Statistik und Data Science für die Informatik

Übung 3: Grundlagen der Wahrscheinlichkeitsrechnung, Bedingte Wahrscheinlichkeit & Naive Bayes

Hermann Stolte

Lehrstuhl “Datenbanken und Informationssysteme”

Übungstermine

- Es gibt über das Semester 7 Übungen mit verschiedenen Themen aus der VL
- Struktur: Vorstellung Übungsblätter, Besprechung von Musterlösungen, Beispiel-Aufgaben

Wochen	Start	Ende	Thema
16-17	15.04	26.04	Einführung
18-19	29.04	10.05	1. Aufgabenblatt
20-21	**13.05**	**24.05**	**2. Aufgabenblatt**
22-23	27.05	07.06	3. Aufgabenblatt
24-25	10.06	21.06	4. Aufgabenblatt
26-27	24.06	05.07	5. Aufgabenblatt
28-29	08.07	19.07	Klausurvorbereitung

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Aufgabe 1 (Features eines Datensatzes)

Wir betrachten den Tornado Datensatz mit 7 Spalten:

Jahr, Monat, Bundesstaat, Magnitude (Fujita-Skala), Streckenlänge (in Meilen), Breite (in Metern) und Anzahl der Verletzten.

a) Nennen Sie jeweils die Spalten, auf die die folgenden Eigenschaften zutreffen:

(1) diskret, (2) nominal, (3) ordinal, (4) kontinuierlich

Jahr	Monat	Bundesstaat	Magnitude	Streckenlaenge	Breite	Verletzte
2021	1	FL	1	21.37	200	0
2021	2	GA	0	0.50	50	0
2021	2	GA	2	1.81	600	5
2021	3	KS	2	12.15	100	0
2021	4	TX	0	2.61	50	0
2021	6	OH	2	5.60	200	0
2021	7	WI	1	2.44	125	0
2021	10	TN	0	2.51	25	0
2021	11	LA	1	0.16	100	0
2021	12	TN	4	168.53	2600	515
2021	12	KY	3	29.26	440	63
2021	12	TX	1	1.57	100	0
2021	12	FL	1	1.29	50	1
2021	12	AL	1	0.95	50	0
2021	12	GA	1	2.75	150	0
2021	12	GA	1	2.50	75	6

Lösung:

- Diskret:
  Jahr, Monat, Bundestaat, Magnitude, Breite, Verletzte
- Nominal:
  Bundesstaat
- Ordinal:
  Monat, Magnitude
- Kontinuierlich:
  Streckenlänge

Aufgabe 1 (Features eines Datensatzes)

Wir betrachten den Tornado Datensatz mit 7 Spalten:

Jahr, Monat, Bundesstaat, Magnitude (Fujita-Skala), Streckenlänge (in Meilen), Breite (in Metern) und Anzahl der Verletzten.

b) Bestimmen Sie die relative Frequenz von (1) Tornados mit einer Magnitude ≥ 2 und (2) Tornados mit mindestens einer verletzten Person. Geben Sie Ihren Rechenweg an.

Jahr	Monat	Bundesstaat	Magnitude	Streckenlaenge	Breite	Verletzte
2021	1	FL	1	21.37	200	0
2021	2	GA	0	0.50	50	0
2021	2	GA	2	1.81	600	5
2021	3	KS	2	12.15	100	0
2021	4	TX	0	2.61	50	0
2021	6	OH	2	5.60	200	0
2021	7	WI	1	2.44	125	0
2021	10	TN	0	2.51	25	0
2021	11	LA	1	0.16	100	0
2021	12	TN	4	168.53	2600	515
2021	12	KY	3	29.26	440	63
2021	12	TX	1	1.57	100	0
2021	12	FL	1	1.29	50	1
2021	12	AL	1	0.95	50	0
2021	12	GA	1	2.75	150	0
2021	12	GA	1	2.50	75	6

Lösung:

Anzahl von Tornados (Tabellenzeilen)

= 16

Anzahl von Tornados mit Magnitude ≥ 2

= 5

Anzahl von Tornados mit Verletzte ≥ 1

= 5

Frequenz von (1) und (2)

= 5/16

Aufgabe 1 (Features eines Datensatzes)

Wir betrachten den Tornado Datensatz mit 7 Spalten:

Jahr, Monat, Bundesstaat, Magnitude (Fujita-Skala), Streckenlänge (in Meilen), Breite (in Metern) und Anzahl der Verletzten.

c) Berechnen Sie (1) den Median und Mittelwert der Anzahl von verletzten Personen und (2) den Modus des Monats. Geben Sie Ihren Rechenweg an.

Jahr	Monat	Bundesstaat	Magnitude	Streckenlaenge	Breite	Verletzte
2021	1	FL	1	21.37	200	0
2021	2	GA	0	0.50	50	0
2021	2	GA	2	1.81	600	5
2021	3	KS	2	12.15	100	0
2021	4	TX	0	2.61	50	0
2021	6	OH	2	5.60	200	0
2021	7	WI	1	2.44	125	0
2021	10	TN	0	2.51	25	0
2021	11	LA	1	0.16	100	0
2021	12	TN	4	168.53	2600	515
2021	12	KY	3	29.26	440	63
2021	12	TX	1	1.57	100	0
2021	12	FL	1	1.29	50	1
2021	12	AL	1	0.95	50	0
2021	12	GA	1	2.75	150	0
2021	12	GA	1	2.50	75	6

Lösung:

Median der Anzahl von verletzten Personen

= 0 , d.h. 50% der Tonados haben keine (≤ 0) Verletzte

Mittelwert der Anzahl von verletzten Personen

= (5 + 515 + 63 + 1 + 6) / 16

= 590 / 16

= 36.875

Modus des Monats

= 12 (häufigster Monat mit 7 Tornados.)

