2. Statistische Entscheidungstheorie

2.1 Ziel und Aufgabe

Ziel in Anwendungphase (Arbeitsphase):

- optimale Entscheidungen treffen (Klassifikation).

Aufgabe in der Trainingsphase (Lernphase):

- statistisch auswertbare Messungen/Beobachtungen (Trainingselemente) aus dem Problembereich erfassen, und
- statistisch beschreibbares Wissen über den Problembereich einbeziehen,

um einen Klassifikator zu erwerben.

In der Praxis sollten die Lern- und die Anwendungsphase in einem Zyklus ablaufen.

2.2 Begriffe zur Wahrscheinlichkeit

Überblick:

- Problembereich und Ereignisse
- Wahrscheinlichkeit und Verbundwahrscheinlichkeit
- Einschub: Gesetze der Wahrscheinlichkeit
- Bedingte Wahrscheinlichkeit
- Bayes-Formel
- Stochastische(r) Variable und Prozess
- Vektoren von Zufallsvariablen
- Erwartungswert versus Mittelwert

Problembereich und Ereignisse

$\boldsymbol{\Omega}$: Problembereich.
Umfasst alle möglichen Situationen, Muster, Ereignisse, etc.

$\boldsymbol{\Psi} \subset \boldsymbol{\Omega}$: Stochastisches Ereignis.
Ergebnis eines zufälligen Versuchs.

$\boldsymbol{v} \in \boldsymbol{\Psi}$: Elementarereignis.
Ereignis, das nicht weiter zerlegbar ist.

$$\boldsymbol{\Omega} := \cup_{\boldsymbol{m}} \boldsymbol{v}^{\boldsymbol{m}}$$

Sprechweise:

$\boldsymbol{\Psi}$ ist eingetreten, wenn $\boldsymbol{v} \in \boldsymbol{\Psi}$ eingetreten ist.

Wahrscheinlichkeit und Verbundwahrscheinlichkeit

$\boldsymbol{P(\Psi)}$: Wahrscheinlichkeit für Ereignis $\boldsymbol{\Psi}$.

$\boldsymbol{P(\Psi^m, \Psi^n)}$: Wahrscheinlichkeit, dass zwei Ereignisse $\boldsymbol{\Psi^m, \Psi^n}$ gleichzeitig eingetreten sind, d.h. $\boldsymbol{P(v \in \Psi^m \cap \Psi^n)}$ ist Verbundwahrscheinlichkeit (joint probability).

Äquivalente Schreibweise:

$$\boldsymbol{P(\Psi^m, \Psi^n) = P(\Psi^m \cap \Psi^n)}$$

Einschub: Gesetze der Wahrscheinlichkeit

$$0 \leq P(\psi) \leq 1$$

$$P(\psi \cup \bar{\psi}) = 1, \quad P(\psi \cap \bar{\psi}) = 0$$

$$P(\psi_1 \cup \psi_2) = P(\psi_1) + P(\psi_2) - P(\psi_1 \cap \psi_2)$$

Bedingte Wahrscheinlichkeit

$$\boldsymbol{P(\Psi^m|\Psi^n) := \frac{P(\Psi^m,\Psi^n)}{P(\Psi^n)}}$$

$$\begin{aligned} \boldsymbol{P(\Psi^m|\Psi^n) \cdot P(\Psi^n)} &= \\ &\boldsymbol{P(\Psi^m \cap \Psi^n) = P(\Psi^n \cap \Psi^m)} \\ &\qquad \boldsymbol{= P(\Psi^n|\Psi^m) \cdot P(\Psi^m)} \end{aligned}$$

Bayes-Formel

$$P(\Psi^m|\Psi^n) := \frac{P(\Psi^n|\Psi^m) \cdot P(\Psi^m)}{P(\Psi^n)}$$

Beispiel:

P(Fieber|Grippe)	Schätzung ist einfach
P(Grippe|Fieber)	schwierig schätzbar
P(Grippe)	Existiert eine 'Grippewelle' ?
P(Fieber)	Temperaturmessung

$$P(\text{Grippe}|\text{Fieber}) := \frac{P(\text{Fieber}|\text{Grippe}) \cdot P(\text{Grippe})}{P(\text{Fieber})}$$

Stochastische(r) Variable und Prozess

$\boldsymbol{X} : \boldsymbol{\Omega} \to \mathbb{R}$, $\boldsymbol{X}$ stochastische Variable (Zufallsvariable).

