Datenzentrierte Informatik

Prof. Dr. Elena Demidova

Arbeitsgruppe Data Science & Intelligente Systeme (DSIS)

Abteilung III: Informationssysteme und Künstliche Intelligenz
Institut für Informatik
Rheinische Friedrich-Wilhelms-Universität Bonn

WS 25/26

Bayessche Klassifikation

Thematische Einordnung und Lernziele

- Thematische Einordnung:
  - Datenzentrierte Informatik $\rightarrow$ Maschinelles Lernen
- notwendige Vorkenntnisse:
  - Grundlagen der Wahrscheinlichkeitstheorie
- Lernziele:
  - Bayessche Klassifikation
    - Bernoulli-Verteilung
    - Bayesscher Klassifikator
    - Naiver Bayesscher Klassifikator
    - Anwendung auf Dokumentkategorisierung
- Vorlesung basiert auf Kapiteln 3 und 5.7 in [1]

Einführung: Bayessche Klassifikation

- Daten stammen oft aus komplexen Prozessen, deren Dynamik nicht vollständig bekannt sind
- Selbst bei deterministischen Prozessen fehlt oft die vollständige Kenntnis der zugrundeliegenden Mechanismen und Bedingungen
- Daher modellieren wir solche Prozesse als stochastisch und verwenden Wahrscheinlichkeitstheorie zur Analyse

Einführung: Beobachtbare und unbeobachtbare Variablen

- beobachtbare Variablen (engl. observable variables): Merkmale, die direkt gemessen oder beobachtet werden können
  - Beispiele:
    - Ergebnis eines Münzwurfs (Kopf/Zahl)
    - Einkommen und Ersparnisse von Kunden bei einer Bank
    - Häufigkeiten der Wörter in Textdokumenten (z. B. im Bag-of-Words-Modell)

Einführung: Beobachtbare und unbeobachtbare Variablen

- unbeobachtbare/latente Variablen (engl. latent variables): Einflussfaktoren, die nicht direkt beobachtet oder gemessen werden können
  - Beispiele:
    - Zusammensetzung einer Münze (Gewichtsverteilung)
    - Gesundheitsinformationen der Kunden für die Bank
- Der Wert einer beobachtbaren Variable X ist eine Funktion der Werte von unbeobachtbaren Variablen $\mathbf{Z}$

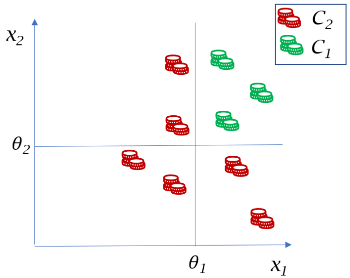
$$x = f(\mathbf{z})$$

  - $f(\cdot)$ ist eine deterministische Funktion
  - $\mathbf{z}$ ist unbekannt und nicht direkt beobachtbar
  - daher können wir $f(\cdot)$ nicht direkt berechnen

Beispiele von ML-Anwendungen: Klassifikation

Problem: Einschätzung des Altersarmutrisikos basierend auf Einkommen (X_1) und Altersvorsorge (X_2)

- **Ziel: Klassifikation in zwei Klassen**
  - $\mathcal{C}_1$: niedriges Risiko (z. B. finanziell abgesichert)
  - $\mathcal{C}_2$: hohes Risiko (z. B. drohende Altersarmut)
- beobachtbare Variablen:
  - Einkommen X_1, Altersvorsorge X_2
- Gibt es weitere relevante (latente) Variablen?



Einführung: Bernoulli-Variable

- Eine **Zufallsvariable X** modelliert ein zufälliges Ergebnis eines Experiments
- Die möglichen Realisierungen von X werden gemäß einer Wahrscheinlichkeitsverteilung $P(X = x)$ generiert, die den zugrunde liegenden Prozess beschreibt
  - **Beispiel Münzwurf:** X kann zwei Werte annehmen

$$\begin{cases} X = 1, & \text{für Kopf} \\ X = 0, & \text{für Zahl} \end{cases}$$

  - Bei einer Klassifikation ist X entweder:

$$\begin{cases} X = 1, & \text{für ein positives Beispiel der Klasse} \\ X = 0, & \text{ein negatives Beispiel (keine Instanz der Klasse)} \end{cases}$$