Aufgabe 2 (Korrelation)

Wir betrachten Messenger-App Datensatz mit zwei Spalten:

X (Stunden)	1	2	3	4	5
Y (Nachrichten)	9.5	21.86	30.65	45.52	49.77

(a) Wie hoch ist die Kovarianz s_{XY} sowie die Pearson Korrelation r_{XY} zwischen X und Y?

Lösung:

- Kovarianz: $s_{XY} = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \hat{x}) \cdot (y_i - \hat{y}) = \frac{1}{4} \cdot \sum_{i=1}^{n}(X_i - 3) \cdot (Y_i - 31.46) = 26.05$
- Korrelation: $r_{XY} = \frac{s(X,Y)}{s_X s_Y} = \frac{26.05}{1.58 \cdot 16.64} \approx 0.99$

Aufgabe 2 (Korrelation)

Wir betrachten Messenger-App Datensatz mit zwei Spalten:

X (Stunden)	1	2	3	4	5
Y (Nachrichten)	9.5	21.86	30.65	45.52	49.77

(b) Berechnen Sie die lineare Regression $f(x) = k \cdot x + d$, die Stunden (X) auf gesendete Nachrichten (Y) abbildet. Welchen Wert liefert *f(3)*?

Lösung:

- $k = r_{XY} \cdot \frac{s_Y}{s_X} = 0.99 \cdot \frac{16.64}{1.58} \approx 10.42$
- $d = \hat{y} - k \cdot \hat{x} = 31.46 - 10.42 \cdot 3 \approx 0.2$
- $f(3) = 10.42 \cdot 3 + 0.2 = 31.46$

Aufgabe 2 (Korrelation)

Wir betrachten Messenger-App Datensatz mit zwei Spalten:

X (Stunden)	1	2	3	4	5
Y (Nachrichten)	9.5	21.86	30.65	45.52	49.77

(c) Wie viele Nachrichten werden laut Ihrer Funktion *f(x)* pro Stunde verschickt?

Lösung:

- Nachrichten pro Stunde sind genau die Steigung der Funktion *f(x)*
- Wir leiten $f'(x) = 10.42$ ab und erhalten die Änderungsrate pro Stunde
- Dementsprechend werden 10.42 Nachrichten pro Stunde verschickt

Aufgabe 3 (Multivariate Lineare Regression)

Wir betrachten Honda City Gebrauchtwagen-Datensatz mit drei Spalten:

- Baujahr, Kilometerstand und Verkaufspreis (in Tausend Dollar)

(a) Implementieren Sie eine lineare Regression, die mithilfe des Baujahres und des Kilometerstandes den Verkaufspreis schätzt.

Lösung:

```
6.51072259
7.42856019
6.5267987
9.14025518
8.33969668
...
```

```python
def teilaufgabe_a():
    year_built, km_driven, selling_price = load_honda_city_dataset()

    # Erstellung der Featurematrix X mit einer zusätzlichen Spalte mit Einsen für den Intercept
    intercept = np.ones((year_built.shape[0], 1))
    features = np.column_stack((intercept, year_built, km_driven))

    # Umwandlung von selling_price in einen Spaltenvektor
    y = selling_price.reshape(-1, 1)

    # Berechnung der Koeffizienten anhand der algebraischen Lösung
    theta = np.linalg.inv(features.T @ features) @ features.T @ y

    # Berechnung der Schätzungen
    y_pred = features @ theta

    # Schätzungen aus Konsistenzgründen als 1D-Array zurückgeben
    return y_pred.flatten()
```

Aufgabe 3 (Multivariate Lineare Regression)

Wir betrachten Honda City Gebrauchtwagen-Datensatz mit drei Spalten:

- Baujahr, Kilometerstand und Verkaufspreis (in Tausend Dollar)

(b) Berechnen Sie den ”Root Mean Square Error“ der Schätzwerte. Was bedeutet dieser Fehlerwert im Kontext der Aufgabe?

