Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Mathematische Grundlagen (Bayes‘sches Lernen)

Tobias Scheffer
Michael Großhans
Paul Prasse

Anwendungsbeispiel 1: Diagnostik

- Neuer Test wurde entwickelt
- Frage: Wie sicher ist die Person krank, wenn positives Testergebnis vorliegt?
- Studie: Auf kranken und gesunden Testpersonen (Zustand ist bekannt) wird Test angewandt

Anwendungsbeispiel 2: Impfstoff

- Neuer Impfstoff wurde entwickelt
- Frage: Wie wirksam ist er? Wie oft verhindert er eine Infektion?
- Studie: Testpersonen werden geimpft; später wird untersucht, ob sie sich angesteckt haben

Was untersucht man?

- *Deskriptive Statistik*: Beschreibung, Untersuchung von Eigenschaften von Stichproben (langweilig).
  - Welcher Anteil der Testpersonen ist gesund geblieben? (= abzählen)

- *Induktive Statistik*: Welche Schlussfolgerungen über die Grundgesamtheit lassen sich aus Stichproben ziehen? (spannend, maschinelles Lernen).
  - Wie viele Personen werden in Zukunft gesund bleiben?
  - Wie sicher sind wir uns dessen?

Wahrscheinlichkeiten

- Frequentistische „objektive“ Wahrscheinlichkeiten
  - Wahrscheinlichkeit als relative Häufigkeit mit der ein Ereignis bei einer großen Zahl unabhängiger und wiederholter Experimente eintritt.
- Bayes‘sche, „subjektive“ Wahrscheinlichkeiten
  - Wahrscheinlichkeit als persönliche Überzeugung, dass ein Ereignis eintritt.
  - Unsicherheit bedeutet hier Mangel an Information.
    - Wie wahrscheinlich ist es, dass der Impfstoff wirkt?
    - Neue Informationen (z.B. Studienergebnisse) können diese subjektive Wahrscheinlichkeiten verändern.

Wahrscheinlichkeitstheorie

- *Zufallsexperiment*: Definierter Prozess in dem eine Beobachtung ω erzeugt wird (Elementarereignis).
- *Ereignisraum* Ω: Menge aller möglichen Elementarereignisse; Anzahl aller Elementarereignisse ist |Ω|.
- *Ereignis* *A*: Teilmenge des Ereignisraums.
- *Wahrscheinlichkeit* *P*: Funktion welche Wahrscheinlichkeitsmasse auf Ereignisse *A* aus Ω verteilt.

$$P(A) := P\left(\{\omega \in A\}\right)$$

Wahrscheinlichkeitstheorie

- Wahrscheinlichkeit = *normiertes Maß*
- definiert durch Kolmogorow-Axiome:
  - Wahrscheinlichkeit von Ereignis $A \subseteq \Omega$: $0 \leq P(A) \leq 1$
  - Sicheres Ereignis: $P(\Omega) = 1$
  - Wahrscheinlichkeit dass Ereignis $A \subseteq \Omega$ <u>oder</u> Ereignis $B \subseteq \Omega$ eintritt mit $A \cap B = \emptyset$ (beide Ereignisse sind *inkompatibel*): $P(A \cup B) = P(A) + P(B)$
  - Allgemein gilt: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Zufallsvariablen

- *Zufallsvariable X* ist Abbildung eines elementaren Ereignisses auf numerischen Wert $X : \omega \in \Omega \mapsto x \in \mathbb{R}$ bzw. auf *m*-dimensionalen Vektor $X : \omega \in \Omega \mapsto \mathbf{x} \in \mathbb{R}^m$
  - Maschinelles Lernen: auch Abbildungen auf Bäume und andere Strukturen möglich.
  - Maschinelles Lernen: gleichgesetzt mit Ereignisraum.

- Bild der Zufallsvariable: $\mathcal{X} := \{ X(\omega) \mid \omega \in \Omega \}$

Diskrete Zufallsvariablen

- X nennt man eine diskrete Zufallsvariable, wenn sie nur diskrete Werte annehmen kann.