$\boldsymbol{x} := \boldsymbol{X}(\boldsymbol{v})$, Realisierung einer Zufallsvariable.

Realisierungen werden durch Messungen/Beobachtungen erhalten.

Zufallsvariable ist zusätzlich Funktion der Zeit, $\boldsymbol{t} \in \boldsymbol{T}$, weil die Messung selbst unstabil ist.

Neudefinition: $\boldsymbol{X}(\boldsymbol{v}, \boldsymbol{t}) : \boldsymbol{\Omega} \times \boldsymbol{T} \to \mathbb{R}$, stochastischer Prozess.

Vektoren von Zufallsvariablen

- Zum Problembereich erhält man somit via Messungen/ Beobachtungen im Computer eine Datenmenge.
- Realisierung $\boldsymbol{x}$ kann auch ein Vektor der Dimension $\boldsymbol{I}$ sein, dann wird durch $\boldsymbol{X} : \boldsymbol{\Omega} \rightarrow \mathbb{R}^{\boldsymbol{I}}$ ein Vektor von Zufallsvariablen definiert.
- Der Vektor von Zufallsvariablen kann auch zweigeteilt in einen sogenannten Eingabe- und einen Ausgabeteil sein.
  Dies ist relevant für Aufgabenstellungen der Klassifikation oder Regression (siehe später).

Erwartungswert versus Mittelwert

Erwartungswert $\boldsymbol{E}$ einer Zufallsvariable $\boldsymbol{X}$:

$$\boldsymbol{E(X) := \int_{-\infty}^{\infty} x \cdot P(x)dx}$$

Spezialfall: Endlich viele Realisierungen $\boldsymbol{x^1, \ldots, x^M}$ und $\boldsymbol{P(x^m) = \frac{1}{M}}$, $\forall \boldsymbol{m}$.

$$\boldsymbol{E(X) := \frac{1}{M} \sum_{m=1}^{M} x^m} \quad \text{Mittelwert}$$

2.3 Statistischer Klassifikator

Überblick:

- Klassifizierte Trainingsmenge
- Repräsentation von Klassen
- Trainingselemente und Wahrscheinlichkeiten
- Klassenbedingte Verteilungsdichte
- Beispiele zu klassenbedingten Verteilungsdichten
- Ziel eines statistischen Klassifikators

Klassifizierte Trainingsmenge

Gegeben sei eine klassifizierte Trainingsmenge:

$$\{(\boldsymbol{x}^m, \boldsymbol{y}^{k_m}) | \boldsymbol{m} = \mathbf{1}, \ldots, \boldsymbol{M}\}$$

$\boldsymbol{I}$-dimensionaler Messvektor:

$$\boldsymbol{x}^m := (\boldsymbol{x}_1^m, \ldots, \boldsymbol{x}_I^m)^T$$

Klassen des Problembereichs:

$$\boldsymbol{y}^{k_m} \in \{\boldsymbol{c}^k | \boldsymbol{k} = \mathbf{1}, \ldots, \boldsymbol{K}\}$$

Repräsentation von Klassen

1. Beispiel: $\boldsymbol{c}^k := \boldsymbol{k}$;
   Probleme mit Metrik: Gewichte werden 'überladen'

2. Beispiel: $\boldsymbol{c}^k := (\underbrace{\mathbf{0}, \ldots, \mathbf{0}, \mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}}_{\boldsymbol{K}\text{ Komp., Wert }\mathbf{1}\text{ für }\boldsymbol{k}\text{. Komp.}})$