- Zufallsvariablen mit genau zwei möglichen Ergebnissen (z. B. 0 oder 1) folgen der Bernoulli-Verteilung

Bernoulli-Verteilung

- Die **Bernoulli-Verteilung (engl. Bernoulli density)** modelliert ein Zufallsexperiment mit genau zwei möglichen Ergebnissen

$$\begin{cases} X = 1: & \text{ein Ereignis tritt ein} \\ X = 0: & \text{ein Ereignis tritt nicht ein} \end{cases}$$

- **Parameter** p ist die Wahrscheinlichkeit für das Ergebnis $X = 1$:

$$\begin{cases} P(X = 1) = p, & \text{die Wahrscheinlichkeit des Eintretens} \\ P(X = 0) = 1 - p, & \text{die Wahrscheinlichkeit des Nichteintretens} \end{cases}$$

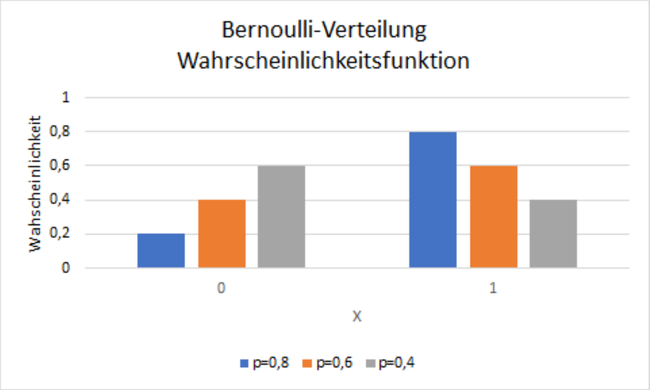
  - Notation:
    - X: eine Zufallsvariable
    - x: ein konkreter Wert, den X annehmen kann

- **Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung**:

$$P(x) = p^x(1-p)^{1-x}, x \in \{0, 1\}$$

  - für $x = 1$: $P(X = x) = p$
  - für $x = 0$: $P(X = x) = 1 - p$

- **Beispiel:** Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung in Abhängigkeit vom Parameter p



  - Erwartungswert: $E[x] = p$, Varianz: $Var(x) = p(1-p)$

Beispiel: Altersarmutrisiko als Bernoulli-Variable

- Wir modellieren das Altersarmutrisiko einer Person mit einer Bernoulli-Variablen $\mathcal{C}$, bedingt durch zwei beobachtbaren Variablen $\mathbf{X} = [X_1, X_2]^T$
  - $\mathcal{C} = 1$: niedriges Risiko (finanziell abgesichert)
  - $\mathcal{C} = 0$: hohes Risiko (Gefahr der Altersarmut)
- Wenn wir die bedingte Wahrscheinlichkeit $P(\mathcal{C}|X_1, X_2)$ kennen, können wir bei einer Person mit gegebenen Werten $X_1 = x_1$ und $X_2 = x_2$ eine Entscheidung treffen:

$$\text{wähle} \begin{cases} \mathcal{C} = 1, & \text{falls } P(\mathcal{C} = 1|x_1, x_2) > P(\mathcal{C} = 0|x_1, x_2) \\ \mathcal{C} = 0, & \text{sonst} \end{cases}$$

- Die Fehlerwahrscheinlichkeit dieser Entscheidung (die Wahrscheinlichkeit, dass die Klassifikation falsch ist):

$$1 - max(P(\mathcal{C} = 1|x_1, x_2), P(\mathcal{C} = 0|x_1, x_2))$$

Satz von Bayes

- Sei $\mathbf{x} = [x_1, x_2]^T$ der Vektor mit den Werten der beobachteten Variablen
  - Jede Komponente x_i repräsentiert ein Merkmal einer Person (z. B. Einkommen, Altersvorsorge, Alter)
- Ziel: gegeben $\mathbf{x}$, Schätzung der bedingten Wahrscheinlichkeit $P(\mathcal{C} \mid \mathbf{x})$
  - Wie wahrscheinlich ist die Klasse $\mathcal{C}$, wenn wir die Werte der beobachteten Merkmale $\mathbf{x}$ kennen?
- Wir nutzen den Satz von Bayes, um $P(\mathcal{C} \mid \mathbf{x})$ zu berechnen