- Wahrscheinlichkeitsfunktion *P* weist jedem möglichen Wert einer Zufallsvariable eine Wahrscheinlichkeit $P(X = x) \in [0;1]$ zu.
  - Summe der Verteilungsfunktion über alle Werte

$$\sum_{x \in \mathcal{X}} P(X = x) = 1$$

Stetige Zufallsvariablen

- X nennt man eine stetige Zufallsvariable, wenn sie sie kontinuierliche Werte annehmen kann.
- Die Werte der Verteilungsfunktion *P* entsprechen an jeder Stelle den kumulierten Wahrscheinlichkeiten $P_X(x) = P(X \leq x) \in [0;1]$
- Die Werte der Dichtefunktion *p* entsprechen an jeder Stelle den Änderungen der Verteilungsfunktion

$$p_X(a) = \left.\frac{\partial P_X(x)}{\partial x}\right|_{x=a} \quad \text{mit} \quad \int_{-\infty}^{\infty} p_X(x)\,dx = 1$$

Feinheiten der Notation

- $P(X)$
  p_X
  Wahrscheinlichkeitsfunktion bzw. Dichtefunktion über alle möglichen Werte von X

- $P(X = x)$
  $p_X(x)$
  konkreter Wahrscheinlichkeitswert bzw. konkreter Wert der Dichtefunktion

- $P(x)$
  $p(x)$
  verkürzte Schreibweise von $P(X = x)$ bzw. von $p_X(x)$ wenn eindeutig ist, welche Zufallsvariable gemeint ist. Meist werden Wahrscheinlichkeitsfunktion und Dichtefunktion nicht getrennt.

Erwartungswert

- Der Erwartungswert $E(X)$ ist der gewichtete Mittelwert der möglichen Werte von X
  - diskrete Zufallsvariable:
    $$E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$$
  - kontinuierliche Zufallsvariable:
    $$E(X) = \int_{\mathcal{X}} xp_X(x)\,dx$$
- Die Varianz $Var(X)$ ist der erwarte quadratische Abstand zum Erwartungswert von X

$$Var(X) = E\left[\left(X - E(X)\right)^2\right]$$

Erwartungswert: Beispiel

- St. Petersburger Spiel:
  - Werfen einer Münze, bis sie zum ersten Mal „Kopf“ zeigt
  - passiert dies gleich beim ersten Wurf, gewinnt man 1 Euro
  - falls nicht, verdoppelt sich der Gewinn so oft man „Zahl“ geworfen hat
  - der Gewinn den man am Ende erhält ist Zufallsvariable X
  - Erwarteter (durchschnittlicher) Gewinn:

$$E(X) = \sum_{x \in \mathcal{X}} x P(X = x)$$

$$= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \ldots = \infty$$

Gemeinsame Wahrscheinlichkeit

- $P(X_1, X_2)$ ist die gemeinsame Wahrscheinlichkeitsverteilung der Zufallsvariablen X_1 und X_2
- gemeinsamer Wertebereich:

kartesisches Produkt

$$\mathcal{X}_1 \times \mathcal{X}_2$$

- z.B.:

$\mathcal{X}_1 \times \mathcal{X}_2 = \{$(infiziert,infiziert), (infiziert, gesund), (gesund, infiziert), (gesund, gesund) $\}$

Bedingte Wahrscheinlichkeiten

- Bedingte Wahrscheinlichkeit: Wahrscheinlichkeit eines der möglichen Werte von X mit Zusatzinformation:
  - Diskrete Zufallsvariable:

$$P\left(X = x \mid \text{zusätzliche Information}\right)$$

  - Stetige Zufallsvariable:

$$p_X\left(x \mid \text{zusätzliche Information}\right)$$

Abhängige Zufallsvariablen

- Zufallsvariablen X_1 und X_2 können abhängig oder unabhängig sein
- Unabhängig: $P(X_1, X_2) = P(X_1)\, P(X_2)$
  - Beispiel:
    - 2 aufeinanderfolgende Münzwürfe
    - Ergebnis des zweiten hängt nicht vom ersten ab
  - Impliziert: $P(X_2 \mid X_1) = P(X_2)$
- Abhängig: $P(X_1, X_2) = P(X_1)\, P(X_2 \mid X_1) \neq P(X_1)\, P(X_2)$
  - Beispiel:
    - Grippeinfektionen von 2 Sitznachbarn

Bedingte Unabhängigkeit

- Zufallsvariablen können abhängig sein, jedoch unabhängig gegeben eine weitere Zufallsvariable
- Die Zufallsvariablenvariablen X_1 und X_2 heißen bedingt unabhängig gegeben Y wenn gilt:
  - $P(X_1,X_2|Y) = P(X_1|Y)\,P(X_2|Y)$
- Beispiel:
  - Wirkrate des Impfstoffs bekannt → Infektionswahrscheinlichkeiten der Personen unabhängig
  - Wirkrate des Impfstoffs unbekannt → Beobachtung eines Teils der Testpersonen gibt Information über restliche Testpersonen