   (1 aus $\boldsymbol{K}$)-Codierung

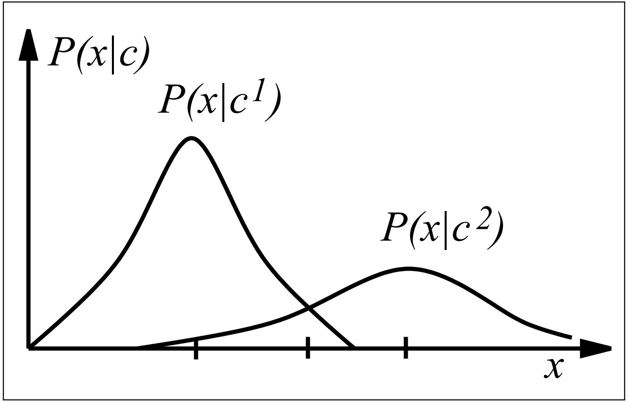
Trainingselemente und Wahrscheinlichkeiten

- Die Menge aller Trainingselemente ist das Ergebnis eines stochastischen Prozesses.
- Die Tupel $(\boldsymbol{x}, \boldsymbol{c}^k)$ sind zufällig aber nicht regellos entstanden.
- Die Regelhaftigkeiten drücken sich durch die Verbundwahrscheinlichkeit $\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{c}^k)$ aus.
- Es gilt: $\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{c}^k) = \boldsymbol{P}(\boldsymbol{x}|\boldsymbol{c}^k) \cdot \boldsymbol{P}(\boldsymbol{c}^k)$

Klassenbedingte Verteilungsdichte

- $\boldsymbol{P(x|c^k)}$ drückt die Unsicherheit der Zugehörigkeit des Trainingselements $\boldsymbol{x}$ zu gegebener Klasse $\boldsymbol{c^k}$ aus.
- Jede Entscheidung beinhaltet ein Risiko, einen Fehler zu begehen.
- Form und Nachbarschaft der klassenbedingten Verteilungsdichten $\boldsymbol{P(x|c^k)}$ wichtig.

Beispiele zu klassenbedingten Verteilungsdichten



Ziel eines statistischen Klassifikators

Bestimmung der Klassenzugehörigkeit von Messvektoren, sodass eine Minimierung der Wahrscheinlichkeit der Fehlklassifikation erreicht wird (minimising a loss function).

2.4 A posteriori Wahrscheinlichkeit

Überblick:

- Bayes-Formel zur Klassifikation
- Maximale a posteriori Wahrscheinlichkeit
- Bestandteile der Bayes-Formel
- Satz der totalen Wahrscheinlichkeit
- Normalisierte a posteriori Wahrscheinlichkeiten

Bayes-Formel zur Klassifikation

Bayes-Formel für die Klassifikation beschreibt, wie bei Beobachtung $\boldsymbol{x}$, unter Einbezug von Erfahrung/Empirie, die a priori Wahrscheinlichkeit einer Klasse $\boldsymbol{c}^k$ in die a posteriori Wahrscheinlichkeit umgewandelt wird.

$$\underbrace{\boldsymbol{P}(\boldsymbol{c}^k|\boldsymbol{x})}_{\text{a posteriori}} := \underbrace{\frac{\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{c}^k)}{\boldsymbol{P}(\boldsymbol{x})}}_{\text{Erfahrung}} \underbrace{\boldsymbol{P}(\boldsymbol{c}^k)}_{\text{a priori}}$$

Bestandteile der Bayes-Formel

$\boldsymbol{P(c^k)}$: A priori Wahrscheinlichkeit der Klasse $\boldsymbol{c^k}$.