Satz von Bayes

Mit dem Satz von Bayes (engl. Bayes' rule) erhalten wir:

$$P(\mathcal{C}|\mathbf{x}) = \frac{P(\mathcal{C})p(\mathbf{x}|\mathcal{C})}{p(\mathbf{x})}$$

- $P(\mathcal{C}|\mathbf{x})$: a-posteriori-Wahrscheinlichkeit
- $P(\mathcal{C})$: a-priori-Wahrscheinlichkeit der Klasse
  - vor der Beobachtung von $\mathbf{x}$
- $p(\mathbf{x}|\mathcal{C})$: Klassen-Likelihood
  - Wahrscheinlichkeit der beobachteten Daten gegeben der Klasse
- $p(\mathbf{x})$: Evidenz
  - die Gesamtwahrscheinlichkeit der beobachteten Daten (über alle Klassen hinweg)

Notation: $P(X)$: Wahrscheinlichkeitsfunktion, wenn X diskret ist
$p(X)$: Wahrscheinlichkeitsfunktion, wenn X stetig ist

Satz von Bayes: a-priori-Wahrscheinlichkeit

- a-priori-Wahrscheinlichkeit (engl. prior probability) $P(\mathcal{C} = 1)$ ist die Wahrscheinlichkeit dafür, dass $\mathcal{C}$ den Wert 1 annimmt, unabhängig von den beobachteten Werten $\mathbf{x}$
  - $P(\mathcal{C})$ stellt das Wissen oder die Annahmen dar, die wir über den Wert $\mathcal{C}$ haben, bevor wir die beobachtbaren Variablen betrachten
  - in diesem Beispiel: $P(\mathcal{C} = 1)$ beschreibt den Anteil der Hochrisikopersonen in der Population
- Bei zwei Klassen gilt die Normalisierungsbedingung:

$$P(\mathcal{C} = 0) + P(\mathcal{C} = 1) = 1$$

Satz von Bayes:

$$\boxed{P(\mathcal{C}|\mathbf{x}) = \frac{P(\mathcal{C})p(\mathbf{x}|\mathcal{C})}{p(\mathbf{x})}}$$

Satz von Bayes: Klassen-Likelihood

- Klassen-Likelihood (engl. class likelihood) $p(\mathbf{x}|\mathcal{C})$ ist die bedingte Wahrscheinlichkeit, dass eine Instanz in der Klasse $\mathcal{C}$ den beobachteten Wert $\mathbf{x}$ hat
  - Klassen-Likelihood beschreibt, wie wahrscheinlich es ist, die beobachteten Merkmale $\mathbf{x}$ zu sehen, wenn wir wissen, dass das Beispiel der Klasse $\mathcal{C}$ angehört
  - Beispiel:
  $$p(x_1, x_2|\mathcal{C} = 1)$$
  ist die Wahrscheinlichkeit, dass eine Person mit niedrigem Risiko die beobachteten Merkmale $X_1 = x_1$ und $X_2 = x_2$ hat
  - diese Informationen über der Klasse erhalten wir typischerweise aus den Daten (s. später)

Satz von Bayes:

$$\boxed{P(\mathcal{C}|\mathbf{x}) = \frac{P(\mathcal{C})p(\mathbf{x}|\mathcal{C})}{p(\mathbf{x})}}$$

Satz von Bayes: Evidenz

- Evidenz (engl. evidence) $p(\mathbf{x})$ ist die Randwahrscheinlichkeit dafür, dass eine Beobachtung $\mathbf{x}$ gemacht wird, unabhängig davon, ob es sich um ein positives oder negatives Beispiel handelt
  - Bei zwei Klassen:

$$p(\mathbf{x}) = p(\mathbf{x}|\mathcal{C}=1)P(\mathcal{C}=1) + p(\mathbf{x}|\mathcal{C}=0)P(\mathcal{C}=0)$$

  - Verallgemeinerung auf K Klassen:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$$

    - hier wird für jede Klasse $\mathcal{C}_k$ die Klassen-Likelihood $p(\mathbf{x} \mid \mathcal{C}_k)$ mit der a-priori-Wahrscheinlichkeit der Klasse $P(\mathcal{C}_k)$ gewichtet

