# Maschinelles Lernen und Datenanalyse

*In der Mess- und Prüftechnik* PD Stefan Bosse

Universität Bremen - FB Mathematik und Informatik

# Probalistisches Lernen

# Wahrscheinlichkeiten und Bayes Regel

- In einem probalistischen Ansatz sind die Dateninstanzen gemessene Ereignesse oder Beobachtungen. [Witten, DMPMLTaT, pp. 335]
  - Die Dateninstanzen in $\boldsymbol{D}$ bilden Zufallsvariablen ab!

Es sei $A$ eine Zufallsvariable mit diskreten Werten $\{a_i\}$. Dann ist $P(A)$ oder kurz $P(a)$ die Wahrscheinlichkeitsfunktion für das Auftreten eines $a_i \in A$!

Es sei $x$ eine Zufallsvariable mit kontinuierlichen Werten $[v_0,v_1]$. Dann ist $p(x)$ die Wahrscheinlichkeitsverteilung der Werte $x \in [v_0,v_1]$.

Dann ist $p(x{=}x_\mathrm{i})$ die Wahrscheinlichkeit des Auftretens des Wertes $x_\mathrm{i}$ von $x$.

- Besondere Rolle nehmen binäre Ereignisse ein (also $A{=}\{0,1\}$). Etwas tritt ein oder ist wahr oder nicht.

Wenn $A$ und $B$ diskrete Zufallsvariablen sind, dann kann man über eine Produktregel die gemeinsame (vereinte) Wahrscheinlichkeit für das Auftreten von $A$ und $B$ bestimmen:

$$P(A,B) = P(A|B)P(B)$$

Die gemeinsame Wahrscheinlichkeit ist ein statistisches Maß, das die Wahrscheinlichkeit berechnet, dass zwei Ereignisse zusammen und gleichzeitig auftreten. Gemeinsame Wahrscheinlichkeit ist die Wahrscheinlichkeit, dass Ereignis $B$ gleichzeitig mit Ereignis $A$ Auftritt.

- $P(A)$ ist die Wahrscheinlichkeit eines Ereignisses $A$
- $P(A|B)$ ist die bedingte Wahrscheinlichkeit von $A$ mit dem bedingten Ereignis $B$ (d.h. $B$ muss eintreten damit auch $A$ eintritt): $B \rightarrow A$
- $P(B|A)$ ist dann das "inverse Problem": $A \rightarrow B$

## Bayes Regel (Umkehr der Schlussfolgerung)

- Interessant wäre es zu wissen wenn $P(A|B)$ bekannt wie es mit $P(B|A)$ aussieht:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

# Binäre Klassifikation

Hat man ein Experiment *n*-mal wieder holt (*n* Fälle oder Dateninstanzen in der Datentabelle), gibt es vier verschiedene Zähler bei einer binären Klassifiation mit $y$: Inferenz durch Modell/Vorhersage , $y_0$: Vorgabe/Wirklicher Wert:

- Ergebnis: {True, False}
- Zähler:
  - True Positiive (**TP**): Die Anzahl der Fälle, in denen $y$=$y_0$=True ist.
  - False Positive (**FP**): Die Anzhal der Fälle, in denen $y_0$=False und $y$=True ist.
  - True Negative (**TN**): Die Anzahl der Fälle, in denen $y$=$y_0$=False ist.
  - False Negative (**FN**): Die Anzhal der Fälle, in denen $y_0$=True und $y$=False ist.

# Binäre Klassifikation

## Statitische Parameter

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$f_1 = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

# Ein Beispiel: Der Mythos des Infektionstests

- Es gibt einen Test auf eine Virusinfektion
  - Dieser wird im Labor getestet: 100 Proben Klasse negativ (kein Virus, $\neg V$), 100 Proben Klasse positiv (mit Virus, $V$)
  - Der Test zeigt $T$ (positiv) an, ansonsten $\neg T$ (negativ)
  - Die Analyse der Testexperimente zeigt: TP=99, FP=2, FN=1, TN=98

