weg: Lineare Regression auf nichtlinearen Basisfunktionen
  - Idee: Nicht auf den ursprünglichen $\mathbf{x}$ arbeiten, sondern auf nichtlinearer Transformation $\phi(\mathbf{x})$
  - Vorteil: Berechnung von posterior und Bayes'scher Vorhersage im Prinzip unverändert
- Basisfunktionen

$$\phi_1, ..., \phi_d : \mathbb{R}^m \to \mathbb{R}$$

$\mathbb{R}^m$ ursprünglicher Instanzenraum $\mathcal{X}$

$$\phi(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \dots \\ \phi_d(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^d$$

$$\phi : \mathbb{R}^m \to \mathbb{R}^d$$

meistens $d \gg m$

Nichtlineare Basisfunktionen

- Lineare Regression in den Basisfunktionen

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$
$$= w_0 + \sum_{i=1}^{d} w_i \phi_i(\mathbf{x})$$

*f(**x**)* ist lineare Kombination von Basisfunktionen

- Anschauung: Abbildung in höherdimensionalen Raum $\phi(\mathcal{X})$, lineare Regression dort

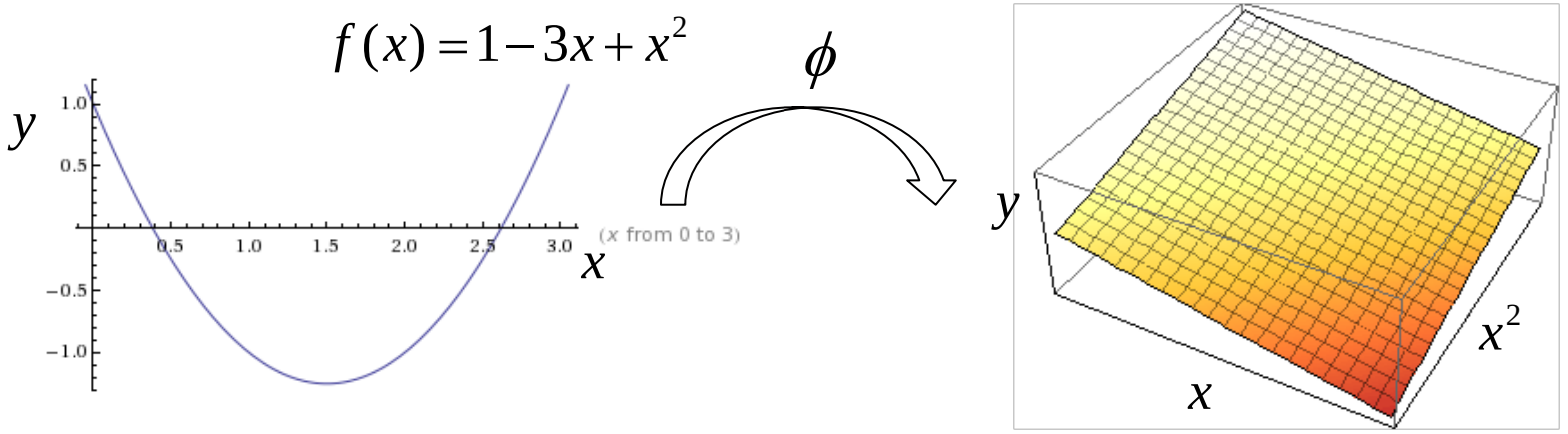
Nichtlineare Basisfunktionen: Beispiel

- Beispiel

$$\mathcal{X} = \mathbb{R} \qquad \phi_1(x) = x \qquad \phi_2(x) = x^2$$

$$f(x) = w_0 + w_1\phi_1(x) + w_2\phi_2(x)$$

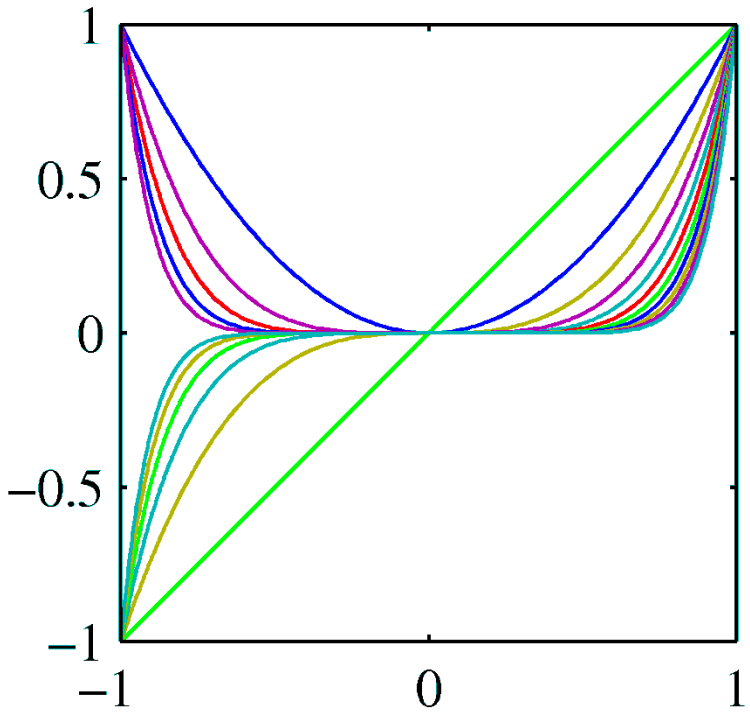
- Nichtlineare Funktion in x darstellbar als lineare Funktion in $\phi(x)$