Lösung:

```
def teilaufgabe_b():
    year_built, km_driven, selling_price = load_honda_city_dataset()
    y_pred = teilaufgabe_a()

    # Berechnung des mittleren quadratischen Fehlers
    mse = np.mean((selling_price - y_pred)**2)

    # Berechnung der mittleren quadratischen Fehlerwurzel
    rmse = np.sqrt(mse)

    return rmse
```

Bedeutung:

- RMSE gibt an, wie groß die durchschnittlichen Schätzfehler des Modells sind, gemessen in Tausend Euro
- RMSE von ~0.94 bedeutet, dass die durchschnittlichen Schätzfehler des Modells ungefähr 940 Dollar betragen

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Zufällige Ereignisse

Konzeptuell: Ein ungewisses Vorkommnis in einem Experiment

- Allerlei Messungen (z.B. in den Naturwissenschaften)
- Glücksspiele (z.B. Kartenspiele, Würfelspiele)
- Physikalische Phänomene (z.B. Erdbeben, Waldbrände, Vulkanausbrüche)

Mathematische Grundbegriffe:

- Ω (Omega): Menge aller möglichen Versuchsausgänge (sicheres Ereignis)
- Ereignis A: Teilmenge von Ω, $A \subseteq \Omega$
- Elementarereignis: einelementige Teilmenge von Ω
- unmögliches Ereignis: ∅
- Komplementärereignis: $\overline{A} = \Omega \setminus A$
- Potenzmenge P(Ω): Menge der auftretenden Ereignisse

Zufällige Ereignisse (Beispiel)

Experiment: Eine Glühbirne testen

- Elementarereignisse {{Funktioniert(f)}, {Funktioniert_nicht(fn)}}
- Sicheres Ereignis: Ω (Omega): {f,fn}
- Ereignis A: Teilmenge von Ω, z.B {f} der {f,fn}
- Elementarereignis: entweder {f} oder {fn}
- Unmögliches Ereignis: Ø = {f} ∩ {fn}
- Komplementärereignis: $\bar{f}$ = fn sowie $\overline{\text{fn}}$= f
- Potenzmenge P(Ω): {Ø, {f}, {fn}, Ω}

Ereignisfeld

Konzeptuell: Ein Ereignisfeld ε (Epsilon) ist ein System von Teilmengen der Menge Ω

- strukturierte Sammlung von möglichen Ereignissen innerhalb eines Wahrscheinlichkeitsexperiments
- Ein Ereignis ist eine Kombination von verschiedenen Elementarereignissen

Formell gilt: $\varepsilon \subseteq P(\Omega)$ heißt Ereignisfeld über Ω, falls gilt:

1. $\Omega \in \varepsilon$
2. Gilt $A_i \in \varepsilon$ für $i \in \mathbf{N}$, dann folgt $\bigcap A_i \in \varepsilon$
3. Wenn $A \in \varepsilon$, dann ist auch $\bar{A} \in \varepsilon$

Für Ereignisse gelten Vereinigung, Schnitt und Komplement

Kommutativgesetz, Assoziativgesetz, Distributivgesetz, De‘Morgansche Regeln

Ereignisfeld

Grundlegende Eigenschaften:

1. Elementarereignisse schließen sich gegenseitig aus
2. Es tritt immer nur genau ein Elementarereignis ein
3. Ein Ereignis tritt genau dann ein, wenn eines seiner Elementarereignisse eintritt

Ereignisfeld

Grundlegende Eigenschaften:

1. **Elementarereignisse schließen sich gegenseitig aus**
2. Es tritt immer nur genau ein Elementarereignis ein
3. Ein Ereignis tritt genau dann ein, wenn eines seiner Elementarereignisse eintritt

Beispiel: Wenn wir einen Würfel werfen und eine 1 erhalten, können wir keine der anderen Zahlen (2, 3, 4, 5, 6) gleichzeitig erhalten

Ereignisfeld

Grundlegende Eigenschaften:

1. Elementarereignisse schließen sich gegenseitig aus
2. **Es tritt immer nur genau ein Elementarereignis ein**
3. Ein Ereignis tritt genau dann ein, wenn eines seiner Elementarereignisse eintritt

Beispiel: Bei jedem Wurf eines fairen Würfels tritt genau eines der Elementarereignisse (1, 2, 3, 4, 5, 6) ein

Ereignisfeld

Grundlegende Eigenschaften:

1. Elementarereignisse schließen sich gegenseitig aus
2. Es tritt immer genau ein Elementarereignis ein
3. **Ein Ereignis tritt genau dann ein, wenn eines seiner Elementarereignisse eintritt**

Beispiel: Angenommen, wir definieren ein Ereignis A als "wir werfen eine gerade Zahl". Ereignis A besteht aus den Elementarereignissen {2, 4, 6}. Wenn wir den Würfel werfen und eine 4 erhalten, tritt genau eines der Elementarereignisse von Ereignis A ein, nämlich die 4, und damit tritt auch Ereignis A ein.

Wahrscheinlichkeit

Sei ε ein Ereignisfeld, dann heißt die Abbildung P: ε → $\mathbb{R}$ Wahrscheinlicht, falls gilt:

1. Wahrscheinlichkeitswerte liegen immer im Bereich von 0 bis 1: Für alle A ∈ E gilt: $0 \leq P(A) \leq 1$
2. Die Gesamtwahrscheinlichkeit des Ereignisraums beträgt 1: $P(\Omega) = 1$
3. Sind die Ereignisse A1, A2, . . . paarweise unvereinbar, gilt die σ–Additivitätseigenschaft: $P(\cup Ai) = \Sigma P(Ai)$

Wahrscheinlichkeit (Beispiel)

Sei ε ein Ereignisfeld, dann heißt die Abbildung P: ε → $\mathbb{R}$ Wahrscheinlicht, falls gilt:

1. **Wahrscheinlichkeitswerte liegen immer im Bereich von 0 bis 1: Für alle A ∈ E gilt: 0 ≤ P(A) ≤ 1**
2. Die Gesamtwahrscheinlichkeit des Ereignisraums beträgt 1: P(Ω) = 1
3. Sind die Ereignisse A1, A2, . . . paarweise unvereinbar, gilt die σ–Additivitätseigenschaft: P(∪Ai) = Σ P(Ai)

Beispiel: Beim Werfen eines fairen Würfels beträgt die Wahrscheinlichkeit, eine "6" zu würfeln, 1/6, da es sechs mögliche Ergebnisse gibt und jedes Ergebnis gleich wahrscheinlich ist.