Satz von Bayes:

$$\boxed{P(\mathcal{C}|\mathbf{x}) = \frac{P(\mathcal{C})p(\mathbf{x}|\mathcal{C})}{p(\mathbf{x})}}$$

Satz von Bayes: Verallgemeinerung auf K Klassen

- Im allgemeinen Fall haben wir K sich gegenseitig ausschließende Klassen $\mathcal{C}_k, k = 1, ..., K$
  - Jede Klasse repräsentiert eine mögliche Kategorie der Instanz
- Angenommen, wir kennen die a-priori-Wahrscheinlichkeiten $P(\mathcal{C}_k)$ und die Klassen-Likelihoods $p(\mathbf{x}|\mathcal{C}_k)$ für jede Klasse $\mathcal{C}_k$
  - diese können aus einer Stichprobe geschätzt werden
- Die a-priori-Wahrscheinlichkeiten $P(\mathcal{C}_k)$ erfüllen folgende Bedingungen:
  - $P(\mathcal{C}_k) \geq 0$
  - bei K Klassen: $\sum_{k=1}^{K} P(\mathcal{C}_k) = 1$

Schätzung der a-priori-Wahrscheinlichkeit

Wir wollen die a-priori-Wahrscheinlichkeiten der Klassen anhand einer Stichprobe schätzen

- Gegeben eine Trainingsstichprobe (Trainingsmenge) $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), ..., (\mathbf{x}^n, \mathbf{y}^n)\}$, mit

$$y_k^i = \begin{cases} 1, & \text{falls } \mathbf{x}^i \in \mathcal{C}_k \\ 0, & \text{sonst} \end{cases}$$

  - y_k^i: binäre Indikatorvariable für die Zugehörigkeit von Trainingsinstanz $\mathbf{x}^i$ zu einer bestimmten Klasse $\mathcal{C}_k$
- a-priori-Wahrscheinlichkeit der Klasse $P(\mathcal{C}_k)$ kann durch den Anteil der Instanzen der Klasse $\mathcal{C}_k$ in der Stichprobe geschätzt werden:

$$\hat{P}(\mathcal{C}_k) = \frac{\sum_i y_k^i}{N}$$

  - durch $\hat{..}$ (Zirkumflex) wird eine Schätzung bezeichnet
  - N: Anzahl der Instanzen in der Trainingsmenge
- dies ist ein Beispiel der Maximum-Likelihood-Schätzung
  - mehr dazu in der nächsten Vorlesung

Satz von Bayes: a-posteriori-Wahrscheinlichkeit

- Die a-posteriori-Wahrscheinlichkeit der Klasse $\mathcal{C}_k$ (engl. posterior probability)
  - beschreibt die Wahrscheinlichkeit, dass $\mathcal{C}_k$ zutrifft, nachdem wir die Beobachtung $\mathbf{x}$ gemacht haben
  - errechnet sich aus den a-priori-Wahrscheinlichkeiten und Klassen-Likelihoods:

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{\sum_{j=1}^{K} p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)}$$

- Durch die Normalisierung durch die Evidenz $p(\mathbf{x})$ summieren sich die a-Posteriori-Werte über alle Klassen zu 1

Bayesscher Klassifikator

- Bayesscher Klassifikator (engl. Bayes' classifier) selektiert die Klasse mit der höchsten a-posteriori-Wahrscheinlichkeit:

$$\text{wähle } \mathcal{C}_k \text{ falls } P(\mathcal{C}_k|\mathbf{x}) = \max_j P(\mathcal{C}_j|\mathbf{x})$$

  - $P(\mathcal{C}_k|\mathbf{x})$: die a-posteriori-Wahrscheinlichkeit der Klasse $\mathcal{C}_k$ gegeben die Daten $\mathbf{x}$

Diskriminanzfunktionen

- Klassifikation kann als eine Menge von Diskriminanzfunktionen (engl. discriminant functions)

$$g_k(\mathbf{x}), k = 1, ..., K$$

  gesehen werden
- Ziel: die Klasse wählen, deren Diskriminanzfunktion den höchsten Wert für den gegebenen Vektor $\mathbf{x}$ liefert

$$\text{wähle } \mathcal{C}_k \text{ falls } g_k(\mathbf{x}) = \max_j g_j(\mathbf{x})$$