Nichtlineare Basisfunktionen

- Beispiele für nicht-lineare Basisfunktionen
  - Polynome

$$\phi_j(x) = x^j$$



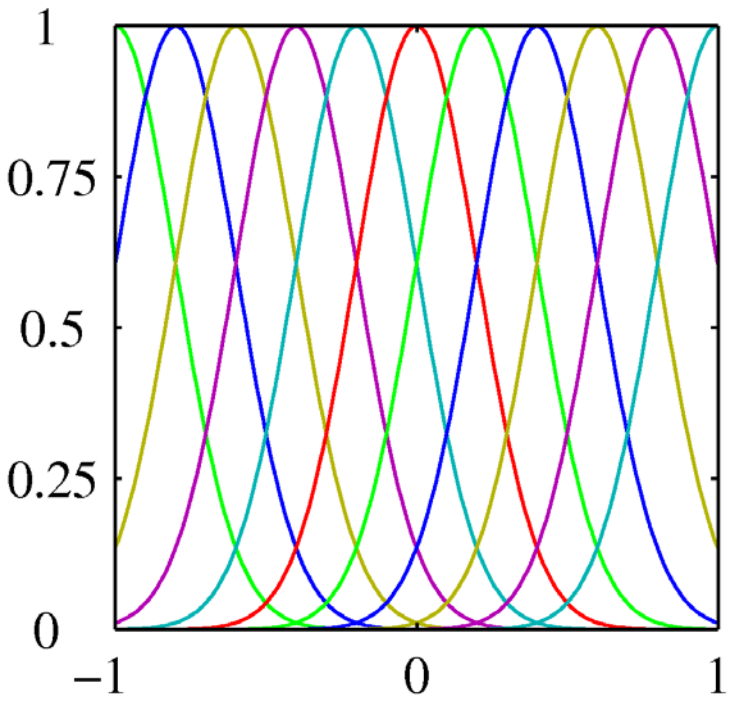
Nichtlineare Basisfunktionen

- Beispiele für nicht-lineare Basisfunktionen
  - Gauss-Kurven

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$

$\mu_1, ..., \mu_d$ Mittelpunkte

s^2 feste Varianz



Nichtlineare Basisfunktionen

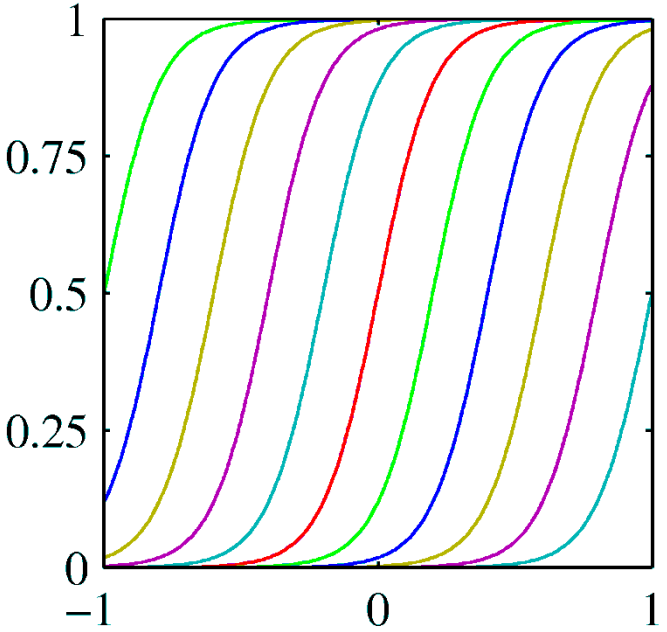
- Beispiele für nicht-lineare Basisfunktionen
  - Sigmoide

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

$\mu_1, \ldots, \mu_d$ Mittelpunkte

s feste Skalierung



Regression mit Basisfunktionen

- Funktion ϕ bildet m-dimensionalen Eingabevektor $\mathbf{x}$ auf d-dimensionalen Merkmalsvektor $\phi(\mathbf{x})$ ab.

- Regressionsmodell: $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x})$

- Optimale Vorhersage wie bisher, mit $\phi(\mathbf{x})$ statt $\mathbf{x}$.

Transformierte Testinstanz

$$P(y \mid \mathbf{x}, L) = \mathcal{N}\left(y \mid \phi(\mathbf{x})^T \overline{\mathbf{w}}, \phi(\mathbf{x})^T A^{-1} \phi(\mathbf{x})\right)$$

$$y_* = \arg\max_y \; p(y \mid \mathbf{x}, L) = \phi(\mathbf{x})^T \overline{\mathbf{w}}$$

Transformierte Datenmatrix

$$A = \sigma^{-2} \Phi\Phi^T + \Sigma_p^{-1}, \quad \overline{\mathbf{w}} = \sigma^{-2} A^{-1} \Phi \mathbf{y} \;\text{ und }\; \Phi = \phi(X)$$

Beispiel Regression mit Nichtlinearen Basisfunktionen

- Beispiel für Regression mit nicht-linearen Basisfunktionen
  - Generiere nicht-lineare Datenpunkte durch

