Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayessches Lernen

Christoph Sawade/Niels Landwehr
Jules Rasetaharison
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.
- Zufallsvariable X: Abbildung von Elementarereignissen auf numerische Werte.

$$X : \Omega \mapsto \mathbb{R}$$

$$X : \omega \mapsto x$$

Wahrscheinlichkeitstheorie, Zufallsvariablen

- Experiment weist Zufallsvariable (Großbuchstabe) einen Wert (Kleinbuchstabe) 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 $P(X)$“

Diskrete Zufallsvariablen

- Diskrete Zufallsvariablen:

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

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

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

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

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

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

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

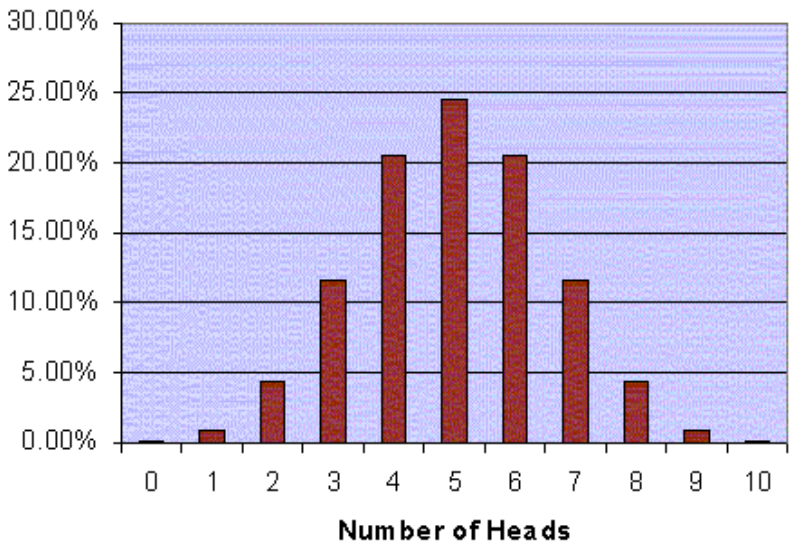
$$\mathrm{Bin}(X \mid N, \mu) = ?$$

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

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

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



Kontinuierliche Zufallsvariablen

- Kontinuierliche Zufallsvariablen
  - Unendlich (meist 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

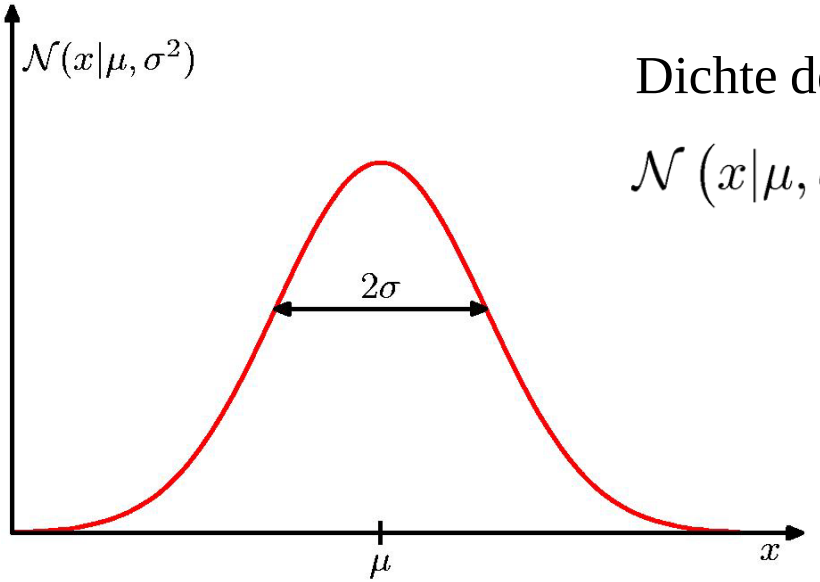
$\forall x : f_X(x) \geq 0, \qquad \int_{-\infty}^{\infty} f_X(x) = 1 \qquad f_X(x) > 1$ möglich

- 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“)
  - $X \sim N(x \mid \mu, \sigma^2)$



Dichte der Normalverteilung

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

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

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

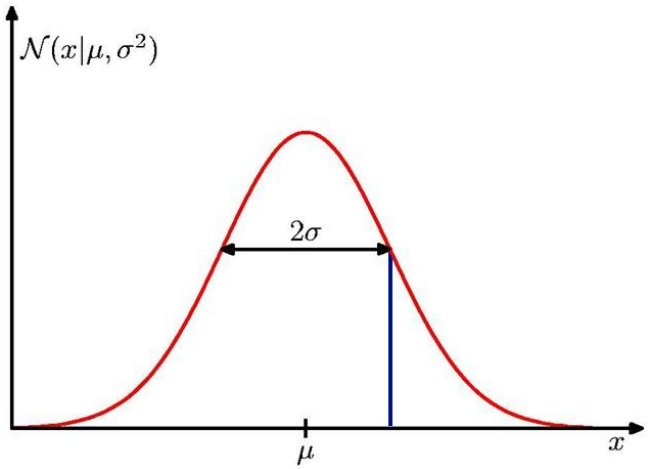
z.B. $\mu = 170, \sigma = 10$

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?