Rechenregeln

- Produktregel:
  $$P(X,Y) = P(X)P(Y|X)$$
  - verallgemeinert:
    $$P(X_1, X_2, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_1, \ldots, X_{i-1})$$
- Summenregel:
  - Sind Ereignisse A, B inkombatibel:
    $$P(A \cup B) = P(A) + P(B)$$
- Randverteilung:
  $$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y) = \sum_{y \in \mathcal{Y}} P(X | Y = y) P(Y = y)$$

Rechenregeln

- Satz von Bayes:
  - Inferiere $P(X \mid Y)$ aus $P(Y \mid X), P(X)$ und $P(Y)$

$$P(X,Y) = P(Y,X)$$

$$\Leftrightarrow P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

$$\Leftrightarrow P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Satz von Bayes: Beispiel

- Diagnostik-Beispiel:
  - $P(positiv \mid krank) = 0{,}98$
  - $P(positiv \mid gesund) = 0{,}05$
  - $P(krank) = 0{,}02$
- Gesucht für Testergebnis $Test$:
  - Wahrscheinlichkeit, dass der Patient krank ist:

    $$P\left(krank \mid Test\right)$$

  - Plausibelste Ursache

    $$\arg\max_{P \in \{krank, gesund\}} P\left(Test \mid P\right)$$

  - Wahrscheinlichste Ursache

    $$\arg\max_{P \in \{krank, gesund\}} P\left(P \mid Test\right)$$

Satz von Bayes

- Wahrscheinlichkeit der Ursache *Urs.* für eine Beobachtung *Beob.*:

$$P(\text{Urs.} \mid \text{Beob.}) = P(\text{Beob.} \mid \text{Urs.}) \frac{P(\text{Urs.})}{P(\text{Beob.})}$$

$$P(\text{Beob.}) = \sum_{u \in Ursachen} P(\text{Beob.} \mid u) P(u)$$

- $P(\text{Urs.})$: A-Priori-Wahrscheinlichkeit, „Prior“.
- $P(\text{Beob.}|\text{Urs.})$: Likelihood.
- $P(\text{Urs.}|\text{Beob.})$: A-Posteriori-Wahrscheinlichkeit, „Posterior“.

Prior, Likelihood und Posterior

- Subjektive Einschätzung, **bevor** man die Daten gesehen hat (a priori): Prior-Verteilung über die Modelle
  - $P(Krankheit)$
  - $P(\theta), \quad \theta -$ Wirksamkeit des Impfstoffes
- Wie gut passen die Daten zum Modell: Likelihood
  - $P(Test \mid Krankheit)$
  - $P(Studie \mid \theta),$
- Subjektive Einschätzung, **nachdem** man die Daten gesehen hat (a posteriori): Posterior-Verteilung
  - $P(Krankheit \mid Test)$
  - $P(\theta \mid Studie)$

Prior

- Woher bekommt man die Prior-Verteilung?
  - $P(Krankheit)$ relativ naheliegend; diskret
  - $P(\theta)$: schwieriger; stetig; z.B. aus allen bisherigen Studien anderer Impfstoffe schätzen
- Es gibt keine „richtige“ Prior-Verteilung!
  - aber: unterschiedliche Prior-Verteilungen ermöglichen unterschiedlich gute Vorhersagen für die Zukunft
- Posterior-Verteilung ergibt sich deterministisch aus Prior und Likelihood der Beobachtungen
  - durch Satz von Bayes

Beispiel Likelihood: Bernoulli-Verteilung

- Eine diskrete Verteilung mit den 2 möglichen Ereignissen 0 und 1 ist eine Bernoulli-Verteilung
- bestimmt durch genau einen Parameter:

$$\theta \in [0; 1]$$

- Verteilungsfunktion:

$$P(X = 1|\theta) = \theta$$
$$P(X = 0|\theta) = 1 - \theta$$

Beispiel Likelihood: Binomialverteilung

- Zusammenfassung mehrerer Bernoulli-verteilter Zufallsvariablen $X_1, \ldots, X_n$ mit gleichem Parameter θ
  - neue Zufallsvariable Y, die angibt, wie viele der X_i positiv sind: $Y = \sum_{i=1}^{n} X_i$
  - Y ist Binomial-verteilt mit Parametern θ und n
  - Verteilungsfunktion:

$$P(Y = y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Binomialkoeffizient: Anzahl der Möglichkeiten, aus n Elementen y auszuwählen