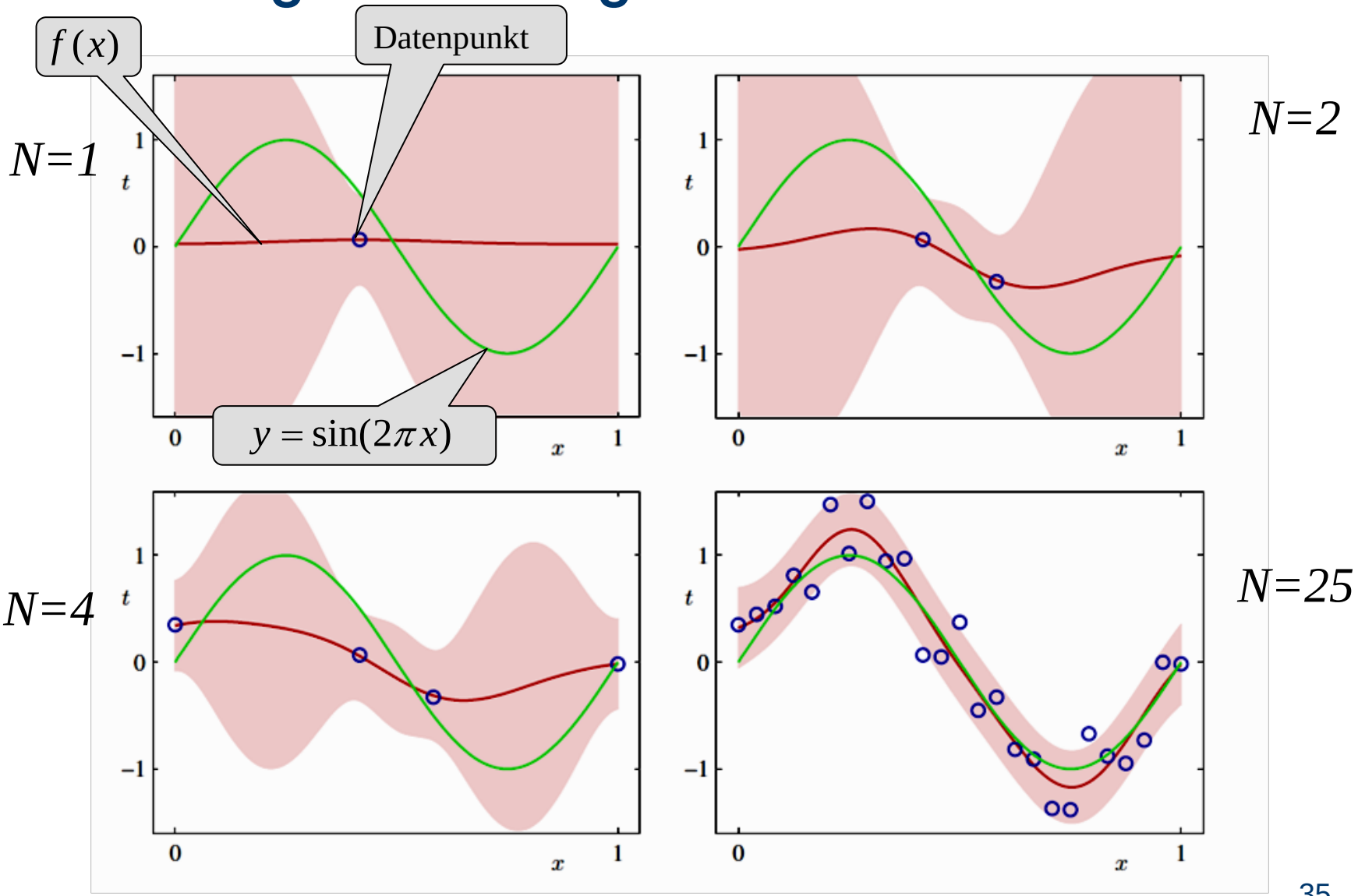
$$y = \sin(2\pi x) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(\varepsilon \mid 0, \sigma^2), \quad x \in [0,1]$$

  - 9 Gaussche Basisfunktionen

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right) \qquad \mu_1 = 0.1, \ldots, \mu_9 = 0.9$$

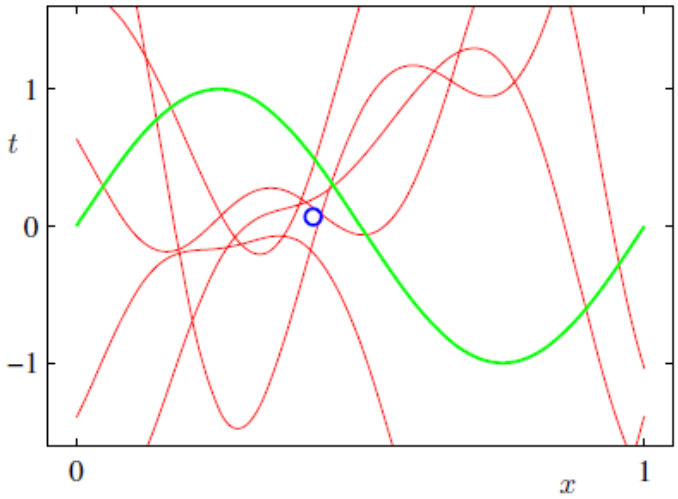
  - Wie sieht der Posterior $P(\mathbf{w} \mid L)$ und die Vorhersageverteilung $P(y \mid \mathbf{x}, L)$ aus?