Wahrscheinlichkeit (Beispiel)

Sei ε ein Ereignisfeld, dann heißt die Abbildung P: ε → $\mathbb{R}$ Wahrscheinlicht, falls gilt:

1. Wahrscheinlichkeitswerte liegen immer im Bereich von 0 bis 1: Für alle A ∈ E gilt: $0 \leq P(A) \leq 1$
2. **Die Gesamtwahrscheinlichkeit des Ereignisraums beträgt 1: $P(\Omega) = 1$**
3. Sind die Ereignisse A1, A2, . . . paarweise unvereinbar, gilt die σ–Additivitätseigenschaft: $P(\cup Ai) = \Sigma P(Ai)$

Beispiel: Beim Ziehen einer Karte aus einem Kartenspiel gibt es 52 mögliche Karten, die gezogen werden können. Die Wahrscheinlichkeit, irgendeine Karte zu ziehen, beträgt 1, da genau eine Karte bei jedem Zug gezogen wird.

Wahrscheinlichkeit (Beispiel)

Sei ε ein Ereignisfeld, dann heißt die Abbildung P: ε → $\mathbb{R}$ Wahrscheinlicht, falls gilt:

1. Wahrscheinlichkeitswerte liegen immer im Bereich von 0 bis 1: Für alle A ∈ E gilt: 0 ≤ P(A) ≤ 1
2. Die Gesamtwahrscheinlichkeit des Ereignisraums beträgt 1: P(Ω) = 1
3. **Sind die Ereignisse A1, A2, . . . paarweise unvereinbar, gilt die σ–Additivitätseigenschaft: P(∪Ai) = Σ P(Ai)**

Beispiel: Angenommen, wir werfen einen fairen Würfel und betrachten die Ereignisse A = "eine gerade Zahl werfen" und B = "eine ungerade Zahl werfen". Da A und B unvereinbar sind, berechnen wir die Wahrscheinlichkeit, entweder A oder B zu erhalten, als P(A ∪ B) = P(A) + P(B). Da P(A) = 1/2 und P(B) = 1/2 gilt, dass P(A ∪ B) = 1/2 + 1/2 = 1.

Übungsaufgabe (Chevalier de Méré, Pascal)

In einem Würfelspiel werden 2 Würfel gleichzeitig geworfen. Der Spieler Chevalier de Méré (C.d.M.) gewinnt das Spiel, wenn innerhalb von 25 Würfen mindestens einmal eine Doppelsechs geworfen wird. Andernfalls gewinnt sein Gegner.

1. Berechnen Sie die Wahrscheinlichkeit, dass eine Doppelsechs in einem Wurf geworfen wird.
2. Berechnen Sie die Wahrscheinlichkeit, dass C.d.M. das Spiel gewinnt.

Übungsaufgabe (Chevalier de Méré, Pascal)

In einem Würfelspiel werden 2 Würfel gleichzeitig geworfen. Der Spieler Chevalier de Méré (C.d.M.) gewinnt das Spiel, wenn innerhalb von 25 Würfen mindestens einmal eine Doppelsechs geworfen wird. Andernfalls gewinnt sein Gegner.

1. Berechnen Sie die Wahrscheinlichkeit, dass eine Doppelsechs in einem Wurf geworfen wird.

Übungsaufgabe (Chevalier de Méré, Pascal)

In einem Würfelspiel werden 2 Würfel gleichzeitig geworfen. Der Spieler Chevalier de Méré (C.d.M.) gewinnt das Spiel, wenn innerhalb von 25 Würfen mindestens einmal eine Doppelsechs geworfen wird. Andernfalls gewinnt sein Gegner.

1. Berechnen Sie die Wahrscheinlichkeit, dass eine Doppelsechs in einem Wurf geworfen wird.
   - Da es insgesamt 6 x 6 = 36 mögliche Ergebnisse gibt und nur 1 davon eine Doppelsechs ist, beträgt die Wahrscheinlichkeit: P(Doppelsechs) = $\frac{1}{36}$

Übungsaufgabe (Chevalier de Méré, Pascal)

In einem Würfelspiel werden 2 Würfel gleichzeitig geworfen. Der Spieler Chevalier de Méré (C.d.M.) gewinnt das Spiel, wenn innerhalb von 25 Würfen mindestens einmal eine Doppelsechs geworfen wird. Andernfalls gewinnt sein Gegner.