$\boldsymbol{P(x|c^k)}$: Klassenbedingte Wahrscheinlichkeit des Trainingselementes $\boldsymbol{x}$ für die Klasse $\boldsymbol{c^k}$ (Klassenbedingte Dichtefunktion).

$\boldsymbol{P(c^k|x)}$: A posteriori Wahrscheinlichkeit dafür, dass $\boldsymbol{x}$ der Klasse $\boldsymbol{c^k}$ zuzuordnen ist.

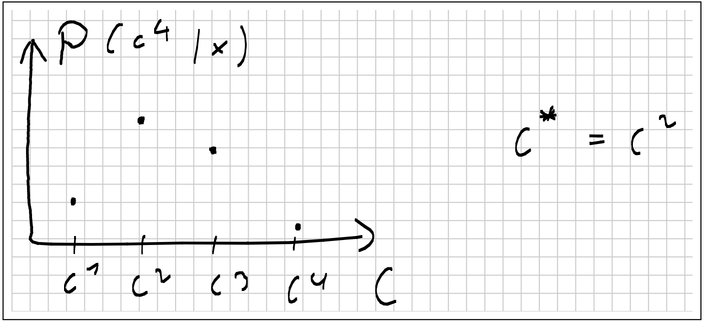
$\boldsymbol{P(x)}$: Wahrscheinlichkeit für das Auftreten von $\boldsymbol{x}$ (Dichtefunktion).

Maximale a posteriori Wahrscheinlichkeit

Die Beobachtung $\boldsymbol{x}$ ist der Klasse $\boldsymbol{c}^k$ zuzuordnen, deren a posteriori Wahrscheinlichkeit maximal ist.

$$\boldsymbol{c}^* := \arg \max_{k \in 1,\ldots,K} \{P(\boldsymbol{c}^k | \boldsymbol{x})\}$$

Beispiel:



Satz der totalen Wahrscheinlichkeit

Unter Verwendung von *Satz der totalen Wahrscheinlichkeit* gilt:

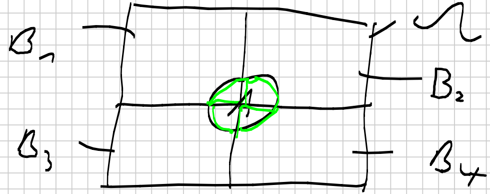
$$\boldsymbol{P(x) = \sum_{k=1}^{K} P(x|c^k) \cdot P(c^k)}$$

Damit kann $\boldsymbol{P(x)}$ als Normierungsfaktor in der Bayes-Formel dienen, so dass gilt:

$$\boldsymbol{\sum_{k=1}^{K} P(c^k|x) = 1}$$

Satz der totalen Wahrscheinlichkeit

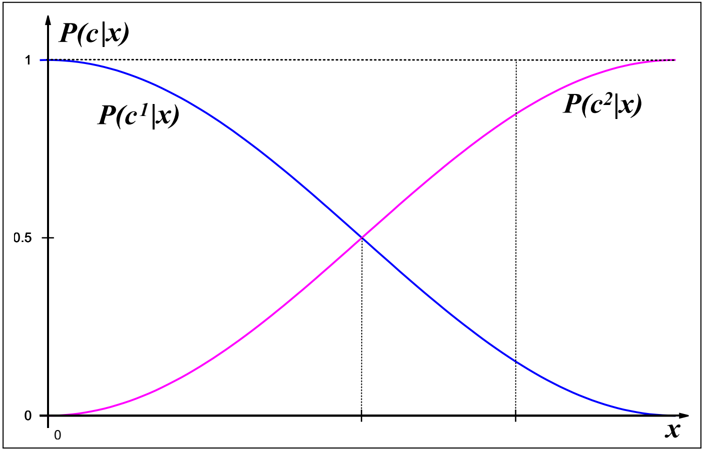
Beispiel:



$$P(A) = \sum_{i=1}^{4} P(A \cap B_i)$$

$$P(A \mid B_i) = \frac{P(A \cap B_i)}{P(B_i)}$$

Normalisierte a posteriori Wahrscheinlichkeiten



2.5 Diskriminanzfunktionen und Merkmals-Teilräume

Überblick:

- Klassenbezogene Diskriminanzfunktionen
- Beispiele für Diskriminanzfunktionen
- Einschub: Logarithmus, Minimum Squared Error, Nearest Neighbor
- Perzeptron Diskriminanzfunktion
- Entscheidungsregionen

Klassenbezogene Diskriminanzfunktionen

Klassifikator wendet $\boldsymbol{K}$ Diskriminanzfunktionen $\boldsymbol{d}^{\boldsymbol{k}}$ an und wählt diejenige Klasse $\boldsymbol{c}^{\boldsymbol{l}}$ mit $\boldsymbol{d}^{\boldsymbol{l}}(\boldsymbol{x}) > \boldsymbol{d}^{\boldsymbol{k}}(\boldsymbol{x})$ für alle $\boldsymbol{k} \neq \boldsymbol{l}$.

Diskriminanzfunktionen sind das Ergebnis eines Lernvorgangs.

Es werden Ungenauigkeiten bei Messungen zugelassen (weil unvermeidlich), und deren Auswirkungen werden bei der Klassifizierung minimiert.

Beispiele für Diskriminanzfunktionen

$$\boldsymbol{d^k := P(c^k|x)} \quad \Rightarrow \quad \text{Maximum a posteriori Wahrscheinlichkeit}$$

$$\boldsymbol{d^k := P(x|c^k)} \quad \Rightarrow \quad \text{Maximum Likelihood}$$

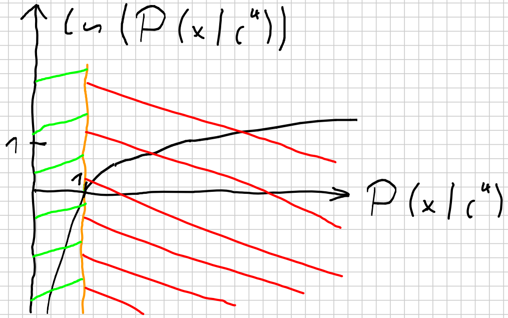
$$\boldsymbol{d^k := \ln P(x|c^k)} \quad \Rightarrow \quad \text{Minimum Squared Error (falls } \boldsymbol{P} \text{ eine Gauss-Verteilung)}$$

$$\boldsymbol{d^k := -R(c^k|x)} \quad \Rightarrow \quad \text{Risikominimierung}$$

Einschub: Logarithmus-Funktion

Beispiel:

$$P(x \mid c^4) \in [0, 1] \subset \mathbb{R}$$



Einschub: Minimum Squared Error

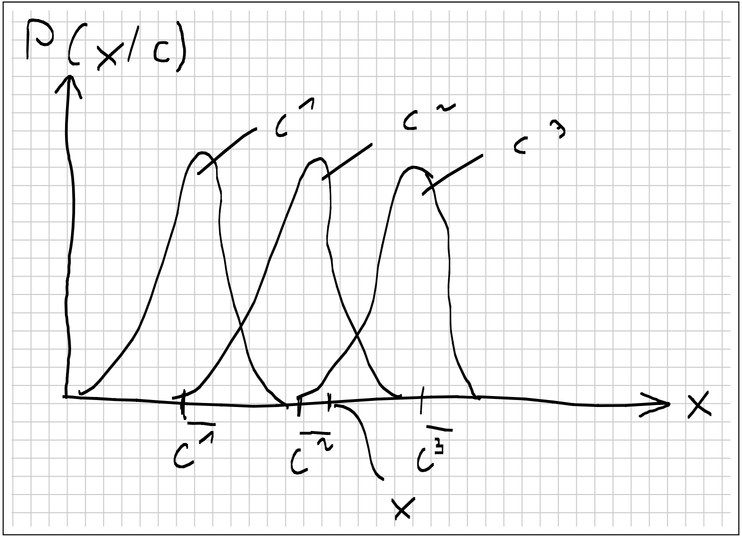
Spezialfall $p(x|c^k) := e^{-(x-\bar{c}^k)^2}$

