



Signal Processing and Machine Learning

1. Math

$\pi \approx 3.141\,59$ $e \approx 2.718\,28$ $\sqrt{2} \approx 1.414$ $\sqrt{3} \approx 1.732$

Binome, Trinome
 $(a \pm b)^2 = a^2 \pm 2ab + b^2$ $a^2 - b^2 = (a - b)(a + b)$
 $(a \pm b)^3 = a^3 \pm 3a^2b + 3ab^2 \pm b^3$
 $(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$

Folgen und Reihen
 $\sum_{k=1}^n k = \frac{n(n+1)}{2}$ $\sum_{k=0}^n q^k = \frac{1-q^{n+1}}{1-q}$ $\sum_{n=0}^{\infty} \frac{z^n}{n!} = e^z$
Arithmetrische SummenformelGeometrische SummenformelExponentialreihe

Mittelwerte (\sum von i bis N) (Median: Mitte einer geordneten Liste)
 $\bar{x}_{ar} = \frac{1}{N} \sum x_i \geq \bar{x}_{geo} = \sqrt[N]{\prod x_i} \geq \bar{x}_{hm} = \frac{N}{\sum \frac{1}{x_i}}$
ArithmetischesGeometrisches MittelHarmonisches

Ungleichungen: Bernoulli-Ungleichung: $(1+x)^n \geq 1+nx$
 $||x| - |y|| \leq |x \pm y| \leq |x| + |y|$ $|\underline{x}^T \cdot \underline{y}| \leq ||\underline{x}|| \cdot ||\underline{y}||$
DreiecksungleichungCauchy-Schwarz-Ungleichung

Mengen: De Morgan: $\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\overline{A \cup B} = \overline{A} \cap \overline{B}$

1.1. Exp. und Log. $e^x := \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$ $e \approx 2,71828$
 $a^x = e^{x \ln a}$ $\log_a x = \frac{\ln x}{\ln a}$ $\ln x \leq x - 1$
 $\ln(x^a) = a \ln(x)$ $\ln(\frac{x}{a}) = \ln x - \ln a$ $\log(1) = 0$

1.2. Matrizen $\underline{A} \in \mathbb{K}^{m \times n}$
 $\underline{A} = (a_{ij}) \in \mathbb{K}^{m \times n}$ hat m Zeilen (Index i) und n Spalten (Index j)
 $(\underline{A} + \underline{B})^T = \underline{A}^T + \underline{B}^T$ $(\underline{A} \cdot \underline{B})^T = \underline{B}^T \cdot \underline{A}^T$
 $(\underline{A}^T)^{-1} = (\underline{A}^{-1})^T$ $(\underline{A} \cdot \underline{B})^{-1} = \underline{B}^{-1} \underline{A}^{-1}$
 $\dim \mathbb{K} = n = \text{rang } \underline{A} + \dim \ker \underline{A}$ $\text{rang } \underline{A} = \text{rang } \underline{A}^T$

1.2.1. Quadratische Matrizen $\underline{A} \in \mathbb{K}^{n \times n}$
regulär/invertierbar/nicht-singulär $\Leftrightarrow \det(\underline{A}) \neq 0 \Leftrightarrow \text{rang } \underline{A} = n$
singulär/nicht-invertierbar $\Leftrightarrow \det(\underline{A}) = 0 \Leftrightarrow \text{rang } \underline{A} \neq n$
orthogonal $\Leftrightarrow \underline{A}^T = \underline{A}^{-1} \Rightarrow \det(\underline{A}) = \pm 1$
symmetrisch: $\underline{A} = \underline{A}^T$ schief-symmetrisch: $\underline{A} = -\underline{A}^T$

1.2.2. Determinante von $\underline{A} \in \mathbb{K}^{n \times n}$: $\det(\underline{A}) = |\underline{A}|$
 $\det \begin{bmatrix} \underline{A} & \underline{0} \\ \underline{C} & \underline{D} \end{bmatrix} = \det \begin{bmatrix} \underline{A} & \underline{B} \\ \underline{0} & \underline{D} \end{bmatrix} = \det(\underline{A}) \det(\underline{D})$
 $\det(\underline{A}) = \det(\underline{A}^T)$ $\det(\underline{A}^{-1}) = \det(\underline{A})^{-1}$
 $\det(\underline{A}\underline{B}) = \det(\underline{A}) \det(\underline{B}) = \det(\underline{B}) \det(\underline{A}) = \det(\underline{B}\underline{A})$
Hat \underline{A} 2 linear abhäng. Zeilen/Spalten $\Rightarrow |\underline{A}| = 0$

1.2.3. Eigenwerte (EW) λ und Eigenvektoren (EV) \underline{v}
 $\underline{A}\underline{v} = \lambda \underline{v}$ $\det \underline{A} = \prod \lambda_i$ $\text{Sp } \underline{A} = \sum a_{ii} = \sum \lambda_i$

Eigenwerte: $\det(\underline{A} - \lambda \underline{1}) = 0$ Eigenvektoren: $\ker(\underline{A} - \lambda_i \underline{1}) = \underline{v}_i$
EW von Dreieck/Diagonal Matrizen sind die Elem. der Hauptdiagonale.