- Eine Diskriminanzfunktion $g_k(\mathbf{x})$ gibt an, wie gut die Beobachtung $\mathbf{x}$ zur Klasse $\mathcal{C}_k$ passt

Diskriminanzfunktionen: Bayesscher Klassifikator

- Bayesscher Klassifikator kann mithilfe von Diskriminanzfunktionen darstellt werden:

$$g_k(\mathbf{x}) = P(\mathcal{C}_k|\mathbf{x}) \equiv p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$$

  - der gemeinsame Evidenz-Term $p(\mathbf{x})$ im Satz von Bayes ändert die Klassenzuordnung nicht, und kann ignoriert werden
- Äquivalent können wir die Log-Variante nutzen:

$$g_k(\mathbf{x}) = log\ p(\mathbf{x}|\mathcal{C}_k) + log\ P(\mathcal{C}_k)$$

  - $log(\cdot)$ konvertiert das Produkt in eine Summe
  - Vereinfachung des Rechenaufwands

Rechenregeln: $log(u \cdot v) = log(u) + log(v)$

Naiver Bayesscher Klassifikator

Wie berechnen wir die Klassen-Likelihood $p(\mathbf{x}|\mathcal{C}_k)$ bei d Variablen, d. h. wenn $\mathbf{x} \in \mathbb{R}^d$ ein d-dimensionaler Vektor ist?

- Naiver Bayesscher Klassifikator (engl. naive Bayes' classifier) nimmt an, dass die Variablen unabhängig sind
  - reduziert das Problem mit mehreren Variablen auf eine Gruppe von univariaten Problemen
- Die Klassen-Likelihood $p(\mathbf{x}|\mathcal{C}_k)$ für die Klasse $\mathcal{C}_k$ wird als Produkt von Klassen-Likelihoods der einzelnen Dimensionen bestimmt:

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{j=1}^{d} p(x_j|\mathcal{C}_k)$$

  - mögliche Abhängigkeiten bzw. Korrelationen zwischen den Variablen werden ignoriert (daher: „naiv“)
  - führt zur Vereinfachung des Berechnungsaufwands

Schätzung der Klassen-Likelihood

Schätzung der Klassen-Likelihood $p(x_j|\mathcal{C}_k)$ (univariat) aus den Daten

- $p(x_j|\mathcal{C}_k)$ ist die Wahrscheinlichkeit, dass für eine Instanz in $\mathcal{C}_k$, $X_j = x_j$ zutrifft
  - Beispiel: die Wahrscheinlichkeit, dass eine Person mit hohem Risiko über Einkommen von $X_j = 50.000$ EUR verfügt
- falls X_j diskret ist:
  - $p(x_j|\mathcal{C}_k)$ kann durch den Anteil der Trainingsdatensätze der Klasse $\mathcal{C}_k$ mit dem Wert $X_j = x_j$ geschätzt werden

$$p(x_j|\mathcal{C}_k) = \frac{\text{Anzahl der Instanzen in } \mathcal{C}_k \text{ mit } X_j = x_j}{\text{Anzahl der Instanzen in } \mathcal{C}_k}$$

- falls X_j stetig ist, gibt es mehrere Möglichkeiten:
  - durch Diskretisierung und Ersetzung durch das entsprechende diskrete Intervall
  - durch die Annahme einer bestimmten Form der Wahrscheinlichkeitsverteilung (z. B. Normalverteilung) und Schätzung der Parameter anhand der Trainingsdaten
    - später in der Vorlesung

NB-Klassifikator für Dokumentkategorisierung

Naiver Bayesscher Klassifikator für Dokumentkategorisierung

- schätzt die Wahrscheinlichkeit, dass ein Dokument zur Klasse gehört, basierend auf den Häufigkeiten der Wörter in den Dokumenten
- Beispiel: Einordnung von Dokumenten in Kategorien Wintersport, Sommersport, Wirtschaft, Unterhaltung, Politik
  - Klassen: Dokumentenkategorien
    - $\mathcal{C}_1$: Wintersport
    - $\mathcal{C}_2$: Sommersport
    - ...
  - Variablen: Wörter (oder "Token") in den Dokumenten