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

Kontinuierliche Zufallsvariablen

- Verteilungsfunktion

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)dx,$$

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

- Dichte ist Ableitung der Verteilungsfunktion

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

- Veranschaulichung Dichte:

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

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:
  - $P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$ 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

$$P(X = x) = \sum_y P(X = x, Y = y)$$ diskret

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

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:

$$E(X) = \sum_x xP(X = x) \qquad X \text{ diskrete ZV}$$

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

- 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

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

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

- 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

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

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

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

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

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

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

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

- Erwartungswert Binomialverteilung

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

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

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

$$= ?$$

Erwartungswert, Varianz Binomialverteilung

- Erwartungswert Bernoulli-Verteilung

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

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

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

- Erwartungswert Binomialverteilung

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

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

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

$$= N\mu$$

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

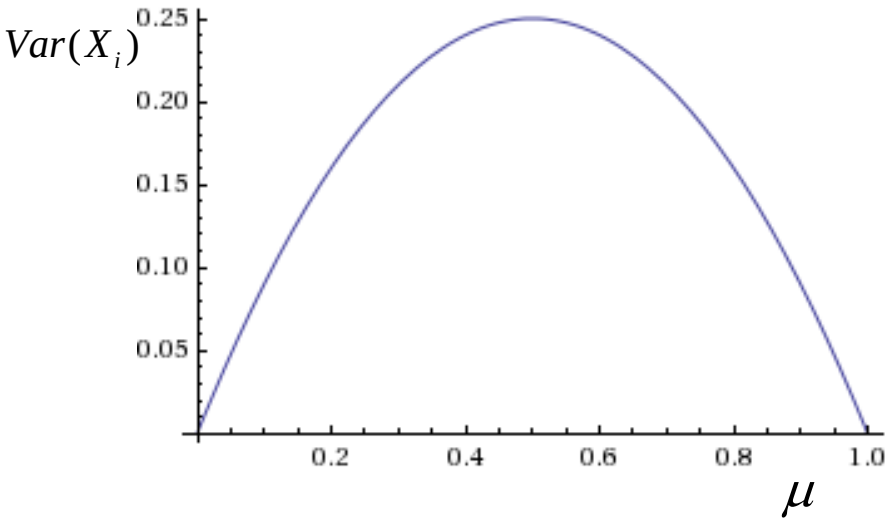
- Varianz Bernoulliverteilung?

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

$$Var(X_i) = ?$$

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

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

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

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

$z = x - \mu$

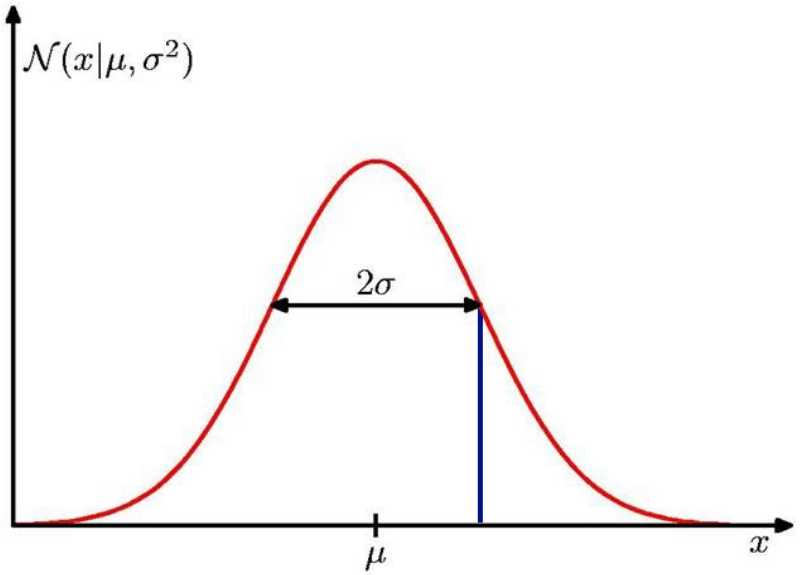
$$= \int_{-\infty}^{\infty} (z+\mu) \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}z^2\right) dz$$