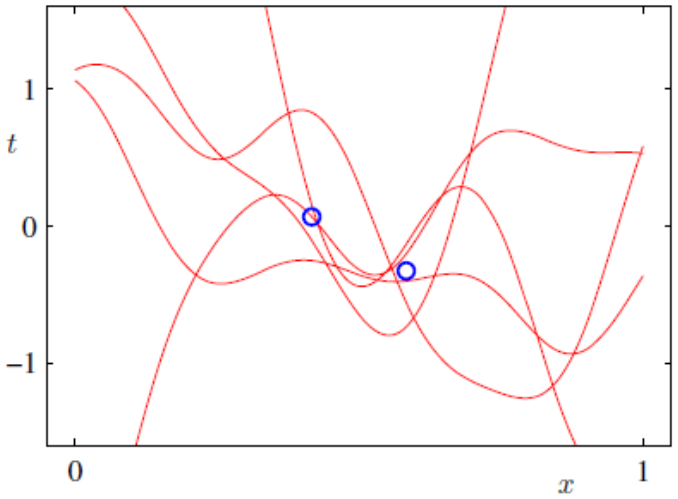
Vorhersageverteilung



Samples aus dem Posterior

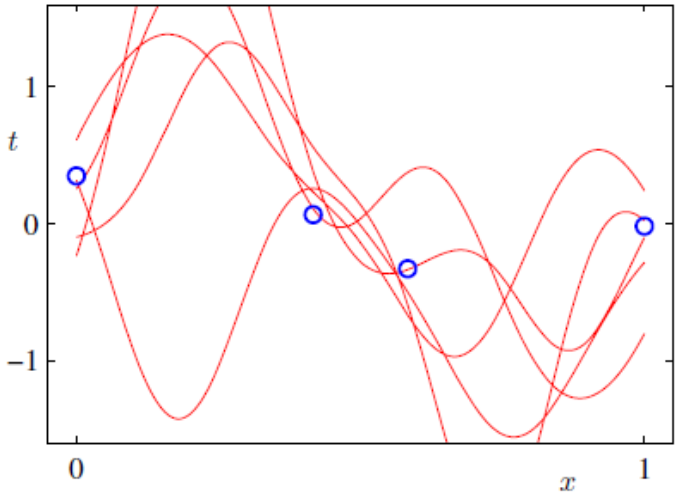
$N=1$

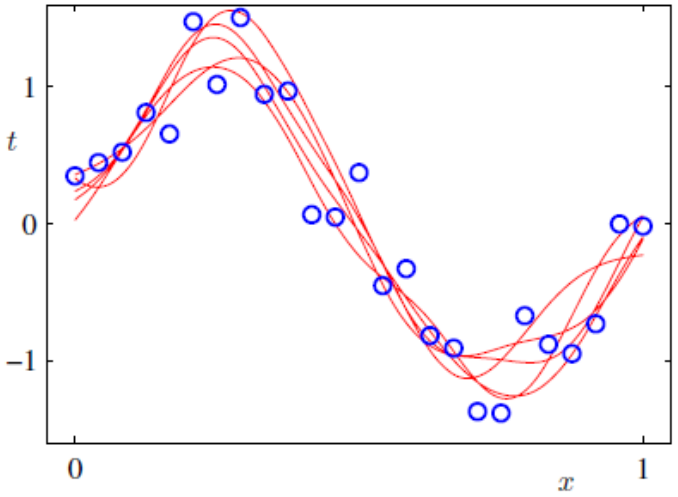




$N=2$

$N=4$





$N=25$

Überblick

- Bayessche Lineare Regression
- Modellbasiertes Klassifikationslernen: Naive Bayes

Klassifikationsprobleme

- Trainingsdaten L

$$L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i$ Merkmalsvektoren
y_i diskrete Klassenlabels

- Matrixschreibweise für Trainingsdaten L

Merkmalsvektoren X

$$X = (\mathbf{x}_1 \quad ... \quad \mathbf{x}_N) = \begin{pmatrix} x_{11} & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1m} & x_{Nm} \end{pmatrix}$$