Sensitivität

$$P(T|V) = \frac{TP}{TP+FN} = \frac{99}{99+1} = 0.99 (CV19: 0.5, 0.7-0.9, \text{Good20})$$

Spezifizität

$$P(\neg T|\neg V) = \frac{TN}{TN+FP} = \frac{98}{98+2} = 0.98 (CV19: 0.99, 0.999, \text{Good20})$$

Genauigkeit

$$Auracy = \frac{TP + TN}{N} = \frac{99 + 98}{200} = 0.985$$

Präzision

$$Precision = \frac{TP}{TP + FP} = \frac{99}{98 + 2} = 0.99$$

- Es gibt eine Vorbedingung (Vorwahrsch.) bei einer Testanwendung: Der Wahrscheinlichkeit einer Infektion *P*(V) wenn eine Stichprobe gemacht wird (also *n*=1). Diese wird mit *P*(V)=0.001 angenommen.

## Anwendung der Bayeseschen Regel

$$P(V|T) = \frac{P(T|V)P(V)}{P(T)},$$
$$P(T) = P(T,V) + P(T,\neg V) =$$
$$P(T|V)P(V) + (1 - P(\neg T|\neg V))(1 - P(V))$$

- Bei P(V)=0.001 (zufällige Stichprobe ohne Anlass und Differentialdiagnose) ergibt sich:

$$P(V|T) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + (1 - \mathbf{0.98})(1 - 0.001)} = 0.047 \approx P(V)$$
$$= \frac{0.70 \times 0.001}{0.70 \times 0.001 + (1 - \mathbf{0.999})(1 - 0.001)} = 0.41$$
$$= \frac{0.50 \times 0.001}{0.50 \times 0.001 + (1 - \mathbf{0.99})(1 - 0.001)} = 0.047$$

$$P(\neg V|\neg T) = \frac{P(\neg T|\neg V)P(\neg V)}{P(\neg T)},$$

$$P(\neg T) = 1 - P(T) =$$

$$P(\neg T|\neg V)P(\neg V) + (1 - P(T|V))P(V)$$

- Bei P(V)=0.001 (zufällige Stichprobe ohne Anlass und Differentialdiagnose) ergibt sich:

$$P(\neg V|\neg T) = \frac{0.98 \times 0.999}{0.98 \times 0.999 + (1 - 0.99)0.001} = 0.999$$

# Naiver Bayes Klassifikator

- Zurück zum Golfspielproblem!

[Witten]

| Outlook | | | Temperature | | | Humidity | | | Windy | | | Play | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | | *Yes* | *No* | *Yes* | *No* |
| Sunny | 2 | 3 | Hot | 2 | 2 | High | 3 | 4 | False | 6 | 2 | 9 | 5 |
| Overcast | 4 | 0 | Mild | 4 | 2 | Normal | 6 | 1 | True | 3 | 3 | | |
| Rainy | 3 | 2 | Cool | 3 | 1 | | | | | | | | |
| Sunny | 2/9 | 3/5 | Hot | 2/9 | 2/5 | High | 3/9 | 4/5 | False | 6/9 | 2/5 | 9/14 | 5/14 |
| Overcast | 4/9 | 0/5 | Mild | 4/9 | 2/5 | Normal | 6/9 | 1/5 | True | 3/9 | 3/5 | | |
| Rainy | 3/9 | 2/5 | Cool | 3/9 | 1/5 | | | | | | | | |

Abb. 1. Die Ein- und Ausgabevariablen mit bedingter Verteilung (also Y|X)

Annahme: Alle Eingabevariblen $\boldsymbol{x}$={Outlook, Temperature, Humidity, Windy} sind unabhängig.

- Wenn es nun eine Messung $E=\boldsymbol{x}$ gibt (Evidenz) mit einer Hypothese vom Ergebnis $H=y$ mit $C_y=\{yes|no\}$, dann gilt quasi als Training (wir kennen ja den Zusammenhang $E \leftrightarrow H$): $P(E|H)$
- Es wird angenommen dass alle Variablen gleichwertig sind → unbekannte Modellannahme oder a-priori Wissen!