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

Erwartungswert, Varianz Normalverteilung

- Varianz Normalverteilung
  - Man kann zeigen dass

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



Überblick

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

Lernen und Vorhersage

- Bisher: Lernproblemstellung getrennt von Vorhersage
  - Lernen:

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

„Wahrscheinlichstes Modell gegeben die Daten“

  - Vorhersage:

$$\mathbf{x} \mapsto f_{MAP}(\mathbf{x})$$

$\mathbf{x}$ neue Testinstanz

„Vorhersage des MAP Modells“

- 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 festlegen - direkt nach der optimalen Vorhersage zu suchen

Lernen und Vorhersage: Beispiel

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

$$P(f_1 \mid L) = 0.3 \qquad P(f_3 \mid L) = 0.25$$

$$P(f_2 \mid L) = 0.25 \qquad P(f_4 \mid L) = 0.2$$

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

Lernen und Vorhersage: Beispiel

- Modelle f_i probabilistische Klassifikatoren:
  - binäre Klassifikation: $P(y=1 \mid \mathbf{x}, f_i) \in [0,1]$

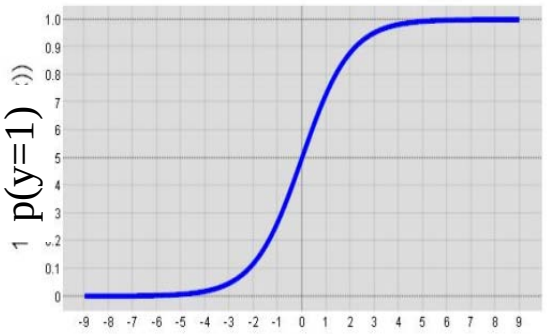
- Z.B lineares Modell:

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

$\mathbf{w}$ Parametervektor

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

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



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 $f_1 : y = 1$
- Andererseits (Rechenregeln der Wsk!):

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

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

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

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

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 des Bayesschen Lernens/Vorhersage!

Bayessches Lernen und Vorhersage

- Problemstellung Vorhersage
- Gegeben:
  - Trainingsdaten L,
  - neue Testinstanz x.
- Gesucht:
  - Verteilung über Werte y für gegebenes x.
  - $P(y \mid x, L)$
- Bayessche Vorhersage: wahrscheinlichstes y.
  - $y_* = \arg\max_y P(y \mid x, L)$
  - Minimiert Risiko einer falschen Vorhersage.
  - Heißt auch Bayes-optimale Entscheidung oder Bayes-Hypothese.

Bayessches Lernen und Vorhersage

- Berechnung Bayessche Vorhersage
  - $y_* = \arg\max_y P(y \mid x, L)$

Summenregel

$= \arg\max_y \int P(y, \theta \mid x, L) d\theta$

θ Modell

Produktregel

$= \arg\max_y \int P(y \mid \theta, x) P(\theta \mid L) d\theta$

Bayesian Model Averaging

Vorhersage, gegeben Modell

Modell gegeben Trainingsdaten

- Bayessches Lernen:
  - Mitteln der Vorhersage über alle Modelle.
  - Gewichtung: wie gut passt Modell zu Trainingsdaten.

Bayessches Lernen und Vorhersage

- Bayessche Vorhersage praktikabel?
  - $$y_* = \arg\max_y P(y \mid x, L)$$
    $$= \arg\max_y \int P(y \mid x, \theta) P(\theta \mid L) d\theta$$
  - 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?

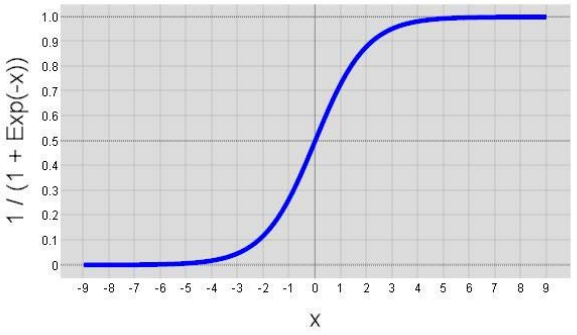
$$y_* = \arg\max_y P(y \mid x, L)$$
$$= \arg\max_y \int P(y \mid x, \theta) P(\theta \mid L) d\theta$$

- Wir brauchen:
  - 1) Wsk für Klassenlabel gegeben Modell, $P(y \mid x, \theta)$

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?