Zugehörige Klassenlabel $\mathbf{y}$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix}$$

- Lernen: MAP Modell

$$\theta_{MAP} = \arg\max_\theta P(\theta \mid L)$$
$$= \arg\max_\theta P(L \mid \theta) P(\theta)$$

Modellbasiertes und Diskriminatives Lernen

- Likelihood $P(L \mid \theta)$: welcher Teil der Daten L wird modelliert?
- Diskriminatives Lernen:

Diskriminative Likelihood

$$\theta_{MAP} = \arg\max_{\theta} P(\theta) P(\mathbf{y} \mid X, \theta)$$

  - θ wird so gewählt, dass es Werte der Klassenvariable $\mathbf{y}$ in den Daten gut modelliert.
  - Klassifikator soll nur $\mathbf{y}$ für jedes $\mathbf{x}$ gut vorhersagen. Wozu also gute Modellierung von X berücksichtigen?
- Generatives (modellbasiertes) Lernen:

Generative Likelihood

$$\theta_{MAP} = \arg\max_{\theta} P(\theta) P(\mathbf{y}, X \mid \theta)$$

  - θ wird so gewählt, dass es Merkmalsvektoren X und Werte der Klassenvariable $\mathbf{y}$ in den Daten gut modelliert

Modellbasiert: Naive Bayes

- Naive Bayes: Modellbasierte Klassifikation

$$\theta_{MAP} = \arg\max_{\theta} P(\theta)P(\mathbf{y}, X \mid \theta)$$

- Likelihood der Daten L: N unabhängige Instanzen mit Klassenlabels

$$P(L \mid \theta) = P(\mathbf{x}_1,...,\mathbf{x}_N, y_1,...,y_N \mid \theta)$$
$$= \prod_{i=1}^{N} P(\mathbf{x}_i, y_i \mid \theta)$$

Modellbasiert: Naive Bayes

- Wie modellieren wir $P(\mathbf{x}, y \mid \theta)$?
- Gemeinsame Verteilung (Produktregel)

$$P(\mathbf{x}, y \mid \theta) = P(y \mid \theta) P(\mathbf{x} \mid y, \theta)$$

Klassenwahrscheinlichkeit:
z.B. *P(spam)* vs *P(nicht spam)*.

x-Verteilung gegeben Klasse:
z.B. Wortverteilung in Spam-Emails

- Wie modellieren wir $P(\mathbf{x} \mid y, \theta)$?
  - $\mathbf{x} = \begin{pmatrix} x_1 \\ \dots \\ x_m \end{pmatrix}$ hochdimensional, 2^m verschiedene Werte (x_i binär)
  - „Naive“ Unabhängigkeitsannahme

Naive Bayes: Unabhängigkeitsannahme

- Bedingte Unabhängigkeitsannahme:

$$P(\mathbf{x} \mid y, \theta) = \prod_{i=1}^{m} P(x_i \mid y, \theta)$$

„Attribute x_i unabhängig gegeben die Klasse y“

- Annahme: zwei Klassen, binäre Attribute $x_i \in \{0,1\}$
- Modellierte Verteilungen (Modellparameter):

$P(y \mid \theta)$ Bernoulli, mit Parameter $\theta^y = P(y = 1 \mid \theta)$

Für $i \in \{1,...,m\}$ (Attribute), $c \in \{0,1\}$ (Klassen):

$P(x_i \mid y = c, \theta)$ Bernoulli, mit Parameter $\theta^{x_i|c} = P(x_i = 1 \mid \theta, y = c)$

Naive Bayes: Likelihood

- Likelihood der Daten L mit bisherigen Annahmen:

$$P(L \mid \theta) = \prod_{j=1}^{N} P(\mathbf{x}_{j,} y_j \mid \theta)$$

Unabhängigkeit Instanzen

$$= \prod_{j=1}^{N} P(y_j \mid \theta) P(\mathbf{x}_j \mid y_j, \theta)$$

Produktregel

$$= \prod_{j=1}^{N} P(y_j \mid \theta^y) \prod_{i=1}^{m} P(x_{ji} \mid y_j, \theta^{x_i \mid y_j})$$

Bedingte Unabhängigkeit Attribute, „zuständige“ Modellparameter

y_j = Klassenlabel j-te Instanz

x_{ji} = Wert i-tes Merkmal j-te Instanz

Naive Bayes: Prior?