Wahrscheinlichkeit, dass y der X_i positiv sind

Wahrscheinlichkeit, dass n-y der X_i negativ sind

Beispiel Prior: Beta-Verteilung

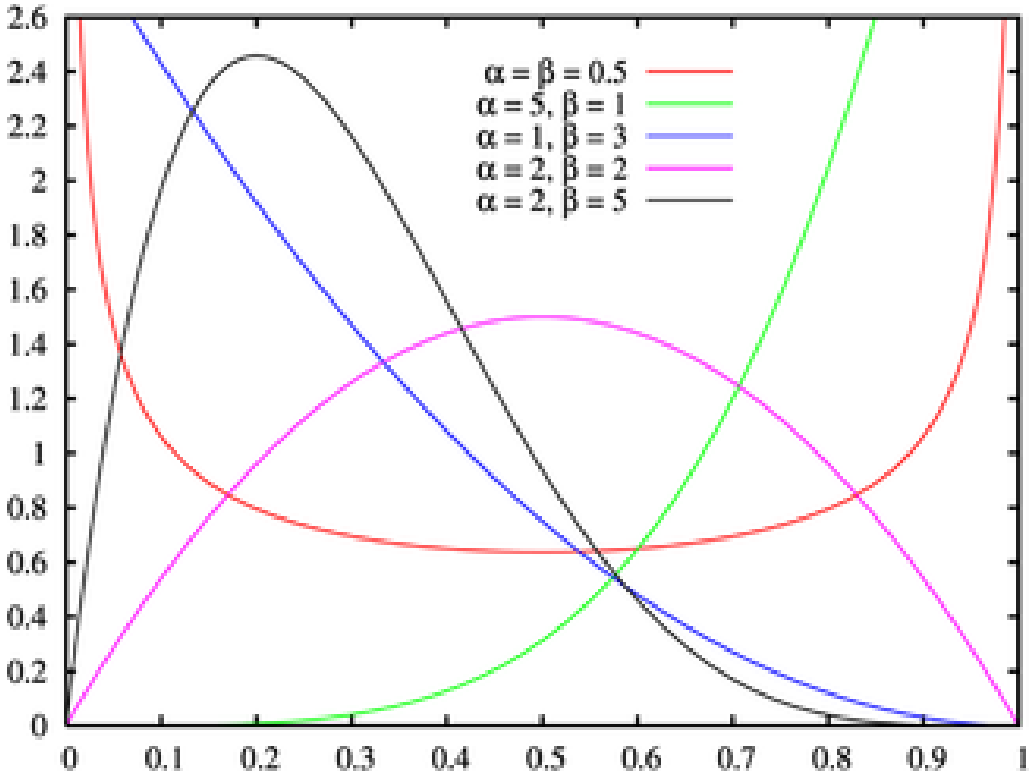
- Verteilung über alle Wirkraten
- keine diskrete, sondern kontinuierliche Verteilung
- $P(\theta)$ beschreibt eine Dichtefunktion

- Häufige Wahl (bei Parameterraum $\theta \in [0; 1]$):
  - Beta-Verteilung
  - definiert durch 2 Parameter α und β

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Beta-Funktion; dient der Normalisierung

Beispiel Prior: Beta-Verteilung



- Spezialfall: $\alpha = \beta = 1$ ist Gleichverteilung

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = \frac{\theta^0(1-\theta)^0}{1} = 1$$

Schema für Ermittlung der Posterior-Verteilung

- Gegeben:
  - Prior-Verteilung $P(\theta)$
  - Beobachtungen $x_1, \ldots, x_n$
  - Likelihood $P(x_1, \ldots, x_n \mid \theta)$
- Gesucht: Posterior-Verteilung $P(\theta \mid x_1, \ldots, x_n)$
- 1. Satz von Bayes anwenden

$$P(\theta|x_1, \ldots, x_n) = P(x_1, \ldots, x_n|\theta)P(\theta)/P(x_1, \ldots, x_n)$$