$$\max_k \{\ln(p(x|c^k))\} = \max_k \{\ln(e^{-(x-\bar{c}^k)^2})\}$$

$$= \max_k \{-(x-\bar{c}^k)^2\} = \min_k \{(x-\bar{c}^k)^2\}$$

nearest neighbor

Einschub: Nearest Neighbor



Perzeptron Diskriminanzfunktion

Spezialfall: 2-Klassen-Aufgabe

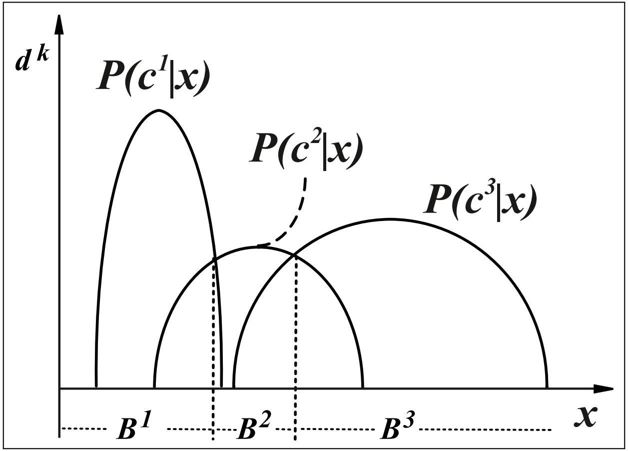
Entscheidung durch Perzeptron (Vorgriff auf Kap.3):

$$\begin{aligned} \boldsymbol{x} \rightarrow \boldsymbol{c}^1 \quad &, \quad \text{wenn } \boldsymbol{d}(\boldsymbol{x}) > \boldsymbol{0} \\ \boldsymbol{x} \rightarrow \boldsymbol{c}^2 \quad &, \quad \text{sonst} \end{aligned}$$

mit $\boldsymbol{d}^1(\boldsymbol{x}) - \boldsymbol{d}^2(\boldsymbol{x}) =: \boldsymbol{d}(\boldsymbol{x}) := \boldsymbol{w}^T \boldsymbol{x} - \boldsymbol{\Theta}$

Entscheidungsregionen

Klassifikator unterteilt Merkmalsraum in Teilräume $B^1, \ldots, B^K$.



2.6 Fehlerwahrscheinlichkeiten

Überblick:

- Wahrscheinlichkeit von Fehlklassifikationen
- Erwartungswert der Fehlerwahrscheinlichkeit
- Illustration der Fehlerwahrscheinlichkeit
- Fehlerwahrscheinlichkeit bei einer Zwei-Klassen-Aufgabe
- Praktisch relevante Evaluationskriterien

Wahrscheinlichkeit von Fehlklassifikationen

Wahrscheinlichkeit der Fehlklassifikation $\boldsymbol{f}$ (Ereignis), wenn ein neuronales Netz den Messvektor $\boldsymbol{x}$ der Klasse $\boldsymbol{c}^k$ zuordnet, sich also für diese Klasse entscheidet ?