- Prior: Parametervektor θ besteht aus

  Parameter für Klassenverteilung θ^y

  Parameter für Merkmalsverteilungen $\theta^{x_i|0}$, $\theta^{x_i|1}$ $(i = 1,...,m)$

- Prior-Verteilung: unabhängig für einzelne Parameter

$$P(\theta) = P(\theta^y) \left(\prod_{i=1}^{m} P(\theta^{x_i|0}) P(\theta^{x_i|1}) \right)$$

  Prior Klassen-verteilung

  Prior Merkmalsverteilungen, gegeben positive/negative Klasse

- Konjugierter Prior Beta-Verteilung

$$P(\theta^y) \sim Beta(\theta^y \mid \alpha_0, \alpha_1)$$

Für $i \in \{1,...,m\}$ (Attribute), $c \in \{0,1\}$ (Klassen):

$$P(\theta^{x_i|c}) \sim Beta(\theta^{x_i|c} \mid \alpha_{x_i|c}, \alpha_{\overline{x}_i|c})$$

Naive Bayes: Posterior

- A-posteriori Verteilung wieder Beta: Standardlösung für Münzwurfszenario
- A-posteriori Verteilung für Parameter $P(\theta^y \mid L)$:

$$P(\theta^y \mid L) = Beta(\theta^y \mid \alpha_0 + N_0, \alpha_1 + N_1)$$

mit
N_0: Anzahl Beispiele mit Klasse 0 in L
N_1: Anzahl Beispiele mit Klasse 1 in L

$$\theta^y_{MAP} = \frac{N_1 + \alpha_1 - 1}{N_0 + \alpha_0 + N_1 + \alpha_1 - 2}$$

Naive Bayes: Posterior

- A-posteriori Verteilung für Parameter $P(\theta^{x_i|c})$:

Für $i \in \{1,...,m\}$ (Attribute), $c \in \{0,1\}$ (Klassen):

$$P(\theta^{x_i|c} \mid L) = Beta(\theta^{x_i|c} \mid \alpha_{x_i|c} + N_{x_i|c}, \alpha_{\overline{x}_i|c} + N_{\overline{x}_i|c})$$

mit
$N_{x_i|c}$: Anzahl Beispiele mit $x_i = 1$ und Klasse c in L
$N_{\overline{x}_i|c}$: Anzahl Beispiele mit $x_i = 0$ und Klasse c in L

$$\theta_{MAP}^{x_i|c} = \frac{N_{x_i|c} + \alpha_{x_i|c} - 1}{N_{x_i|c} + \alpha_{x_i|c} + N_{\overline{x}_i|c} + \alpha_{\overline{x}_i|c} - 2}$$

Naive Bayes: Lernalgorithmus

- Eingabe: $L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$
- Schätze Klassenverteilung:

Zähle
N_1: Anzahl Beispiele mit Klasse 1 in L
N_0: Anzahl Beispiele mit Klasse 0 in L

$$\theta^{y}_{MAP} = \frac{N_1 + \alpha_1 - 1}{N_0 + \alpha_0 + N_1 + \alpha_1 - 2}$$

- Für Klassen $y=0$ und $y=1$, für alle Attribute i:

Zähle
$N_{x_i|y}$: Anzahl Beispiele mit $x_i = 1$ und Klasse y in L
$N_{\overline{x}_i|y}$: Anzahl Beispiele mit $x_i = 0$ und Klasse y in L

$$\theta^{x_i|y}_{MAP} = \frac{N_{x_i|y} + \alpha_{x_i|y} - 1}{N_{x_i|y} + \alpha_{x_i|y} + N_{\overline{x}_i|y} + \alpha_{\overline{x}_i|y} - 2}$$

- Alle Modellparameter gelernt!