- 2. Randverteilung für kontinuierliche Parameter einsetzen

$$P(x_1, \ldots, x_n) = \int P(x_1, \ldots, x_n|\theta)P(\theta)d\theta$$

Ermittlung der Posterior-Verteilung: Beispiel

- Gegeben:
  - Modellparameterraum $\theta \in [0;1]$
  - Beta-Prior mit Parametern α und β: $P(\theta)=Beta(\theta \mid \alpha,\beta)$
  - Bernoulli-Likelihood
  - binäre Beobachtungen $x_1,\ldots,x_n$, bedingt unabhängig gegeben Modellparameter θ
    - a positive Beobachtungen, b negative
- Gesucht:
  - Posterior $P(\theta \mid x_1,\ldots,x_n)$

Ermittlung der Posterior-Verteilung

$$P(\theta \mid x_1,\ldots,x_n)$$

$$= P(x_1,\ldots,x_n \mid \theta)\,P(\theta) / P(x_1,\ldots,x_n)$$

Satz von Bayes

$$= \left[\prod_{i=1}^{n} P(x_i \mid \theta)\right] P(\theta) / P(x_1,\ldots,x_n)$$

Bedingte Unabhängigkeit

$$= P(X=1 \mid \theta)^a\, P(X=0 \mid \theta)^b\, P(\theta) / P(x_1,\ldots,x_n)$$

a positive, b negative

$$= \theta^a (1-\theta)^b \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} / P(x_1,\ldots,x_n)$$

Bernoulli- und
Beta-Verteilung einsetzen

$$= \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} / \left[\int \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} d\theta\right]$$

Terme zusammenfassen,
Randverteilungsformel

$$= \frac{\theta^{a+\alpha-1}(1-\theta)^{b+\beta-1}}{B(\alpha,\beta)} / \left[\frac{B(a+\alpha,b+\beta)}{B(\alpha,\beta)}\right]$$

Definition der
Beta-Funktion

$$= Beta(\theta \mid a+\alpha, b+\beta)$$

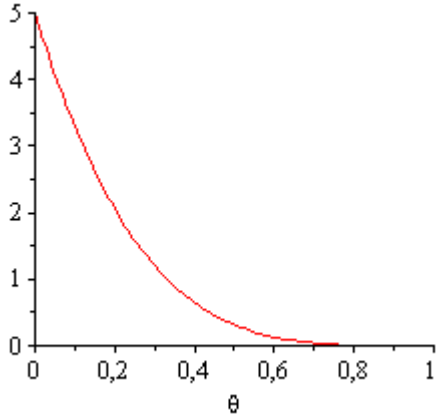
Kürzen, Definition
der Beta-Verteilung

Konjugierter Prior

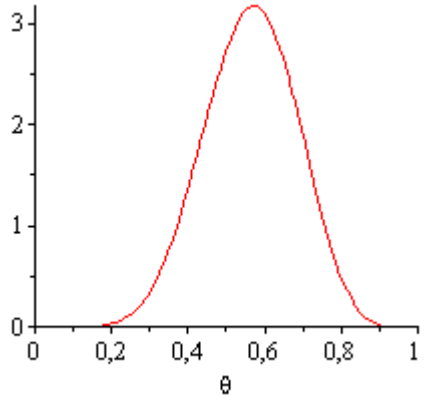
- Im vorherigen Beispiel:
  - Übergang vom Prior $Beta(\theta \mid \alpha, \beta)$
  - durch a positive und b negative Beobachtungen
  - zum Posterior $Beta(\theta \mid \alpha + a, \beta + b)$
  - algebraische Form von Posterior und Prior identisch
- Die Beta-Verteilung ist der konjugierte Prior zur Bernoulli-Likelihood
- Immer vorteilhaft, den konjugierten Prior zu verwenden, um zu garantieren, dass der Posterior effizient berechenbar ist

Rechenbeispiel: Impfstudie

- Prior: Beta mit α=1, β=5



- 8 gesunde Testpersonen, 2 infizierte
- ergibt Posterior: Beta mit α=9, β=7



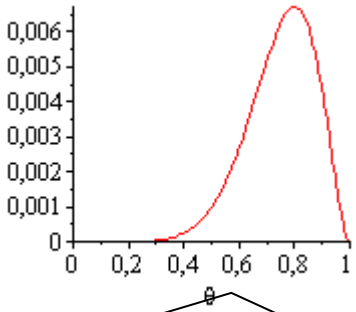
Parameterschätzung

- Bayes'sche Inferenz liefert keinen Modellparameter, sondern Verteilung über Modellparameter
- Ermittlung des Modells mit der höchsten Wahrscheinlichkeit: MAP-Schätzung
  - „maximum-a-posteriori" = maximiert den Posterior
  - $\theta_{MAP} = \text{argmax}_\theta P(\theta \mid Beobachtungen)$
- Im Gegensatz dazu: *plausibelstes* Modell = ML-Schätzung
  - „maximum-likelihood" = maximiert die Likelihood
  - ohne Berücksichtigung des Priors
  - $\theta_{ML} = \text{argmax}_\theta P(Beobachtungen \mid \theta)$