Beispiel: Dokumentkategorisierung

- Ziel: Klassifizierung der Dokumenten mithilfe des Naiven Bayesschen Klassifikators
  - Schritt 1: Berechnung der Parameter für den multinomialen Naiven Bayesschen Klassifikator aus der Trainingsmenge
    - basierend auf den Häufigkeiten der Wörter in den Dokumenten in der Trainingsmenge
  - Schritt 2: Anwendung des Klassifikators auf die Testdokumenten (hier: D_5), um eine Zuordnung zu einer der Klassen $\mathcal{C}_1$ oder $\mathcal{C}_2$ zu treffen

	Dokument	**Wörter**	**Klasse**
Trainingsmenge	D_1	Sport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_2	Wintersport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_3	Wintersport	$\mathcal{C}_1$
	D_4	Radfahren	$\mathcal{C}_2$
Testmenge	D_5	Sport, Radfahren	?

A-priori-Wahrscheinlichkeit

- Die a-priori-Wahrscheinlichkeit einer Klasse $\mathcal{C}_k$ ist die Wahrscheinlichkeit, dass ein Dokument zufällig dieser Klasse zugeordnet wird, bevor wir Merkmale (z. B. Wörter) betrachten
- die Schätzung basiert auf der Häufigkeit der Klassen in der Trainingsmenge:

$$\hat{P}(\mathcal{C}_1) = \frac{3}{4}$$

  - drei von vier Trainingsdokumenten gehören zur Klasse $\mathcal{C}_1$

$$\hat{P}(\mathcal{C}_2) = \frac{1}{4}$$

  - ein von vier Trainingsdokumenten gehört zur Klasse $\mathcal{C}_2$

	Dokument	**Wörter**	**Klasse**
Trainingsmenge	D_1	Sport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_2	Wintersport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_3	Wintersport	$\mathcal{C}_1$
	D_4	Radfahren	$\mathcal{C}_2$
Testmenge	D_5	Sport, Radfahren	?

Merkmalsselektion für Dokumentkategorisierung

- Vokabular V bestehend aus d Wörtern, die relevanten Informationen für die jeweiligen Klassen liefern
- nützliche Wörter sollen eine hohe Wahrscheinlichkeit für eine (oder wenige) Klasse(n) haben, und niedrige Wahrscheinlichkeiten für alle anderen Klassen
- Wörter, deren Wahrscheinlichkeiten für verschiedene Klassen ähnlich sind, liefern wenig Information
  - Beispiel: Stoppwörter – Wörter, die in der Sprache häufig vorkommen (z. B. "und", "der")
  - Beispiel: mehrdeutige Wörter – Wörter mit verschiedenen Bedeutungen, je nach Kontext (z. B. "Jaguar": Auto vs. Tier)
- Beispiel hier: V besteht aus allen Wörtern aus den Dokumenten in der Trainingsmenge