Nun gibt es ein weiteres unbekanntes Beispiel:

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Cool | High | True | ? |

$$P(y=yes) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$$
$$P(y=no) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.026$$

- Es gilt also nach der Bayes Regel:

$$P(C=yes|E) = \frac{\Pi_i P(E_i|C=yes) \times P(C=yes)}{P(E)}$$

$$P(C=no|E) = \frac{\Pi_i P(E_i|C=no) \times P(C=no)}{P(E)}$$

- Das ist ein einfacher Klassifikator

## Funktionale Beschreibung

Es gilt:

$$c \in C = h_{bayes}(x) = \arg\max_{j=1,m} P(c_j) \prod_{i=1,n} P(x_i|c_j)$$

mit *c* als eine Klasse aus allen möglichen Klassenwerten *C* der Zielvariable *y* und *x* als eine Dateninstanz.

# Bayes Netzwerke

Naive Bayes Netzwerke bieten einen einfachen Ansatz mit klarer Semantik, um probabilistisches Wissen darzustellen, zu verwenden und zu lernen. Naiv $\leftrightarrow$ Alle $x_\mathrm{i}$ sind unabhängig!

**Ein bayesisches Netzwerk ist eine Form eines grafischen Modells zur Darstellung multivariater Wahrscheinlichkeitsverteilungen.**

[Witten]

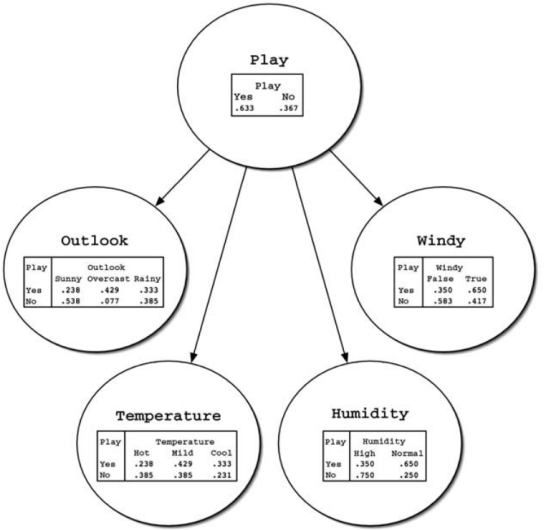



Abb. 2. Bayes Netzwerke sind eine Kombination aus Bäumen (gerichteten Graphen) und Wahrscheinlichkeitstabellen (Look-up Tabelle) die dann durch bedingte Wahrscheinlichkeiten das Ergebnis (der Hypothese von $y$) abschätzen.

[Witten]

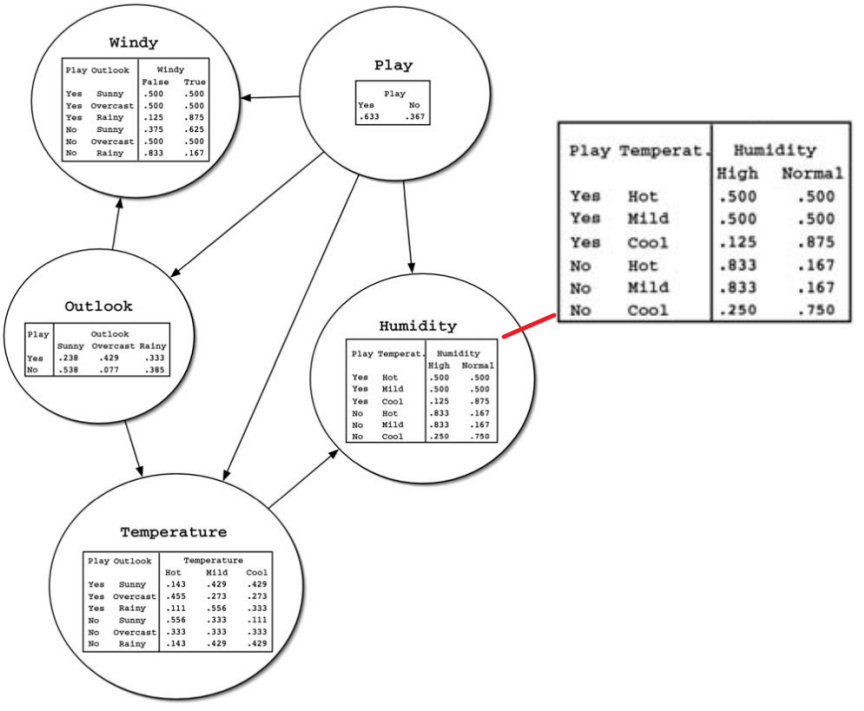
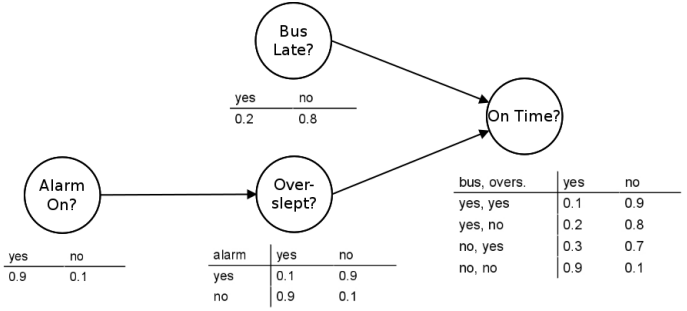