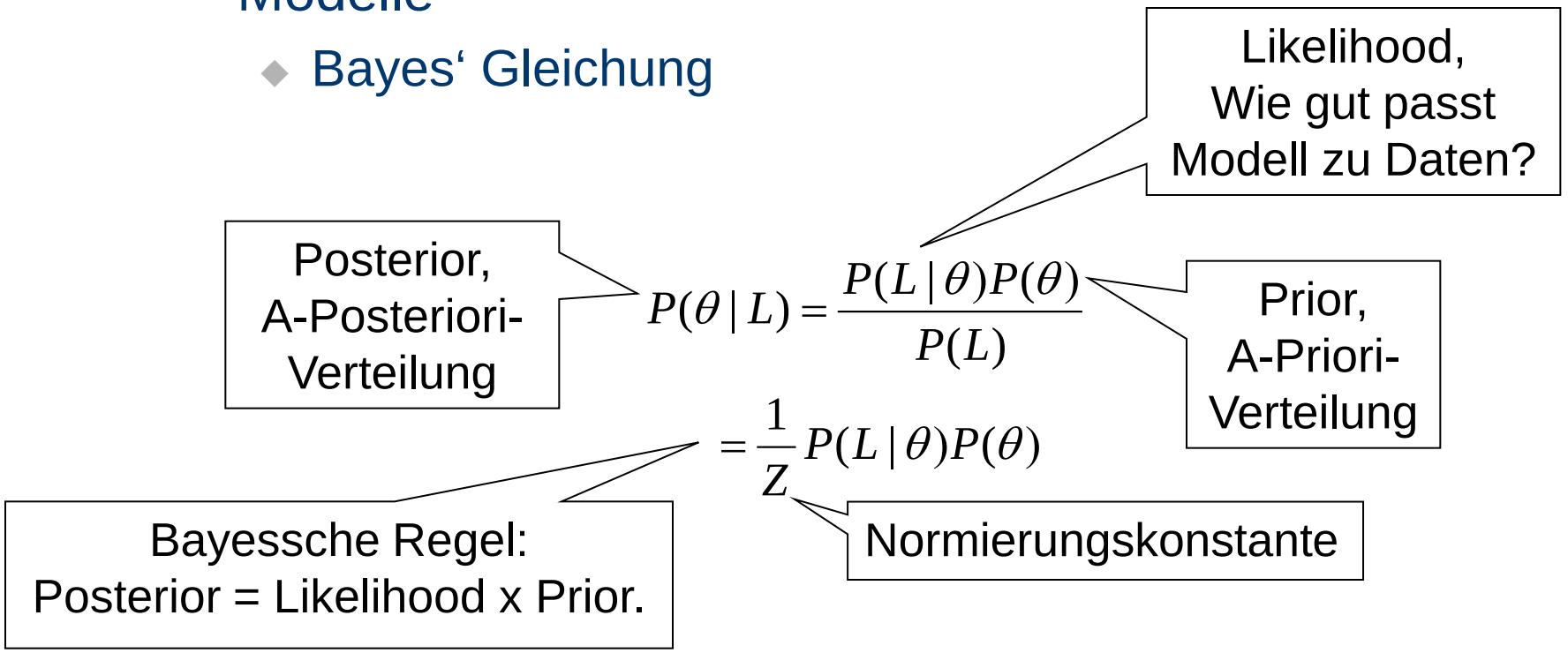
$$y_* = \arg\max_y P(y \mid x, L)$$
$$= \arg\max_y \int P(y \mid x, \theta) P(\theta \mid L) d\theta$$

- 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



Bayessche Regel

- Bayes‘ Gleichung

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

- 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:

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

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

Wahrscheinlichkeit des in L beobachteten Klassenlabels gegeben Modell θ

Bayessche Regel

- Bayes‘ Gleichung

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

- 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

- Bayes‘ Gleichung

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

- 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:
  - $|\mathbf{w}|^{\wedge 2}$ möglichst niedrig ($\mathbf{w} = \theta$)

Bayessche Regel

- Bayes‘ Gleichung

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

- 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

- Bayes‘ Gleichung

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

- 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/MAP/ML-Hypothese

- Um Risiko einer Fehlentscheidung zu minimieren: wähle Bayessche Vorhersage

$$y_* = \arg\max_y \int P(y \mid x, \theta) P(\theta \mid L) d\theta$$

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

$$\theta_* = \arg\max_\theta P(\theta \mid L)$$
$$y_* = \arg\max_y P(y \mid x, \theta_*)$$

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

Zusammenfassung Bayessche/MAP/ML-Hypothese

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