Wahrscheinlichkeit von Fehlklassifikationen

1. Beispiel: Zwei-Klassen-Aufgabe

$$\boldsymbol{P(f|x) := P(c^l|x), l \neq k}$$

2. Beispiel: Mehr-Klassen-Aufgabe

$$\begin{aligned} \boldsymbol{P(f|x)} &:= \sum_{\substack{l=1 \\ l \neq k}}^{K} \boldsymbol{P(c^l|x)} \\ &:= \boldsymbol{1 - P(c^k|x)} \end{aligned}$$

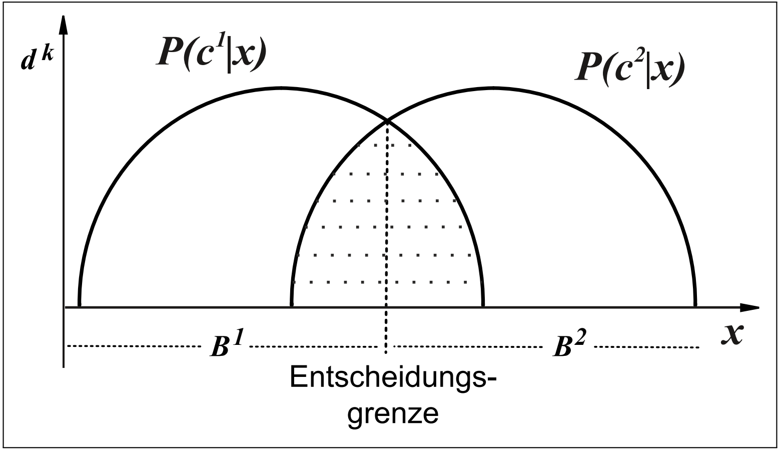
Erwartungswert der Fehlerwahrscheinlichkeit

Erwartungwert der Wahrscheinlichkeit, einen Fehler bei der Klassifizierung zu begehen:

Integration über alle möglichen Messvektoren.

$$\boldsymbol{P(f) := E(P(f|x)) = \int\limits_{-\infty}^{\infty} P(f|x) \cdot P(x) dx}$$

Illustration der Fehlerwahrscheinlichkeit



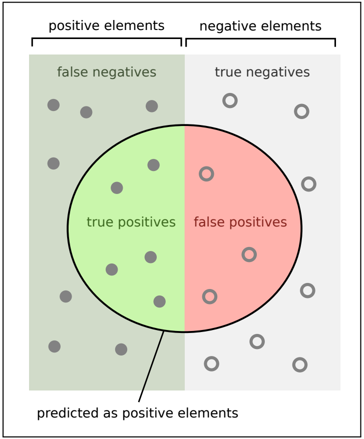
Fehlerwahrscheinlichkeit bei einer Zwei-Klassen-Aufgabe

$$\begin{aligned}
\boldsymbol{P(f)} &:= \boldsymbol{P(x \in B^1, c^2) + P(x \in B^2, c^1)} \\
&:= \int_{\boldsymbol{B^1}} \boldsymbol{P(x|c^2) \cdot P(c^2)} + \\
&\qquad \int_{\boldsymbol{B^2}} \boldsymbol{P(x|c^1) \cdot P(c^1)}
\end{aligned}$$

Praktisch relevante Evaluationskriterien

Beispiel für eine Klassifikation:

$\boldsymbol{TP}$: True Positives
$\boldsymbol{FP}$: False Positives
$\boldsymbol{TN}$: True Negatives
$\boldsymbol{FP}$: False Negatives

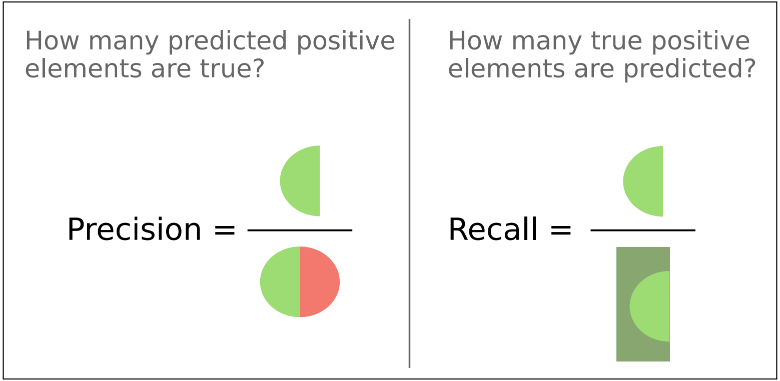


(adaptiert aus Wikipedia, Precision and Recall)