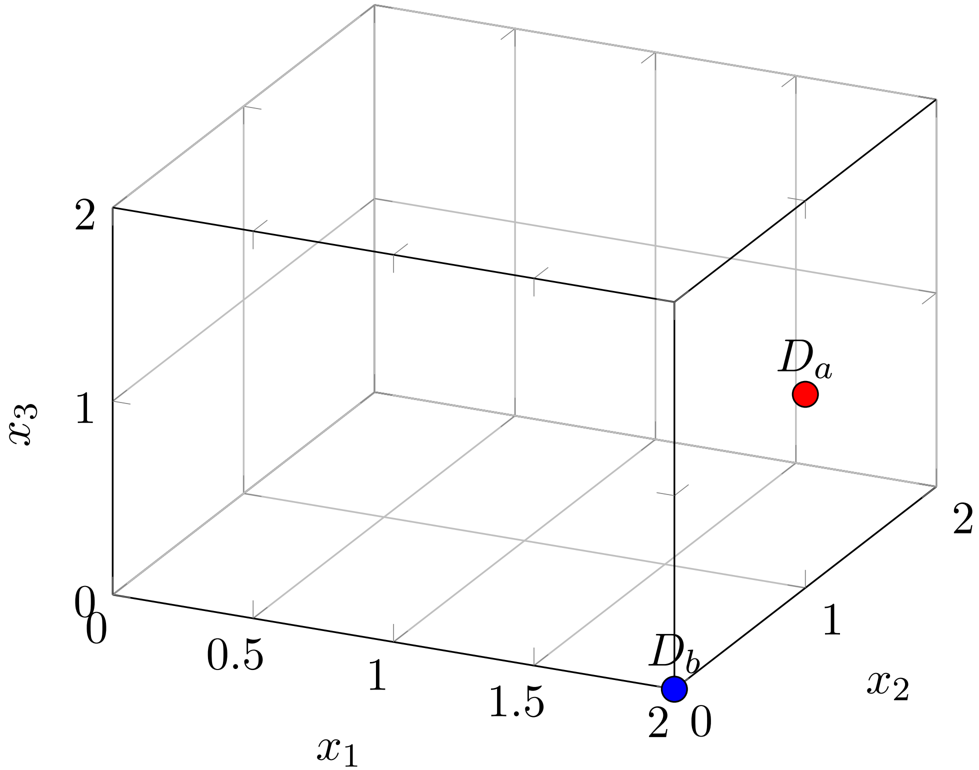
Bag-of-Words-Repräsentation

Bag-of-Words-Repräsentation (engl. bag of words)

- Ein Dokument wird als Vektor $\mathbf{x} \in \mathbb{R}^d$ im d-dimensionalen Raum dargestellt
- Dimensionen des Vektorraums entsprechen den Wörtern im Vokabular V
- jede Dimension des Dokumentenvektors repräsentiert die Häufigkeit eines bestimmten Wortes im Dokument

Bag-of-Words-Repräsentation

- Beispiel: $V = \{\text{Apfel, Mango, Banane}\}$
- Welche Wörter beinhalten die abgebildeten Dokumente?



Schätzung der Klassen-Likelihood

- Schätzung der Klassen-Likelihood $p(x_j|\mathcal{C}_k)$, dass das Wort x_j in einem Dokument der Klasse $\mathcal{C}_k$ vorkommt, anhand einer Trainingsmenge $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}$
- Intuitiv: $p(x_j|\mathcal{C}_k)$ kann durch die relative Häufigkeit des Wortes x_j in den Trainingsdokumenten der Klasse $\mathcal{C}_k$ geschätzt werden:

$$\hat{p}(x_j|\mathcal{C}_k) = \frac{\sum_i f_j^i \cdot y_k^i}{\sum_i \sum_{v \in V} f_v^i \cdot y_k^i}$$

  - f_j^i: Häufigkeit des Wortes x_j im Dokument i
  - y_k^i: binäre Indikatorvariable für die Klasse $\mathcal{C}_k$
  - V: das Vokabular

$$\hat{p}(x_j|\mathcal{C}_k) = \frac{\text{Häufigkeit des Wortes } x_j \text{ in den Dokumenten aus } \mathcal{C}_k}{\text{gesamte Anzahl der Wörter in den Dokumenten aus } \mathcal{C}_k}$$

Beispiel: Dokumentkategorisierung

- Die Likelihood $\hat{p}(x_j|\mathcal{C}_k)$, dass das Wort x_j in einem Dokument der Klasse $\mathcal{C}_k$ vorkommt:

$$\hat{p}(Sport|\mathcal{C}_1) = 1/5 \quad \hat{p}(Radfahren|\mathcal{C}_1) = 0$$

  - "Radfahren" kommt in keinem Dokument der Klasse $\mathcal{C}_1$ in der Trainingsmenge vor!

$$\hat{p}(Sport|\mathcal{C}_2) = 0 \quad \hat{p}(Radfahren|\mathcal{C}_2) = 1$$

- Problem: Klassen-Likelihoods für Dokument D_5 werden zu Null!

$$\hat{p}(D_5|\mathcal{C}_1) = 1/5 \cdot 0 = 0 \quad \hat{p}(D_5|\mathcal{C}_2) = 0 \cdot 1 = 0$$

- Grund: **Datenknappheit**, nicht alle Wörter kommen in den Trainingsdokumenten einer Klasse vor

	Dokument	**Wörter**	**Klasse**
Trainingsmenge	D_1	Sport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_2	Wintersport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_3	Wintersport	$\mathcal{C}_1$
	D_4	Radfahren	$\mathcal{C}_2$
Testmenge	D_5	Sport, Radfahren	?