Parameterschätzung: Beispiel

- Impfstudie:
  - Prior: Beta mit α=1, β=5
  - 8 gesunde Testpersonen, 2 infizierte
  - ergibt Posterior: Beta mit α=9, β=7
- ML-Schätzung:
  - $\theta_{ML} = \text{argmax}_\theta P(Beob. \mid \theta)$
  - $\theta_{ML} = \arg\max_\theta \theta^8 (1-\theta)^2 = \frac{4}{5}$
- MAP-Schätzung:
  - $\theta_{MAP} = \text{argmax}_\theta P(\theta \mid Beob.)$

$$\theta_{MAP} = \arg\max_\theta \frac{\theta^8 (1-\theta)^6}{B(9,7)} = \frac{4}{7}$$



Likelihood-Funktion
(keine Wahrscheinlichkeitsverteilung)

Vorhersage

- Welche Beobachtungen kann man in Zukunft erwarten, gegeben die Beobachtungen der Vergangenheit?
  - Vorhersage für Testdaten, gegeben eine Menge von Trainingsdaten $P(X_{neu} \mid X_{alt})$
- Vorhersage mit MAP-Schätzung:
  - erst θ_{MAP} bestimmen durch $\theta_{MAP} = argmax_{\theta} P(\theta \mid X_{alt})$
  - dann $P(X_{neu} \mid \theta_{MAP})$ bestimmen (Likelihood-Verteilung)
  - Mit Informationsverlust verbunden:
    - θ_{MAP} nicht „echter“ Parameter, sondern wahrscheinlichster
    - ignoriert, dass auch andere Modelle in Frage kommen

Bayes-optimale Vorhersage

- Kein Zwischenschritt über das MAP-Modell, sondern direkte Herleitung der Vorhersage:

$$P\left(X_{neu} \mid X_{alt}\right)$$

1. Randverteilung $$= \int_{\theta} P\left(X_{neu} \mid \theta, X_{alt}\right) P\left(\theta \mid X_{alt}\right) d\theta$$

2. bedingte Unabhängigkeit $$= \int_{\theta} P\left(X_{neu} \mid \theta\right) P\left(\theta \mid X_{alt}\right) d\theta$$

mitteln über *alle* Modelle (Bayesian Model-Averaging)

gewichtet durch: wie gut passt das Modell zu den früheren Beobachtungen? (Posterior)

Vorhersage gegeben Modell

Vorhersage: Beispiel

- Impfstudie: Mit welcher Wahrscheinlichkeit bleibt neue Person gesund, gegeben die Studie?
- Vorhersage mit MAP-Modell:
  - $\theta_{MAP} = \text{argmax}_\theta P(\theta \mid Beob.) = 4/7$
  - $P(gesund \mid \theta_{MAP}) = \theta_{MAP} = 4/7$
- Bayes-optimale Vorhersage:

$$P\left(gesund \mid X_{alt}\right) = \int_\theta P\left(gesund \mid \theta\right) P\left(\theta \mid X_{alt}\right) d\theta$$

$$= \boxed{\int_\theta \theta \cdot Beta\left(\theta \mid 9,7\right) d\theta} = \frac{9}{16}$$

Erwartungswert einer Beta-Verteilung

Rekapitulation

- Bayes‘sches Lernen:
  - subjektiver Prior: Ausgangsverteilung über die Modelle
  - Beobachtungen aus der Vergangenheit: Likelihood gegeben Modellparameter
  - ergibt durch Satz von Bayes Posterior: Verteilung über Modelle gegeben die Beobachtungen
  - Mögliche Wege für Vorhersagen in der Zukunft:

einfacher →
- MAP-Modell berechnen (Maximierung des Posteriors), dann Vorhersage mit MAP-Modell

besser →
- Bayes-optimale Vorhersage: über alle Modelle mitteln, gewichtet mit ihrer Posterior-Wahrscheinlichkeit

Fragen?