Praktisch relevante Evaluationskriterien

Definition von Precision und Recall:

$$\text{Precision} := \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}\ , \quad \text{Recall} := \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$



(adaptiert aus Wikipedia, Precision and Recall)

2.7 Datenpartitionierung und Kreuzvalidierung

Überblick:

- Lernmenge, Evaluationsmenge, Anwendungsmenge
- Trainingsmenge, Validierungsmenge
- Kreuzvalidierung

Lernmenge, Evaluationsmenge, Anwendungsmenge

Partitionierung des Problembereichs (drei disjunkte Datenmengen):

$$\boldsymbol{\Omega} := \boldsymbol{\Omega}_L \cup \boldsymbol{\Omega}_E \cup \boldsymbol{\Omega}_A$$

$\boldsymbol{\Omega}_L$: Lernmenge zum Lernen eines Klassifikators

$$\boldsymbol{\Omega}_L := \{(\boldsymbol{x}^m, \boldsymbol{c}^{k_m}) | \boldsymbol{m} \in \{1, \ldots, \boldsymbol{M}_L\}\}$$

Klassifizierte Elemente, d.h.
Messvektoren und Klassenzugehörigkeiten.
Grundlage für das Lernen des Klassifikators.

Lernmenge, Evaluationsmenge, Anwendungsmenge

$\boldsymbol{\Omega_E}$: Evaluationsmenge zum Evaluieren eines Klassifikators

$$\boldsymbol{\Omega_E := \{(x^m, c^{k_m}) | m \in \{1, \ldots, M_E\}\}}$$

Klassifizierte Elemente, d.h.

Messvektoren und Klassenzugehörigkeiten.

Evaluation des Klassifikators und Publikation.

$\boldsymbol{\Omega_A}$: Anwendungsmenge bei Anwendung eines Klassifikators

$$\boldsymbol{\Omega_A := \{x^m | m \in \{1, \ldots, M_A\}\}}$$

Unklassifizierte Elemente, d.h.

nur Messvektoren.

Prädiktion der Klassenzugehörigkeiten.

Trainingsmenge, Validierungsmenge

Weitere disjunkte Partitionierung der Lernmenge $\boldsymbol{\Omega_L := \Omega_T \cup \Omega_V}$,
in Trainingsmenge $\boldsymbol{\Omega_T}$ und Validierungsmenge $\boldsymbol{\Omega_V}$.

$\boldsymbol{\Omega_T}$: Trainiere Klassifikator so lange, bis die Entscheidung $\boldsymbol{e : x \rightarrow c^k}$ mit einem akzeptablen Minimum an Fehlentscheidungen erfolgt.

$\boldsymbol{\Omega_V}$: Validiere, ob Fehlerrate des Klassifikators bei $\boldsymbol{\Omega_V}$ höher ist als die Fehlerrate bei $\boldsymbol{\Omega_T}$:

- nein $\rightarrow$ Beendigung des Trainings.
- ja $\rightarrow$ Fortsetzung des Trainings, evtl. Adaption von Netztopologie und/oder Hyperparameter erforderlich.

Kreuzvalidierung

Engl.: Cross Validation

Praktikabler Lernansatz:

- Variables Aufspalten der Lernmenge $\boldsymbol{\Omega_L := \Omega_{T_j} \cup \Omega_{V_j}}$.
- Iteratives Verändern der Zusammensetzung von Trainings- und Validierungsmenge.
- Jeweils trainieren mit aktueller Trainingsmenge und Evaluation bezüglich aktueller Validierungsmenge.
- Zusammenführen aller Ergebnisse.

Kreuzvalidierung

Beispiel: