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

Maximum-Likelihood-Schätzung & Parametrische Klassifikation

Thematische Einordnung und Lernziele

- Thematische Einordnung:
  - Datenzentrierte Informatik → Maschinelles Lernen
- notwendige Vorkenntnisse:
  - Grundlagen der Wahrscheinlichkeitstheorie
- Lernziele:
  - Maximum-Likelihood-Methode
    - Maximum-Likelihood-Methode (MLE)
    - MLE für verschiedene Verteilungen (Bernoulli-Verteilung, Multinomialverteilung, Normalverteilung)
  - Parametrische Klassifikation
    - Diskriminanzfunktionen
    - Diskriminanzfunktion für Normalverteilung
- Vorlesung basiert auf Kapitel 4 in [1]

Verteilungen an Daten anpassen: Maximum-Likelihood-Methode

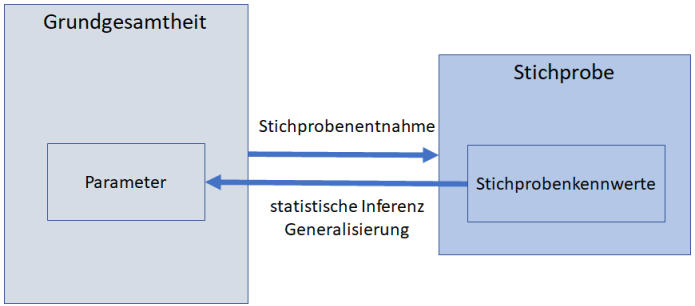
Maximum-Likelihood-Methode

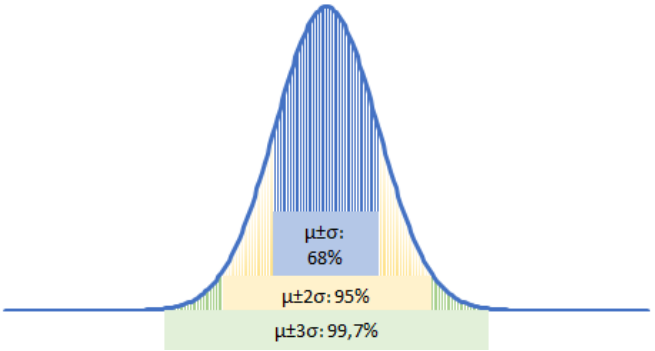
Maximum-Likelihood-Methode (engl. Maximum Likelihood Estimation, MLE) *(R. A. Fisher, 1922)*

- Statistische Methode zur Schätzung von Modellparametern
- **Ziel**:
  - Parameter θ eines Modells so schätzen, dass die beobachteten Daten am wahrscheinlichsten sind
- **Input**:
  - Stichprobe $\mathcal{D} = \{x^1, ..., x^n\}$ mit beobachteten Messungen
  - Modell $p(x|\theta)$ mit unbekannten Parametern θ
- **MLE-Grundidee**:
  - Wähle die Parameter θ so, dass die Wahrscheinlichkeit ("Likelihood") der beobachteten Daten $\mathcal{D}$ maximiert wird

Maximum-Likelihood-Methode: Intuition

- **Beispiel**: eine Stichprobe enthält die Messungen $\mathcal{D} = \{11, 8, 12, 9, 10\}$
- Wir nehmen eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ an
  - Erwartungswert μ und Varianz σ^2 sind unbekannt
- Wir schätzen die Parameter der Verteilung (μ und σ^2)
  - Es ist unwahrscheinlich, dass $\mu = 20$ diese Stichprobenwerte erklärt
  - Welcher μ-Wert wäre plausibler?

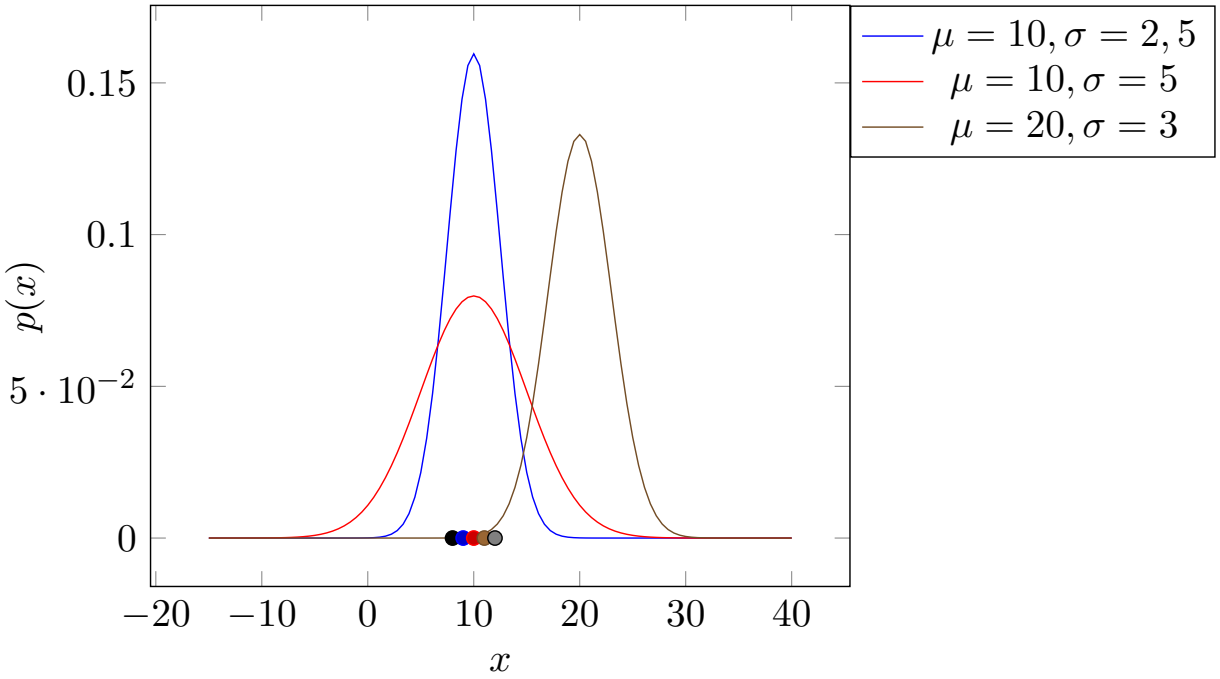




Maximum-Likelihood-Methode: Intuition

- Stichprobe $\mathcal{D} = \{11, 8, 12, 9, 10\}$, Normalverteilung $\mathcal{N}(\mu, \sigma^2)$
  - Zu welchen Parameterwerten gehört diese Stichprobe?

Wahrscheinlichkeitsdichtefunktionen, $\mathcal{N}(\mu, \sigma^2)$



Maximum-Likelihood-Schätzung: Beispiel

Beispiel: MLE in der E-Mail-Klassifikation

- **Ziel**: Schätzung der Wahrscheinlichkeit p, dass eine E-Mail-Nachricht Werbung enthält
- **Vorgehen**:
  - Wähle eine zufällige Stichprobe $\mathcal{D}$
    - Beispiel: $N = 30$ E-Mail-Nachrichten
  - Überprüfe die Nachrichten in $\mathcal{D}$
    - Beispiel: Wir identifizieren 5 Werbenachrichten ($x^i = 1$)
- Wie lässt sich der Anteil der Werbenachrichten in der gesamten E-Mail-Kollektion schätzen?
  - Wir nehmen Bernoulli-Verteilung an
    - Wahrscheinlichkeitsfunktion:

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

  - MLE-Schätzer für p bei der Bernoulli-Verteilung:

$$\hat{p} = \frac{\sum_i x^i}{N} = \frac{5}{30} = 0,17$$

Maximum-Likelihood-Schätzung

Ziel: eine parametrische Verteilung finden, die die gegebene Stichprobe "gut erklärt"

- **Input**:
  - eine unabhängige und identisch verteilte (iid) Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$
  - wir nehmen an, dass die Instanzen x^i aus einer bekannten Wahrscheinlichkeitsdichte $p(x|\theta)$ stammen, die bis auf die Parameter θ definiert ist:

$$x^i \sim p(x|\theta)$$

- **Maximierung der Likelihood**:
  - durch Variieren von θ erhalten wir unterschiedliche Wahrscheinlichkeitsdichten $p(x|\theta)$
  - finde den Parameterwert $\hat{\theta}$, der die Wahrscheinlichkeit der Stichprobe $\mathcal{D}$ maximiert

Likelihood-Funktion

- Likelihood-Funktion $l(\theta|\mathcal{D})$: Likelihood der Stichprobe $\mathcal{D}$ als Funktion der Parameter θ
  - Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$ ist gegeben
  - Parameter θ müssen geschätzt werden
- Likelihood der gesamten iid-Stichprobe $\mathcal{D}$ als Produkt der einzelnen Datenpunkte:

$$l(\theta|\mathcal{D}) \equiv p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^i|\theta)$$

- Maximum-Likelihood-Schätzer für θ: Wir suchen den $\hat{\theta}$-Wert, der die Likelihood für die Stichprobe $\mathcal{D}$ maximiert:

$$\hat{\theta} = \arg\max_{\theta} l(\theta|\mathcal{D})$$

  - $\hat{\cdot}$ (Zirkumflex) kennzeichnet einen Schätzer

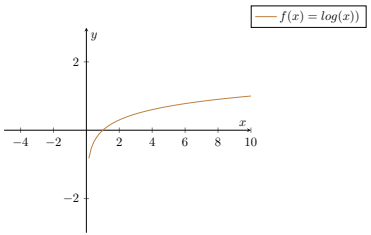
Log-Likelihood-Funktion

Log-Likelihood-Funktion (engl. log likelihood function), $\mathcal{L}(\theta|\mathcal{D})$

- Logarithmus der Likelihood, definiert als:

$$\mathcal{L}(\theta|\mathcal{D}) \equiv log\ l(\theta|\mathcal{D}) = \sum_{i=1}^{N} log\ p(x^i|\theta)$$

- $log(\cdot)$ wandelt das Produkt in eine Summe um
  - monotone Transformation, Maximum bleibt unverändert
  - Vorteil: vereinfacht die Berechnungen



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

Maximum-Likelihood-Schätzer

- Maximum-Likelihood-Schätzer für θ:
  - Gegeben eine Stichprobe $\mathcal{D}$, suchen wir den Parameterwert $\hat{\theta}$, der die Log-Likelihood-Funktion $\mathcal{L}(\theta|\mathcal{D})$ maximiert
- Vorgehen:
  - Log-Likelihood-Funktion $\mathcal{L}(\theta|\mathcal{D})$ aufstellen
  - $\mathcal{L}(\theta|\mathcal{D})$ nach den Parametern θ ableiten
    - (bei mehreren Parametern partielle Ableitungen je Parameter)

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0$$

  - Die Gleichung nach θ auflösen, um den Schätzwert $\hat{\theta}$ zu erhalten
    - verifizieren, ob bei der Nullstelle ein Maximum vorliegt (2. Ableitung überprüfen)
- Wir betrachten MLE-Schätzer für einige Verteilungen
  - Bernoulli-Verteilung
  - Multinomialverteilung
  - Normalverteilung

Bernoulli-Verteilung

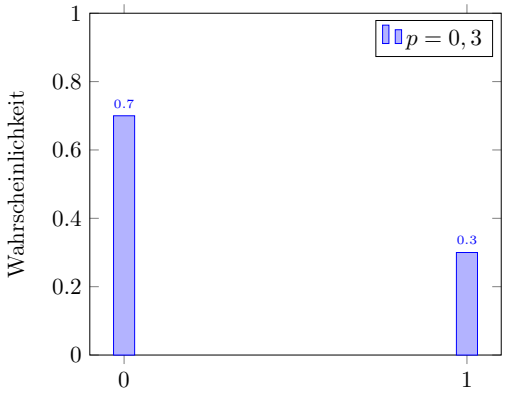
- Bernoulli-Verteilung (engl. Bernoulli distribution) beschreibt ein Zufallsexperiment mit genau zwei möglichen Ergebnissen:

$$X = \begin{cases} 1, & \text{Ereignis tritt ein, mit Wahrscheinlichkeit } p \\ 0, & \text{Ereignis tritt nicht ein, mit Wahrscheinlichkeit } 1-p \end{cases}$$

  - **Beispiel**: Klassifikation einer E-Mail-Nachricht x^i

$$X = \begin{cases} 1, & x^i \text{ ist Werbung} \\ 0, & x^i \text{ ist keine Werbung} \end{cases}$$

  - **Ziel**: Schätzung des Parameters p der Bernoulli-Verteilung gegeben einer Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$

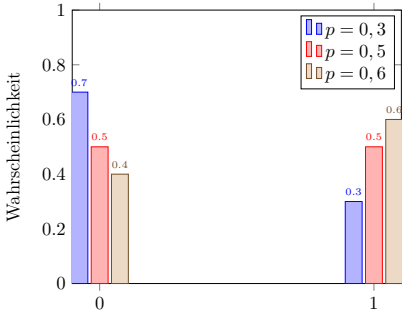


Bernoulli-Verteilung

- Parameter p der Bernoulli-Verteilung ist die Wahrscheinlichkeit, dass das Ergebnis $X = 1$ eintritt:
  - $P(X = 1) = p$
  - $P(X = 0) = 1 - p$: Wahrscheinlichkeit des Nichteintretens
- Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung:

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

  - Erwartungswert: $E[x] = p$
  - Varianz: $Var(x) = p \cdot (1 - p)$
- Beispiele von Verteilungen bei verschiedenen p-Werten



Maximum-Likelihood-Schätzer für Bernoulli-Verteilung

- **Ziel**: Schätzung $\hat{p}$ des Parameters p der Bernoulli-Verteilung mithilfe der Log-Likelihood-Funktion
  - Wahrscheinlichkeitsfunktion der Bernoulli-Verteilung:
$$P(x) = p^x(1-p)^{1-x}$$
  - $\mathcal{D} = \{x^i\}_{i=1}^N$: eine iid-Stichprobe
- Log-Likelihood-Funktion $\mathcal{L}(p|\mathcal{D})$ für Bernoulli-Verteilung:

$$\mathcal{L}(p|\mathcal{D}) = log \prod_{i=1}^{N} p^{x^i} \cdot (1-p)^{1-x^i}$$

$$= \sum_i x^i \cdot log(p) + \left(N - \sum_i x^i\right) \cdot log(1-p)$$

- Interpretation:
  - $\sum_i x^i$ zählt, wie oft $X = 1$ in der Stichprobe $\mathcal{D}$ vorkommt
  - $N - \sum_i x^i$ zählt, wie oft $X = 0$ vorkommt

Rechenregeln:

$log(u \cdot v) = log(u) + log(v); log(u^v) = v \cdot log(u)$

Maximum-Likelihood-Schätzer für Bernoulli-Verteilung

- Log-Likelihood-Funktion $\mathcal{L}(p|\mathcal{D})$ für Bernoulli-Verteilung:

$$\mathcal{L}(p|\mathcal{D}) = \sum_i x^i \cdot log(p) + \left(N - \sum_i x^i\right) \cdot log(1-p)$$

- Um den Wert $\hat{p}$ für die Maximierung der $\mathcal{L}(p|\mathcal{D})$ zu finden, wird $\mathcal{L}$ nach dem Parameter p differenziert und nach p aufgelöst:

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\sum_i x^i}{p} - \frac{N}{1-p} + \frac{\sum_i x^i}{1-p} = 0$$

  - Wie kann man überprüfen, ob es sich bei dieser Nullstelle um ein Maximum handelt?

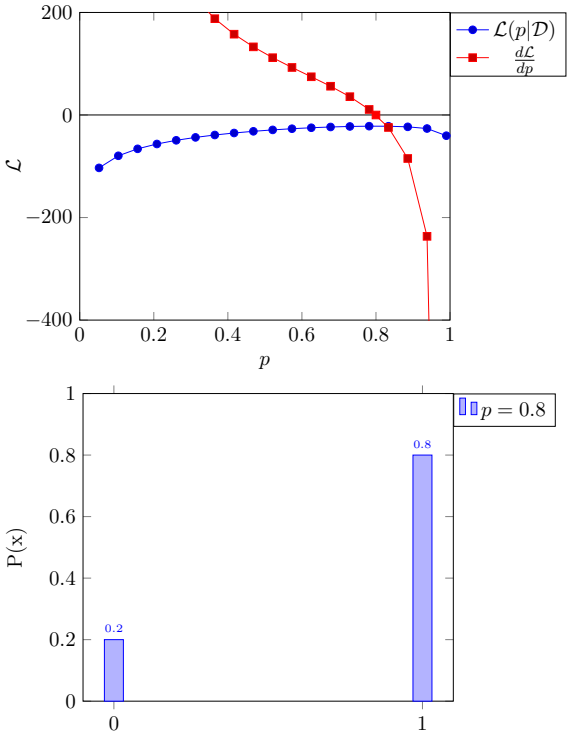
- Daraus ergibt sich die Schätzung für p: $\boxed{\hat{p} = \frac{\sum_i x^i}{N}}$
  - Interpretation: Schätzung für p ist das Verhältnis der Anzahl der Ereignisse $X = 1$ zur Stichprobengröße N
  - In dem Beispiel: enthalten 17% der E-Mails in der Stichprobe Werbung, ist $\hat{p} = 0,17$

Rechenregeln: $log(x)' = \frac{1}{x}; g(h(x))' = g'(h(x)) \cdot h'(x)$

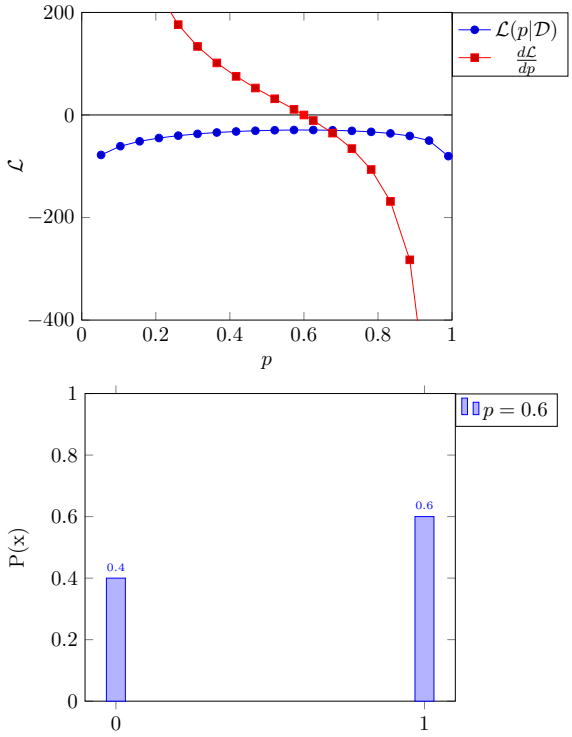
Maximum-Likelihood-Methode: Bernoulli-Verteilung

- Bernoulli-Verteilung, Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$, $x \in \{0, 1\}$

- $N = 100$, $\sum_i x^i = 80$



- $N = 100$, $\sum_i x^i = 60$



Multinomialverteilung

- Multinomialverteilung (engl. multinomial distribution): eine Verallgemeinerung der Bernoulli-Verteilung auf $K \geq 2$ mögliche, sich gegenseitig ausschließende Zustände
  - **Zufallsexperiment**: Ein Ereignis tritt in genau einem von K möglichen Zuständen auf
    - Beispiel: Klassifikation von Dokumenten in K Kategorien (z. B. Nachrichten, Werbung, Spam, usw.)
  - Jeder Zustand k hat eine Wahrscheinlichkeit p_k, wobei gilt:

$$\sum_{k=1}^{K} p_k = 1$$

  - **Indikatorvariablen**: $x_1, x_2, \ldots, x_K$, definiert durch:

$$x_k = \begin{cases} 1, & \text{wenn Zustand } k \text{ eintritt} \\ 0, & \text{sonst} \end{cases}$$

  - Bei einem Experiment: $P(x_1, x_2, ..., x_K) = \prod_{k=1}^{K} p_k^{x_k}$

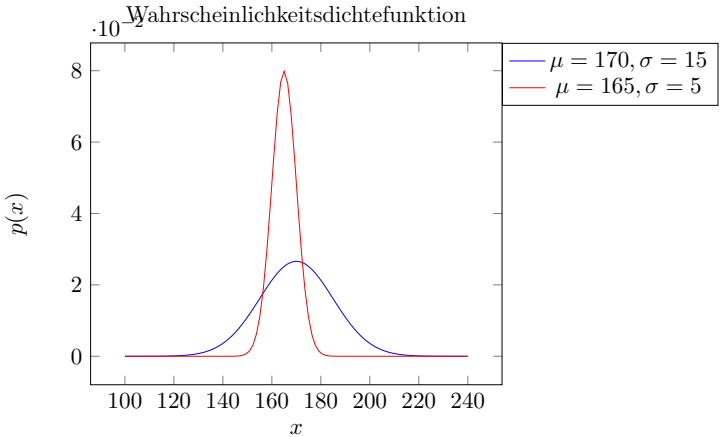
MLE für die Multinomialverteilung

- Wir führen N unabhängige Experimente mit den Ergebnissen $\mathcal{D} = \{\mathbf{x}^i\}_{i=1}^N$ durch, wobei:
  - $x_k^i = \begin{cases} 1, & \text{falls im Experiment } i \text{ der Zustand } k \text{ eintritt} \\ 0, & \text{sonst} \end{cases}$
  - Es gilt: $\sum_k x_k^i = 1$, d.h., jedes Experiment führt zu genau einem Zustand
- **Maximum-Likelihood-Schätzer für** p_k: $\boxed{\hat{p_k} = \frac{\sum_i x_k^i}{N}}$
- **Interpretation**:
  - $\hat{p}_k$ ist die relative Häufigkeit des Auftretens von Zustand k in der Stichprobe
- **Beispiel**:
  - In einer Trainingsmenge: $\hat{p}_k$ = Anteil der Instanzen der Klasse $\mathcal{C}_k$
  - jeder Zustand k kann als ein separates Bernoulli-Experiment betrachtet werden
    - jedes Experiment fällt in einen von K Zuständen

Normalverteilung (Gauß-Verteilung)

- Eine Normalverteilung (Gauß-Verteilung) (engl. normal (Gaussian) distribution) $\mathcal{N}(\mu, \sigma^2)$ ist eine kontinuierliche Wahrscheinlichkeitsverteilung
  - Sie ist parametrisiert durch:
    - Erwartungswert μ
    - Varianz σ^2
- Dichtefunktion der Normalverteilung:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty$$



Log-Likelihood-Funktion für die Normalverteilung

- Gegeben: Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$ mit $x^i \sim \mathcal{N}(\mu, \sigma^2)$
  - Daten sind unabhängig und identisch verteilt und folgen einer Normalverteilung mit unbekanntem Erwartungswert μ und Varianz σ^2
- Log-Likelihood-Funktion beschreibt die Wahrscheinlichkeit, die beobachtete Stichprobe $\mathcal{D}$ zu erhalten, basierend auf den Parametern μ und σ:

$$\mathcal{L}(\mu, \sigma|\mathcal{D}) = \sum_{i=1}^{N} log\ p(x^i|\mu, \sigma)$$

$$= \sum_{i=1}^{N} log\ \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x^i - \mu)^2}{2\sigma^2}\right)$$

$$= -\frac{N}{2}\ log(2\pi) - N\ log\ \sigma - \frac{\sum_i (x^i - \mu)^2}{2\sigma^2}$$

Rechenregeln:

$log(u \cdot v) = log(u) + log(v); log(\frac{u}{v}) = log(u) - log(v); log(u^v) = v \cdot log(u)$

Maximum-Likelihood-Schätzer für Normalverteilung

- **Ziel**: Bestimme den Maximum-Likelihood-Schätzer (MLE) für die Parameter μ und σ^2 der Normalverteilung gegeben eine Stichprobe $\mathcal{D} = \{x^i\}_{i=1}^N$, wobei $x^i \sim \mathcal{N}(\mu, \sigma^2)$
  - **Vorgehen**: Maximiere die Log-Likelihood-Funktion durch Nullsetzen der partiellen Ableitungen:

$$\mathcal{L}(\mu, \sigma|\mathcal{D}) = -\frac{N}{2}\, log(2\pi) - N\, log\, \sigma - \frac{\sum_i (x^i - \mu)^2}{2\sigma^2}$$

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{\sigma^2} \sum_i (x^i - \mu) = 0$$

    - $\frac{d\mathcal{L}}{d\sigma} = 0$ analog, s. nächste Folie
  - Auflösen nach den Parametern μ und σ^2 auf
- MLE für den Erwartungswert μ ist der empirische Mittelwert, d. h. der Mittelwert der gemessenen Werte x^i in der Stichprobe:

$$\hat{\mu} = \frac{\sum_i x^i}{N} = \overline{x}$$

Maximum-Likelihood-Schätzer für Normalverteilung

- Um den Maximum-Likelihood-Schätzer für die Varianz σ^2 zu bestimmen:
  - berechnen wir die partielle Ableitung der Log-Likelihood-Funktion nach dem Parameter σ:

$$\mathcal{L}(\mu, \sigma|\mathcal{D}) = -\frac{N}{2}\, log(2\pi) - N\, log\, \sigma - \frac{\sum_i (x^i - \hat{\mu})^2}{2\sigma^2}$$

$$\frac{d\mathcal{L}}{d\sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_i (x^i - \hat{\mu})^2 = 0$$

  - und lösen nach σ^2 auf
- MLE für die Varianz σ^2 ist die Stichprobenvarianz, d. h. die mittlere quadratische Abweichung der gemessenen Werte x^i vom empirischen Mittelwert $\overline{x}$:

$$\hat{\sigma}^2 = s^2 = \frac{1}{N} \sum_i (x^i - \overline{x})^2$$

Rechenregeln: $log(x)' = \frac{1}{x}; (x^n)' = n \cdot x^{n-1}$

Maximum-Likelihood-Schätzer für Normalverteilung

- Stichprobenvarianz

$$s^2 = \frac{1}{N} \sum_i (x^i - \overline{x})^2$$

  ist ein verzerrter Schätzer für die Populationsvarianz σ^2, weil er systematisch σ^2 unterschätzt
- Verzerrung des Schätzers: systematische Abweichung des geschätzten Wertes σ^2 von Erwartungswert $E[s^2]$[1]

$$E[s^2] = \frac{N-1}{N} \cdot \sigma^2 \neq \sigma^2$$

- verzerrungsfreier Schätzer der Populationsvarianz:

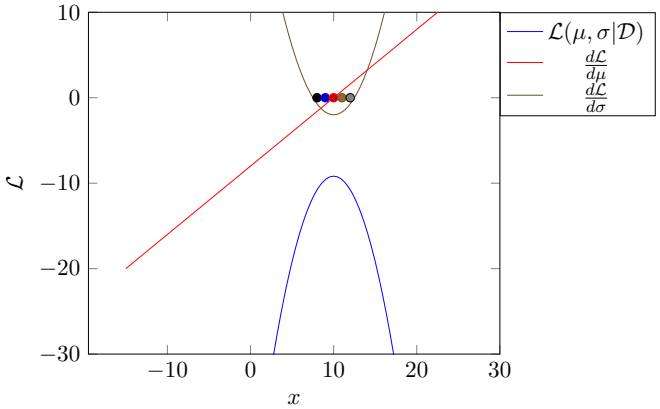
$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1} \sum_i (x^i - \overline{x})^2 = \frac{N}{N-1} s^2$$

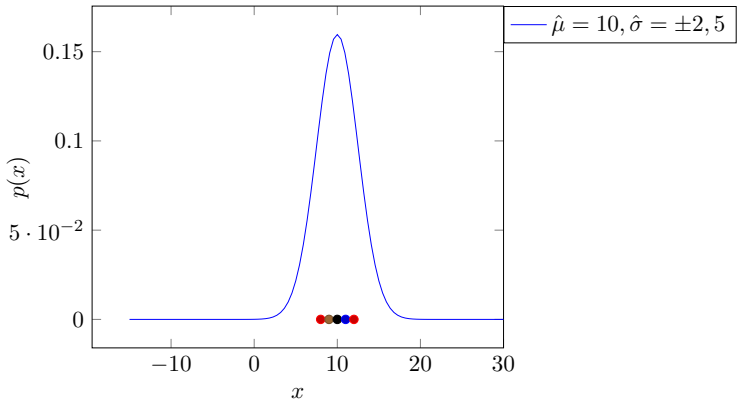
  - Wenn N sehr groß ist, ist der Unterschied unbedeutend
    - d. h. s^2 ist asymptotisch verzerrungsfreier Schätzer

[1] s. Seite 74 in [1] für die Berechnung

Maximum-Likelihood-Methode: Beispiel

- Stichprobe $\mathcal{D} = \{11, 8, 12, 9, 10\}$, Normalverteilung





- MLE für μ: Stichprobenmittelwert

$$\frac{d\mathcal{L}}{d\mu} = \frac{1}{\sigma^2} \sum_i (x^i - \mu) = 0$$

$$\hat{\mu} = \frac{\sum_i x^i}{N} = \overline{x} = 10$$

- MLE für σ^2: Stichprobenvarianz

$$\frac{d\mathcal{L}}{d\sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_i (x^i - \hat{\mu})^2 = 0$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{N} \sum_i (x^i - \overline{x})^2 = 2,5^2$$

Parametrische Klassifikation

Parametrische Klassifikation

Bei der parametrischen Klassifikation (engl. parametric classification) gehen wir davon aus, dass die Daten innerhalb jeder Klasse $\mathcal{C}_k$ einer bestimmten Verteilung folgen, die durch Parameter beschrieben wird

- Schätzung der Parameter:
  - Wir nutzen die Trainingsdaten, um die Parameter für
    - die a-priori-Wahrscheinlichkeiten $P(\mathcal{C}_k)$
    - die Klassen-Likelihoods $p(\mathbf{x}|\mathcal{C}_k)$ zu schätzen
  - dazu wird häufig MLE verwendet
- Berechnung der a-posteriori-Wahrscheinlichkeiten
  - mit dem Satz von Bayes
- Diskriminanten:
  - Für jede Klasse $\mathcal{C}_k$ berechnen wir die Diskriminante
    - a-posteriori-Wahrscheinlichkeiten oder deren Logarithmen
  - Wir wählen die Klasse mit der höchsten Diskriminante

Parametrische Klassifikation: Diskriminanzfunktion

- Die a-posteriori-Wahrscheinlichkeit der Klasse $\mathcal{C}_k$ errechnen wir als:

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

- Die Diskriminanzfunktion:

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

- Äquivalent können wir die Log-Variante nutzen:

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

Diskriminanzfunktion bei Normalverteilung

- Falls $p(x|\mathcal{C}_k)$ einer Normalverteilung unterliegt, dann gilt:
  - Klassen-Likelihood:

$$p(x|\mathcal{C}_k) = \frac{1}{\sigma_k\sqrt{2\pi}} exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

  - Diskriminanzfunktion:

$$\begin{aligned} g_k(x) &= log\ p(x|\mathcal{C}_k) + log\ P(\mathcal{C}_k) \\ &= -\frac{1}{2}\ log\ 2\pi - log\ \sigma_k - \frac{(x-\mu_k)^2}{2\sigma_k^2} + log\ P(\mathcal{C}_k) \end{aligned}$$

- Achtung: Normalität der Verteilung $p(x|\mathcal{C}_k)$ soll zuerst durch einen statistischen Test überprüft werden
  - z. B. Kolmogorov-Smirnov-Test, Chi-Quadrat-Test
  - ansonsten ist $g_k(x)$ in der Form nicht anwendbar!

Parametrische Klassifikation: Anwendungsbeispiel

- Anwendungsbeispiel:
  - Ein Autohaus verkauft $K = 3$ verschiedene Autotypen:
    - $K = \{\mathcal{C}_1 = Kompaktwagen, \mathcal{C}_2 = SUV, \mathcal{C}_3 = Sportwagen\}$
  - zur Vereinfachung nehmen wir an, dass der einzige Faktor für die Kaufentscheidung ist das Jahreseinkommen des Kunden
  - Folgende Daten liegen vor:

	Kunde	**Jahreseinkommen, T. EUR**	**Autotyp**
Trainingsmenge	$Kunde_1$	80	Kompaktwagen
	$Kunde_2$	120	SUV
	$Kunde_3$	200	Sportwagen
	$Kunde_4$	100	SUV
	$Kunde_5$	300	Sportwagen
	$Kunde_6$	70	Kompaktwagen
	$Kunde_7$	75	Kompaktwagen
Testmenge	$Kunde_z$	90	?

Parametrische Klassifikation: Vorgehensweise

Vorgehensweise:

- Überprüfung der Verteilung der Merkmale (Einkommen) in jeder Klasse
  - aufgrund der Überprüfung gehen wir davon aus, dass diese der Normalverteilung folgen
- Modellierung der Klassen-Likelihoods:
  - wir approximieren die Klassen-Likelihoods $p(x|\mathcal{C}_k)$ durch die Normalverteilung $p(x|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \sigma_k^2)$
- Schätzung der Normalverteilungs-Parameter
  - wir schätzen Parameter der Normalverteilung für jede Klasse, $\hat{\mu_k}$ und $\hat{\sigma_k^2}$ anhand der Trainingsdaten $\mathcal{D}$
- Schätzung der a-priori-Wahrscheinlichkeiten
  - anhand der Trainingsdaten $\mathcal{D}$
- Berechnung der Diskriminanzfunktionen der Klassen $g_k(x)$
- Vorhersage: die Klasse, die die höchste $g_k(x)$ für x aufweist

Anwendungsbeispiel: Stichprobe

- Stichprobe: $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{y}^1), (\mathbf{x}^2, \mathbf{y}^2), ..., (\mathbf{x}^n, \mathbf{y}^n)\}$
  - $x \in \mathbb{R}$ ist eindimensionale Eingabe, entspricht dem Jahreseinkommen eines Kunden
  - $\mathbf{y} \in \{0, 1\}^K$ ist ein K-dimensionaler binärer Vektor

$$y_k^i = \begin{cases} 1, & \text{falls } x^i \in \mathcal{C}_k, \\ 0, & \text{falls } x^i \in \mathcal{C}_j, j \neq k \end{cases}$$

Anwendungsbeispiel: Parameterschätzung

- Für jede Klasse $\mathcal{C}_k$ berechnen wir die Schätzungen für den Erwartungswert und Varianz basierend auf MLE für Normalverteilung:
  - Empirischer Mittelwert der Klasse $\mathcal{C}_k$:

$$\hat{\mu}_k = \frac{\sum_i x^i \cdot y_k^i}{\sum_i y_k^i}$$

    - entspricht dem Mittelwert der Jahreseinkommen der Kunden in der Stichprobe, die einen Wagen der Klasse $\mathcal{C}_k$ kaufen
  - Varianz der Klasse $\mathcal{C}_k$:

$$\hat{\sigma}_k^2 = \frac{\sum_i (x^i - \hat{\mu}_k)^2 \cdot y_k^i}{\sum_i y_k^i}$$

    - entspricht der mittleren quadratischen Abweichung der gemessenen Jahreseinkommen-Werte x^i in der Klasse $\mathcal{C}_k$ vom empirischen Mittelwert der Klasse $\hat{\mu}_k$

Beispiel: Schätzung der a-priori-Wahrscheinlichkeiten

- Schätzung der a-priori-Wahrscheinlichkeiten der Klassen:

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

  - entspricht dem Anteil der Kunden in der Stichprobe, die einen Wagen der Klasse $\mathcal{C}_k$ kaufen

Beispiel: Diskriminamzfunktionen

- Wir nutzen die geschätzten Werte für die Schätzung der Diskriminamzfunktionen der Klassen $g_k(x)$

$$g_k(x) = -\frac{1}{2}\ log\ 2\pi - log\ \hat{\sigma}_k - \frac{(x-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2} + log\ \hat{P}(\mathcal{C}_k)$$

  - $-\frac{1}{2}\ log\ 2\pi$ ist eine Konstante, wird weggelassen
  - wenn a-priori-Werte der Klassen gleich sind, wird $log\ \hat{P}(\mathcal{C}_k)$ ebenfalls weggelassen
  - falls auch Varianzen gleich sind, gilt:

$$g_k(x) = -(x-\hat{\mu}_k)^2$$

- unter diesen Annahmen wird x der Klasse mit dem naheliegenden empirischen Mittelwert zugewiesen:

$$\text{wähle } \mathcal{C}_k, \text{ falls } |x-\hat{\mu}_k| = \min_j |x-\hat{\mu}_j|$$

Parametrische Klassifikation: Beispiel

Kompaktwagen			**SUV**			**Sportwagen**		
$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{P}(\mathcal{C}_1)$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{P}(\mathcal{C}_2)$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{P}(\mathcal{C}_3)$
75	4	0,42	110	5	0,29	250	50	0,29

- Welcher Klasse wird ein Kunde mit dem Jahreseinkommen 90 $T.EUR$ zugewiesen?

Parametrische Klassifikation: Beispiel

Kompaktwagen			**SUV**			**Sportwagen**		
$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{P}(\mathcal{C}_1)$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{P}(\mathcal{C}_2)$	$\hat{\mu}_3$	$\hat{\sigma}_3$	$\hat{P}(\mathcal{C}_3)$
75	4	0,42	110	5	0,29	250	50	0,29

- Wir berechnen die Werte der Diskriminanzfunktionen $g_k(x)$ für den Kunden mit dem Einkommen $x = 90$ für jede Klasse

$$g_1(90) = -log\ 4 - \frac{(90-75)^2}{2 \cdot 4^2} + log\ 0,42 = -8,01$$

$$g_2(90) = -log\ 5 - \frac{(90-110)^2}{2 \cdot 5^2} + log\ 0,29 = -9,24$$

$$g_3(90) = -log\ 50 - \frac{(90-250)^2}{2 \cdot 50^2} + log\ 0,29 = -7,36$$

- $\rightarrow$ Der Kunde wird der Klasse $\mathcal{C}_3$ zugewiesen
  - Ist das Ergebnis plausibel? Welche Faktoren führen dazu?

Fazit: Maximum-Likelihood-Schätzung & Parametrische Klassifikation

Die Studierenden sollen in der Lage sein

- Parametrische Klassifikation zu beschreiben und zu diskutieren
- Maximum-Likelihood-Methode zu beschreiben
- die Definition der Likelihood anzugeben
- einen Maximum-Likelihood-Schätzer für eine einfache Dichte analytisch berechnen
  - Bernoulli-Verteilung, Multinomialverteilung, Normalverteilung
- für eine gegebene Stichprobe eine geeignete Verteilung finden und anpassen
- Parametrische Klassifikation 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.