Naive Bayes: Klassifikation

- Eingabe: $\mathbf{x} = \begin{pmatrix} x_1 \\ \dots \\ x_m \end{pmatrix}$
- Rückgabe:

$$y_* = \arg\max_y P(y \mid \mathbf{x}, \theta_{MAP})$$

$$= \arg\max_y \underbrace{P(y \mid \theta^y_{MAP})}_{\text{Klassenverteilung}} \underbrace{\prod_{i=1}^{m} P(x_i \mid y, \theta^{x_i|y}_{MAP})}_{\text{Produkt der Attributverteilungen, gegeben Klasse}}$$

- Laufzeit beim Klassifizieren:

$O(m)$ $\quad m = \text{Anzahl Attribute}$

- Laufzeit beim Lernen:

$O(Nm)$ $\quad N = \text{Anzahl Trainingsinstanzen}$

Naive Bayes: Eigenschaften

- Einfach zu implementieren, effizient, populär.
- Funktioniert ok, wenn die Attribute wirklich unabhängig sind.
- Das ist aber häufig nicht der Fall.
- Unabhängigkeitsannahme und modellbasiertes Training führen häufig zu schlechten Ergebnissen.
- Logistische Regression, Winnow, Perzeptron sind meist besser.

Naive Bayes: Beispiel

- Trainingsdaten:

	x_1 : Schufa pos.	x_2: Student	y : Rückzahlung ok?
Instanz $\mathbf{x}_1$	1	1	1
Instanz $\mathbf{x}_2$	1	0	1
Instanz $\mathbf{x}_3$	0	1	0

- Prior: alle Parameter α in den Beta-Verteilungen setzen wir auf $\alpha=2$ (Pseudocounts: $\alpha-1=1$)

- Gelernte Parameter/Hypothese?

Naive Bayes: Beispiel

- Gelernte Parameter/Hypothese
  - Merkmalsverteilungen $P(x_i \mid y)$

x_1	$P(x_1 \mid y=0)$
0	?
1	?

x_1	$P(x_1 \mid y=1)$
0	?
1	?

x_2	$P(x_2 \mid y=0)$
0	?
1	?

x_2	$P(x_2 \mid y=1)$
0	?
1	?

  - Klassenverteilung $P(y)$

y	$P(y)$
0	?
1	?

Naive Bayes: Beispiel

- Gelernte Parameter/Hypothese
  - Merkmalsverteilungen $P(x_i \mid y)$

x_1	$P(x_1 \mid y=0)$
0	2/3
1	1/3

x_1	$P(x_1 \mid y=1)$
0	1/4
1	3/4

x_2	$P(x_2 \mid y=0)$
0	1/3
1	2/3

x_2	$P(x_2 \mid y=1)$
0	2/4
1	2/4

  - Klassenprior $P(y)$

y	$P(y)$
0	2/5
1	3/5

Naive Bayes: Beispiel

- Testanfrage:

$$\mathbf{x} = (\text{Schufa pos} = 0, \text{Student} = 0)$$

- Vorhersage:

$$y_* = \arg\max_y P(y \mid \mathbf{x}, \theta_{MAP}) = \arg\max_y P(y \mid \theta_{MAP}) \prod_{i=1}^{m} P(x_i \mid y, \theta_{MAP})$$

$$P(y=0)P(\mathbf{x} \mid y=0) = P(y=0)P(x_1 = 0 \mid y=0)P(x_2 = 0 \mid y=0)$$

$$= \frac{2}{5} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{45}$$

$$P(y=1)P(\mathbf{x} \mid y=1) = P(y=1)P(x_1 = 0 \mid y=1)P(x_2 = 0 \mid y=1)$$

$$= \frac{3}{5} \cdot \frac{1}{4} \cdot \frac{2}{4} = \frac{3}{40}$$

$$\frac{4}{45} > \frac{3}{40} \Rightarrow y_* = 0$$