2. Berechnen Sie die Wahrscheinlichkeit, dass C.d.M. das Spiel gewinnt.

Übungsaufgabe (Chevalier de Méré, Pascal)

In einem Würfelspiel werden 2 Würfel gleichzeitig geworfen. Der Spieler Chevalier de Méré (C.d.M.) gewinnt das Spiel, wenn innerhalb von 25 Würfen mindestens einmal eine Doppelsechs geworfen wird. Andernfalls gewinnt sein Gegner.

2. Berechnen Sie die Wahrscheinlichkeit, dass C.d.M. das Spiel gewinnt.
   - Die Wahrscheinlichkeit, dass in einem Wurf keine Doppelsechs geworfen wird, ist:
   - P(keine Doppelsechs) = 1 - P(Doppelsechs) = 1 - $\frac{1}{36}$ = $\frac{35}{36}$
   - Die Wahrscheinlichkeit, dass in 25 Würfen keine Doppelsechs geworfen wird, ist:
   - P(keine Doppelsechs)^{25} = $(\frac{35}{36})^{25}$
   - Nun berechnen wir die Gegenwahrscheinlichkeit, dass mindestens einmal eine Doppelsechs geworfen wird: P(min. eine Doppelsechs) = 1 – P(keine Doppelsechs)^{25}= 1 - $(\frac{35}{36})^{25}$ = 50.6%

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

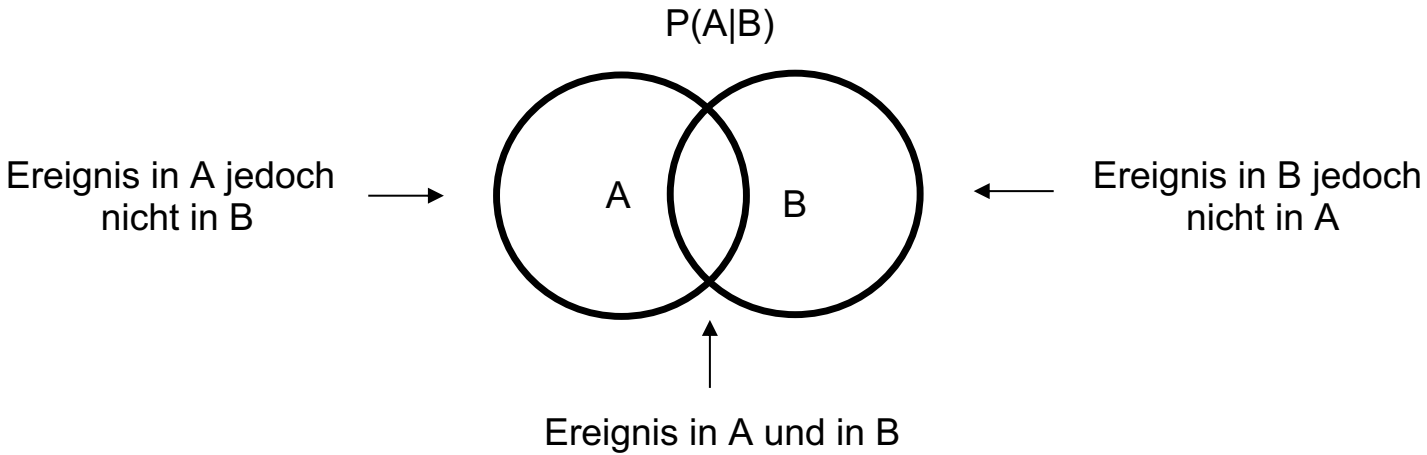
Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Bedingte Wahrscheinlichkeit

- Konzeptuell: Wahrscheinlichkeit P(A|B), dass Ereignis A eintritt, gegeben, dass Ereignis B bereits eingetreten ist
- Diverse Anwendungsfälle aus der Praxis
  - Diagnose-Tests, Fehlerraten, Wettervorhersagen, Risikobewertung
- Visualisierung durch Überlappende Kreise in Venn-Diagrammen



Bedingte Wahrscheinlichkeit

- Mathematisch: $P_B(A)$:= P(A|B) = P(A ∩ B) / P(B), falls P(B) > 0 und A ∩ B ≠ ∅
  - Falls A ⊇ B: P(A|B) = 1
  - Falls A ⊆ B: P(A|B) = P(A) / P(B)
- Unabhängige Ereignisse A und B
  - P(A|B) = P(A) und P(B|A) = P(B)
  - Negierte Kombinationen $(\overline{A}, B)$, (A, $\bar{B}$), $(\overline{A}, \bar{B})$ auch jeweils unabhängig
  - P(A ∩ B) = P(A) * P(B)
- Mathematische Beziehungen
  - Whrkt., dass A eintritt, wenn B und C eingetreten sind: $P_B(A|C)$ = P(A|B ∩ C)
  - Satz der totalen Wahrscheinlichkeit: P(B) = $\sum_{i=1}^{\infty} P(B|A_i) * P(A_i)$, für A_i > 0

Beispiel: Bedingte Wahrscheinlichkeit

- Gut gemischtes Kartenspiel mit 52 Karten
  - 13 Karten jeder Farbe (Kreuz, Pik, Herz, Karo)
  - 3 Bildkarten jeder Farbe (Bube, Dame, König)
- Zwei Ereignisse A und B
  - A: Ziehen einer Bildkarte
  - B: Ziehen einer Herz-Karte
- Bedingte Wahrscheinlichkeit $P(A|B)$, eine Bildkarte zu ziehen, wenn man eine Herz-Karte zieht
  - $P(A|B) = P(A \cap B) / P(B)$
  - Berechnung von $P(A \cap B)$: 3 Bildkarten in Herz (Bube, Dame, König)
  - Berechnung von $P(B)$: 13 Herz-Karten
  - $P(A|B) = 3 / 13 \approx 23.08\%$
- Sind Ereignisse A und B unabhängig?
  - $P(A) = 12/52 = 3 / 13$, $P(B) = 13 / 52 = 1 / 4$
  - $P(A) = P(A|B)$, $P(B) = 3 / 12 = P(B|A)$
  - Ja, A und B sind unabhängig!

Übungsaufgabe: Sensitivität und Spezifität

Sensitivität ist die Fähigkeit eines Tests, *wahre Positive korrekt zu identifizieren*, während Spezifität die Fähigkeit eines Tests ist, *wahre Negative korrekt zu erkennen*. Wir bezeichnen "+" als Krankheit und "-" gesund. Angenommen, es sei bekannt, dass

P(getestet +|+) = 80% (Sensitivität)
P(getestet +|-) = 15% (1 - Spezifität)
P(+) := 10%

1. Wie hoch ist die Wahrscheinlichkeit, dass der Test eine korrekte Diagnose macht?

2. Wie hoch ist die Wahrscheinlichkeit, als krank getestet zu werden?

Übungsaufgabe: Sensitivität und Spezifität

Wir bezeichnen “+” als Krankheit und “-” gesund. Angenommen, es sei bekannt, dass

P(getestet +|+) = 80% (Sensitivität)

P(getestet +|-) = 15% (1 - Spezifität)

P(+) := 10%

1. Wie hoch ist die Wahrscheinlichkeit, dass der Test eine korrekte Diagnose macht?

Übungsaufgabe: Sensitivität und Spezifität

Wir bezeichnen “+” als Krankheit und “-” gesund. Angenommen, es sei bekannt, dass

P(getestet +|+) = 80% (Sensitivität)

P(getestet +|-) = 15% (1 - Spezifität)

P(+) := 10%

1. Wie hoch ist die Wahrscheinlichkeit, dass der Test eine korrekte Diagnose macht?

Lösung:

- P(korrekte Diagnose) = P(getestet -|-)P(-) + P(getestet +|+)P(+)
- = 0.85 * 0.9 + 0.8 * 0.1 = 84.5%

Übungsaufgabe: Sensitivität und Spezifität

Wir bezeichnen “+” als Krankheit und “-” gesund. Angenommen, es sei bekannt, dass

P(getestet +|+) = 80% (Sensitivität)

P(getestet +|-) = 15% (1 - Spezifität)

P(+) := 10%

2. Wie hoch ist die Wahrscheinlichkeit, als krank getestet zu werden?

Übungsaufgabe: Sensitivität und Spezifität

Wir bezeichnen “+” als Krankheit und “-” gesund. Angenommen, es sei bekannt, dass

P(getestet +|+) = 80% (Sensitivität)

P(getestet +|-) = 15% (1 - Spezifität)

P(+) := 10%

2. Wie hoch ist die Wahrscheinlichkeit, als krank getestet zu werden?

Lösung:

- P(getestet +) = P(getestet +|+)P(+) + P(getestet +|-)P(-)
- = 0.8 * 0.1 + 0.15 * 0.9 = 21.5%

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Beispiel: Themenklassifikation

Beispiel: Themenklassifikation (~ topic classification)

- Klassen: Politik, Wirtschaft, Gesellschaft, Kultur, Sport, ….

Keimbelastete Wurst

Ikea nimmt Wilke-Wurstaufschnitt aus Sortiment

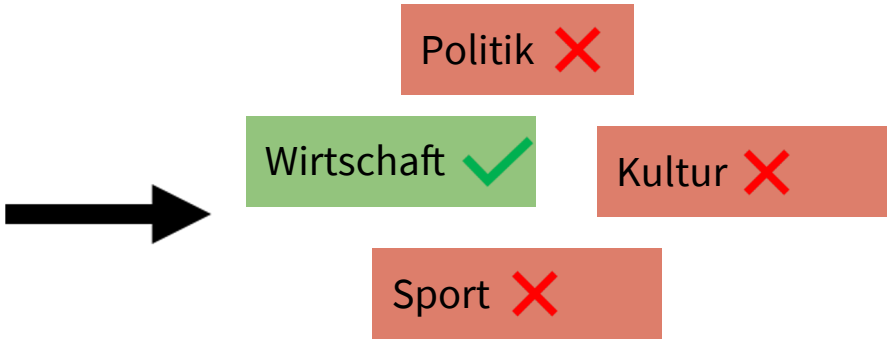
Bundesweit ergreifen Händler wegen keimbelasteter Produkte eines hessischen Herstellers Vorsichtsmaßnahmen. Verbraucherschützer werfen den Behörden Versäumnisse vor.

7. Oktober 2019, 16:59 Uhr / Quelle: ZEIT ONLINE, dpa, AFP, tst / 80 Kommentare

Nach zwei Todesfällen durch keimbelastete Fleischwaren des nordhessischen Wurstproduzenten Wilke ist auch der Möbelkonzern Ikea vom Rückruf betroffen. Über einen Großhändler habe Ikea Deutschland Wurstaufschnitt für Kunden- und Mitarbeiterrestaurants von diesem Hersteller erhalten, sagte eine Sprecherin und bestätigte damit Angaben der Verbraucherorganisation Foodwatch.

Ikea war nach eigenen Angaben am

https://www.zeit.de/wirtschaft/unternehmen/2019-10/keimbelastete-wurst-wilke-ikea-verkaufsstopp



Textklassifikation

Ein Text ist eine beliebige Folge von Token/Wörtern

- Typisch: Bücher, wissenschaftliche Artikel, Nachrichten, E-Mails, Briefe, ...
- Untypisch: Tweets, Berichte mit Bildern und Tabellen, gesprochene Sprache, ...

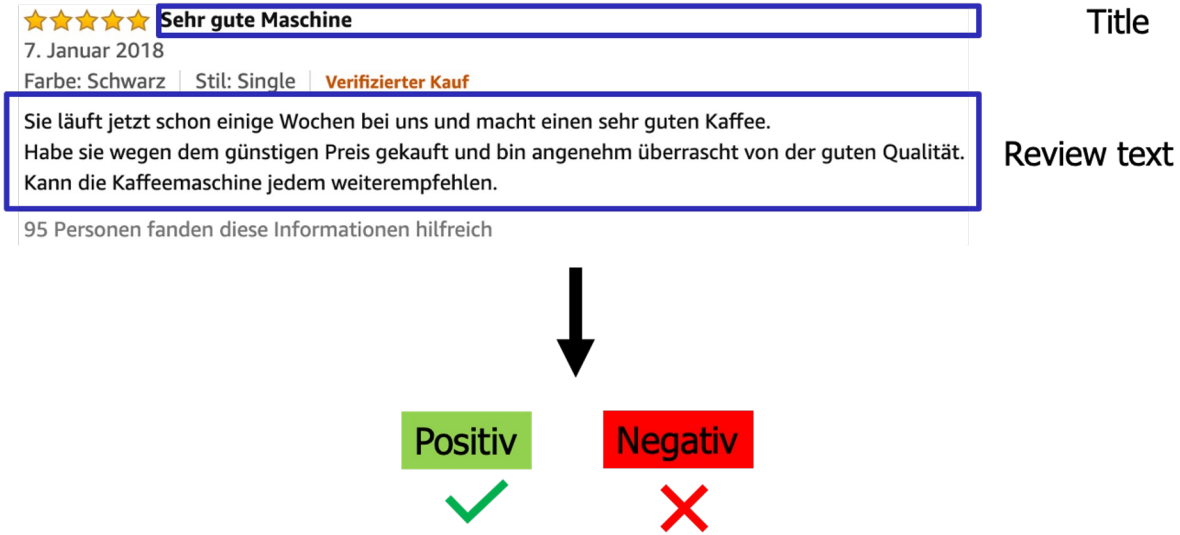
Aufgabe: Zuordnung jedes Textes zu einer Klasse (aus einer Menge an vorgegebenen Klassen)

- Topic identification
- Language identification
- Spam detection
- Sentiment analysis
- Hate speech detection
- Content-based messaging routing
- Author identification
- …

Sentiment Analysis

Analyse der Meinungs- und Stimmungsbildes in einem Text

- Binär: Ist der Text grundlegend positiv oder negativ?
- Multi-Class: Vorhersage der Sternbewertung oder einer anderen Skala



Naive Bayes Textklassifikator

Grundidee: Verwendet Häufigkeiten der Wörter (~ Merkmale) in den verschiedenen Klassen

Gegeben:

- Menge S von Dokumenten und Menge von Klassen C={c_1, c_2, ... c_m}
- Dokumente werden als Menge von Wörtern t_1, …, t_n dargestellt

Wir suchen p(c|d), die Wahrscheinlichkeit, dass ein Dokument d∈S ein Mitglied der Klasse c ist

$$p(c|d) = p(c|t_1, t_2, ..., t_n)$$

Umformulierung und Anwendung des Bayes-Theorems

$$p(c|d) = p(c|t_1, t_2, ..., t_n) = \frac{p(c) \cdot p(t_1, t_2, ..., t_n|c)}{p(t_1, t_2, ..., t_n)} \propto p(c) \cdot p(t_1, t_2, ..., t_n|c)$$

Naive Bayes Textklassifikator

Wir haben für p(c|d) folgendes Modell entwickelt:

$$p(c|d) \propto p(c) \cdot p(t_1, t_2, ..., t_n|c)$$

Der zweite Term auf der rechten Seite kann nicht genau gelernt werden (mit jeder einigermaßen großen, realistischen Trainingsmenge)

- Es gibt 2^n Kombinationen von (binären) Merkmalswerten!