Schätzung mit Wordhäufigkeiten und Laplace-Glättung

- Ergänzung: wir nehmen an, dass jedes Word in jeder Klasse mindestens einmal vorkommt
- Wir nutzen Wordhäufigkeiten mit Laplace-Glättung, um die Klassen-Likelihood zu schätzen:

$$\hat{p}(x_j|\mathcal{C}_k) = \frac{(\sum_i f_j^i \cdot y_k^i) + 1}{(\sum_i \sum_{v \in V} f_v^i \cdot y_k^i) + |V|}$$

  - $+1$ ist die Laplace-Glättung, die fügt zu jeder Zählung eins zu, um Nullen zu eliminieren
  - $|V|$ ist die Größe des Vokabulars

Beispiel: Dokumentkategorisierung

- a-priori-Wahrscheinlichkeiten der Klassen:

$$\hat{P}(\mathcal{C}_1) = 3/4 = 0,75 \quad \hat{P}(\mathcal{C}_2) = 1/4 = 0,25$$

- Klassen-Likelihoods mit Laplace-Glättung ($|V| = 4$):

$$\hat{p}(Sport|\mathcal{C}_1) = \frac{1+1}{5+4} = \frac{2}{9} \quad \hat{p}(Radfahren|\mathcal{C}_1) = \frac{0+1}{9} = \frac{1}{9}$$

$$\hat{p}(Sport|\mathcal{C}_2) = \frac{0+1}{1+4} = \frac{1}{5} \quad \hat{p}(Radfahren|\mathcal{C}_2) = \frac{1+1}{1+4} = \frac{2}{5}$$

- Klassen-Likelihoods für D_5:

$$\hat{p}(D_5|\mathcal{C}_1) = 2/9 \cdot 1/9 = 0,0242 \quad \hat{p}(D_5|\mathcal{C}_2) = 1/5 \cdot 2/5 = 0,08$$

- Diskriminanzfunktionen für D_5:

$$g_1(D_5) = \hat{P}(\mathcal{C}_1) \cdot \hat{p}(D_5|\mathcal{C}_1) = 0,75 \cdot 0,024 = 0,018 \quad g_2(D_5) = 0,25 \cdot 0,08 = 0,02$$

$\rightarrow D_5$ wird der Klasse $\mathcal{C}_2$ zugewiesen

	Dokument	**Wörter**	**Klasse**
Trainingsmenge	D_1	Sport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_2	Wintersport, Schlittschuhlaufen	$\mathcal{C}_1$
	D_3	Wintersport	$\mathcal{C}_1$
	D_4	Radfahren	$\mathcal{C}_2$
Testmenge	D_5	Sport, Radfahren	?

Fazit: Bayessche Klassifikation

Die Studierenden sollen in der Lage sein

- Bayessche Klassifikation zu beschreiben und zu diskutieren
- Naiver Bayesscher Klassifikator zu beschreiben und die Parameter anhand einer Stichprobe zu bestimmen
- Naiver Bayesscher Klassifikator anzuwenden

Literatur und Vorlesungsfolien

Maschinelles Lernen:

1. Ethem Alpaydin: „Maschinelles Lernen“. De Gruyter Studium. 2. Auflage, (2019).
2. Aurélien Géron: „Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme“. O’Reilly, 2. Auflage, (2020).

ML-Bücher sind auch elektronisch in der Bibliothek vorhanden!
`https://bonnus.ulb.uni-bonn.de/`

- Vorlesungsfolien: s. eCampus
    - Die Vorlesungsmaterialien, einschließlich aller Vorlesungsfolien, Übungsmaterialien und Prüfungsfragen, werden ausschließlich für die Teilnehmer*innen des Moduls ”BA-INF 035 Datenzentrierte Informatik” an der Universität Bonn im WS 2025/2026 bereitgestellt. Die Weitergabe an Dritte, die Veröffentlichung und die Verbreitung der Vorlesungsmaterialien sind untersagt.