Abb. 3. Ein anderes Bayes Netzwerk für das gleiche Problem!

- Für das gleiche Problem können verschiedene Bayes Netzwerke aufgebaut werden, die genau die gleiche Wahrscheinlichkeitsverteilung darstellen.
  - Dies geschieht durch Änderung der Art und Weise, wie die gemeinsame Wahrscheinlichkeitsverteilung faktorisiert wird, um bedingte Unabhängigkeiten auszunutzen.

[Barber, BRaML, 2011]

# Bayes Netzwerk: Struktur und Parameter




[https://www.uib.no/en/rg/ml/119695/bayesian-networks]

Abb. 4. Bayessche Netzwerke bestehen aus Struktur und Parametern

- Die Struktur ist ein gerichteter azyklischer Graph (DAG), der bedingte Unabhängigkeiten und Abhängigkeiten zwischen Random-Variablen ausdrückt, die Knoten zugeordnet sind.
- Die Parameter bestehen aus bedingten Wahrscheinlichkeitsverteilungen, die jedem Knoten zugeordnet sind.
- Ein Bayessches Netzwerk ist daher eine kompakte, flexible und interpretierbare Darstellung einer gemeinsamen Wahrscheinlichkeitsverteilung.
- Es ist auch ein nützliches Werkzeug bei der Wissenserlangung, da gerichtete azyklische Graphen die Darstellung kausaler Beziehungen zwischen Variablen ermöglichen.

Typischerweise wird ein Bayessches Netzwerk aus Daten gelernt.

# Belief Netzwerke

Ein "Belief" Netzwerk ist eine Verteilung der Form:

$$p(x_1, x_2, .., x_D) = \Pi_{i=1}^{D} p(x_i | pa(x_i))$$

mit $pa(x_\mathrm{i})$ als die Elternvariable der Variablen $x_\mathrm{i}$.

- Dargestellt als gerichteter Graph, wobei ein Pfeil von einer Elternvariablen auf eine untergeordnete Variable zeigt.
- Ein Belief Netzwerk ist ein gerichteter azyklischer Graph (DAG), wobei der i-te Knoten im Graph dem Faktor $p(x_\mathrm{i}|pa(x_\mathrm{i}))$ entspricht.

Übergang zu Markovschen Entscheidungsprozessen und Graphen!

# Bayes Entscheidungslerner

## Bayes Lerner

Es werden zwei Funktionen zum Erlernen der Netzwerke benötigt (Struktur und Berechnung):

1. Eine Funktion um ein gegebenes Netzwerk zu evaluieren
2. Eine Funktion um geeignete Netzwerke aus dem Raum aller möglichen Netzwerke zu finden (Suche)

[Barber, BRaML, 2011, pp. 288]

[Witten, DMPMLTaT, pp. 335]

Ein einfacher und sehr schneller Lernalgorithmus ist **K2**

- Er beginnt mit einer bestimmten Reihenfolge der Attribute (d.h. Knoten).
- Dann verarbeitet er jeden Knoten der Reihe nach und wird Kanten von zuvor verarbeiteten Knoten zu dem aktuellen Knoten hinzuzufügen.
- In jedem Schritt wird die Kante hinzugefügt, die die Trefferwahrscheinlichekit (Score) des Netzwerks maximiert. Wenn es keine weitere Verbesserung gibt, wird der nächste Knoten bearbeitet.
- Als zusätzlicher Mechanismus zur Vermeidung von Überanpassungen kann die Anzahl der Eltern für jeden Knoten auf ein vordefiniertes Maximum beschränkt werden.

- Da nur Kanten von zuvor verarbeiteten Knoten berücksichtigt werden und es eine feste Reihenfolge gibt, kann dieses Verfahren keine Zyklen einführen.

## Randomisiertes Sampling

- Das Ergebnis hängt jedoch von der anfänglichen Reihenfolge ab, daher ist es sinnvoll, den Algorithmus mehrmals mit unterschiedlichen zufälligen Ordnungen auszuführen.

## Klassifikation und Bewertung

- Bei Entscheidungsbäumen gibt es keine Information wie "sicher" eine Klassifikation ist (Vorhersage- oder Vertrauenwahrscheinlichkeit), die aber häufig sehr wicht ist!
- Bei ANN ist der Ausgangswert eines Neurons zwar ein Indikator für die Aktivierungsstärke, ist aber modellbasiert keine Wahrscheinlichkeit der Vorhersage (höchstens grobe Näherung)
- Bei NB Klassifizieren gibt es als Ergebnis immer eine (statistisch) modellbasierte Wahrscheinlichkeit!

# Anwendungen von Naiven Bayes-Algorithmen

### Echtzeit-Vorhersage

Naive Bayesfunktionen sind ein einfacher und schneller Klassifikator und geben Wahrscheinlichkeiten aus für die Beurteilung. Somit könnte es für Vorhersagen in Echtzeit verwendet werden.

### Mehrklassen-Vorhersage

Dieser Algorithmus ist auch für Mehrklassen-Vorhersage Funktionen verwendbar. Hier können die Wahrscheinlichkeiten mehrerer Klassen von Zielvariablen vorhergesagt werden.

### Textklassifizierung / Spam-Filterung / Stimmungsanalyse

Naive Bayes-Klassifikatoren, die hauptsächlich in der Textklassifizierung verwendet werden (aufgrund eines besseren Ergebnisses bei Problemen mit mehreren Klassen und der Unabhängigkeitsregel), weisen im Vergleich zu anderen Algorithmen eine höhere Erfolgsrate auf. Infolgedessen werden sie häufig in der Spam-Filterung (Identifizierung von spam-e-Mails) und in der Stimmungsanalyse (in der social-media-Analyse) verwendet, um positive und negative Kundenstimmungen zu identifizieren).

### Empfehlungssystem

Der Naive Bayes-Klassifikator und die "kollaborative Filterung" erstellen zusammen ein Empfehlungssystem, das maschinelles Lernen und Data Mining Techniken verwendet, um versteckte Informationen zu filtern und vorherzusagen, ob ein Benutzer eine bestimmte Ressource möchte oder nicht.

# Zusammenfassung

Trainingsdaten

$$x \sim p(x|y)$$
$$y \sim p(y)$$

Vorhersagemodell

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

# Zusammenfassung

- Bayes Klassifikation erfolgt über die Berechnung bedingter und abhängiger Wahrscheinlichkeitsverteilungen
- Das "Training" liefert $P(x|y)$, die Inferenz benötigt die inverse Abhängigkeit und verwendet $P(y|x)$
- Ein naiver Bayes Klassifikator nimmt unabhänge Varibalen $x_\mathrm{i}$ an $\rightarrow$ Entspricht nicht immer "physikalischen Modellen und Kausalitäten"!
- Ein Bayes Klassifikator muss Tabellen mit bedingten Verteilungen/Wahrscheinlichkeiten verwenden.