"Naive" Lösung: Statistische Unabhängigkeit der Terme voraussetzen

$$p(t_1, t_2, ..., t_n|c) = p(t_1|c) \cdot p(t_3|c) \cdot ... \cdot p(t_n|c)$$

$$p(c|d) \propto p(c) \cdot \prod_{i}^{n} p(t_i|c)$$

Varianten der Schätzungen

Zum Schätzen der Wahrscheinlichkeit p(t) existieren unterschiedliche Herangehensweisen:

- Anteil der Dokumente in S die t enthalten
- Häufigkeit von t in Dokumenten von S in Relation zur Summe der Häufigkeit aller (anderen) Wörter
- …

Analog zum Schätzen der bedingten Wahrscheinlichkeit p(t|c):

- Anteil der Dokumente in c die t enthalten
- Häufigkeit von t in Dokumenten der Klasse c in Relation zur Summe der Häufigkeiten aller anderen Wörter in Dokumenten von c
- …

Glättung

Bei der Klassifizierung eines neuen, ungesehenen Textdokuments kann es vorkommen, dass dieses Dokument ein Wort t’ enthält, welches wir in den Trainingsdaten noch nicht gesehen haben

- Für dieses Wort t’ haben wir keine Schätzungen für p(t’) oder p(t’|c)

Anpassung der Wahrscheinlichkeitsverteilung bzw. -schätzung notwendig

- Wir vergeben auch eine gewisse Wahrscheinlichkeit an Ereignisse bzw. Wörter, welche wir bisher noch nicht gesehen haben
- Dies wird auch als Glättten (engl. Smoothing) der Verteilung bezeichnet

Glättung

Möglichkeiten zum Umgang mit einem neuen, bisher ungesehenen Wort t':

- Ignorieren des Wortes (nicht gut!)
- Annahme einer konstanten (i.d.R.) sehr kleinen Wahrscheinlichkeit für p(t') und p(t'|c)
  - Konstante kann auch in Abhängigkeit der Menge an Dokumenten S oder der Anzahl der verschiedenen Wörtern von S gewählt werden

- Laplace-Glättung (auch Additive Smoothing):
  - Addition einer Konstanten λ zu jeder Worthäufigkeit
    - Worthäufigkeit von t in Klasse c mit s aus insgesamt |S| Dokumenten: $\frac{\#_c(t)+\lambda}{s+\lambda\cdot|S|}$
    - Worthäufigkeit von t in Klasse c mit m beobachteten aus n möglichen Wörtern: $\frac{\#_c(t)+\lambda}{m+\lambda\cdot n}$
  - Ungesehene Wörter "erhalten" dann eine Häufigkeit von λ
  - Konstante λ muss bei der Berechnung der Summe aller Häufigkeiten beachtet werden

Naive Bayes: Sentiment Analysis Beispiel

- Nehmen wir an, wir haben 100 Reviews, die wir in die Klassen „positiv“ und „negativ“ eingeteilt haben (60% der Reviews sind positiv, 40% sind negativ). Nun wollen wir ein neues Review „Sehr gute Maschine!“ mit Naive Bayes klassifizieren.

- A-priori Wahrscheinlichkeiten:
  - Klassen: P(positiv) = $\frac{60}{100}$, P(negativ) = $\frac{40}{100}$
  - Wörter: P(Sehr) = $\frac{30}{100}$, P(gute) = $\frac{40}{100}$, P(Maschine) = $\frac{90}{100}$ (Beispiel-Werte!)
- Bedingte Wahrscheinlichkeiten P(Wort|Klasse) + Addiere-1 Laplace-Glättung:
  - Klasse positiv: P(Sehr|positiv) = $\frac{15+1}{60+100}$, P(gute|positiv) = $\frac{30+1}{60+100}$, P(Maschine|positiv) = $\frac{55+1}{60+100}$
  - Klasse negativ: P(Sehr|negativ) = $\frac{15+1}{40+100}$, P(gute|negativ) = $\frac{10+1}{40+100}$, P(Maschine|negativ) = $\frac{35+1}{40+100}$
- Klassenzugehörigkeiten P(Klasse|“Sehr gute Maschine“):
  - P(positiv|“Sehr gute Maschine“) = $\frac{60}{100} \cdot \frac{16}{160} \cdot \frac{31}{160} \cdot \frac{56}{160} \approx 0.004$
  - P(negativ|“Sehr gute Maschine“) = $\frac{40}{100} \cdot \frac{16}{140} \cdot \frac{11}{140} \cdot \frac{36}{140} \approx 0.0009$
- Dementsprechend würden wir dem Review „Sehr gute Maschine“ die Klasse positiv zuordnen

Agenda

Musterlösungen von Übungsblatt 1

Übungsblatt 2

Grundlagen der Wahrscheinlichkeitsrechnung

Bedingte Wahrscheinlichkeit

Naive Bayes

Übungstermine

- Es gibt über das Semester 7 Übungen mit verschiedenen Themen aus der VL
- Struktur: Vorstellung Übungsblätter, Besprechung von Musterlösungen, Beispiel-Aufgaben

Wochen	Start	Ende	Thema
16-17	15.04	26.04	Einführung
18-19	29.04	10.05	1. Aufgabenblatt
20-21	**13.05**	**24.05**	**2. Aufgabenblatt**
22-23	27.05	07.06	3. Aufgabenblatt
24-25	10.06	21.06	4. Aufgabenblatt
26-27	24.06	05.07	5. Aufgabenblatt
28-29	08.07	19.07	Klausurvorbereitung