1.2.4. Spezialfall 2×2 Matrix \underline{A}
 $\det(\underline{A}) = ad - bc$ $\text{Sp}(\underline{A}) = a + d$ $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \underline{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$\lambda_{1/2} = \frac{\text{Sp } \underline{A}}{2} \pm \sqrt{\left(\frac{\text{Sp } \underline{A}}{2}\right)^2 - \det \underline{A}}$

1.2.5. Differentiation
 $\frac{\partial \underline{x}^T \underline{y}}{\partial \underline{x}} = \frac{\partial \underline{y}^T \underline{x}}{\partial \underline{x}} = \underline{y}$ $\frac{\partial \underline{x}^T \underline{a}}{\partial \underline{x}} = \underline{a}$ $\frac{\partial \underline{x}^T \underline{A} \underline{x}}{\partial \underline{A}} = \underline{x} \underline{y}^T$ $\frac{\partial \det(\underline{B} \underline{A} \underline{C})}{\partial \underline{A}} = \det(\underline{B} \underline{A} \underline{C}) (\underline{A}^{-1})^T$

1.2.6. Ableitungsregeln ($\forall \lambda, \mu \in \mathbb{R}$)
Linearität: $(\lambda f + \mu g)'(x) = \lambda f'(x) + \mu g'(x)$
Produkt: $(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$
Quotient: $\left(\frac{f}{g}\right)'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$ ($\frac{\text{NAZ}-\text{ZAN}}{\text{N}^2}$)
Kettenregel: $(f(g(x)))' = f'(g(x))g'(x)$

1.3. Integrale $\int e^x dx = e^x = (e^x)'$
Partielle Integration: $\int u w' = u w - \int u' w$
Substitution: $\int f(g(x))g'(x) dx = \int f(t) dt$

$F(x) - C$	$f(x)$	$f'(x)$
$\frac{1}{q+1} x^{q+1}$	x^q	$q x^{q-1}$
$\frac{2\sqrt{q x^3}}{3}$	$\sqrt{a x}$	$\frac{a}{2\sqrt{a x}}$
$x \ln(ax) - x$	$\ln(ax)$	$\frac{1}{x}$
$\frac{1}{a^2} e^{ax} (ax - 1)$	$x \cdot e^{ax}$	$e^{ax} (ax + 1)$
$\frac{a^x}{\ln(a)}$	a^x	$a^x \ln(a)$
$-\cos(x)$	$\sin(x)$	$\cos(x)$
$\cosh(x)$	$\sinh(x)$	$\cosh(x)$
$-\ln \cos(x) $	$\tan(x)$	$\frac{1}{\cos^2(x)}$

$\int e^{at} \sin(bt) dt = e^{at} \frac{a \sin(bt) + b \cos(bt)}{a^2 + b^2}$
 $\int \frac{dt}{\sqrt{at+b}} = \frac{2\sqrt{at+b}}{a}$ $\int t^2 e^{at} dt = \frac{(a x - 1)^2 + 1}{a^3} e^{at}$
 $\int t e^{at} dt = \frac{at-1}{a^2} e^{at}$ $\int x e^{ax^2} dx = \frac{1}{2a} e^{ax^2}$

1.3.1. Volumen und Oberfläche von Rotationskörpern um x-Achse
 $V = \pi \int_a^b f(x)^2 dx$ $O = 2\pi \int_a^b f(x) \sqrt{1 + f'(x)^2} dx$

2. Probability Theory Basics

2.1. Kombinatorik
Mögliche Variationen/Kombinationen um k Elemente von maximal n Elementen zu wählen bzw. k Elemente auf n Felder zu verteilen:

	Mit Reihenfolge	Reihenfolge egal
Mit Wiederholung	n^k	$\binom{n+k-1}{k}$
Ohne Wiederholung	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Permutation von n mit jeweils k gleichen Elementen: $\frac{n!}{k_1! \cdot k_2! \cdot \dots}$
Binomialkoeffizient $\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k! \cdot (n-k)!}$
 $\binom{n}{0} = 1$ $\binom{n}{1} = n$ $\binom{n}{2} = 6$ $\binom{n}{5} = 10$ $\binom{n}{6} = 15$

2.2. Der Wahrscheinlichkeitsraum $(\Omega, \mathbb{F}, \mathbb{P})$

Ergebnismenge	$\Omega = \{\omega_1, \omega_2, \dots\}$	Ergebnis $\omega_j \in \Omega$
Ereignisalgebra	$\mathbb{F} = \{A_1, A_2, \dots\}$	Ereignis $A_i \subseteq \Omega$
Wahrscheinlichkeitsmaß	$\mathbb{P} : \mathbb{F} \rightarrow [0, 1]$	$\mathbb{P}(A) = \frac{ A }{ \Omega }$

2.3. Wahrscheinlichkeitsmaß \mathbb{P}
 $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

2.3.1. Axiome von Kolmogorow
Nichtnegativität: $\mathbb{P}(A) \geq 0 \Rightarrow \mathbb{P} : \mathbb{F} \mapsto [0, 1]$
Normiertheit: $\mathbb{P}(\Omega) = 1$
Additivität: $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, wenn $A_i \cap A_j = \emptyset, \forall i \neq j$

2.4. Bedingte Wahrscheinlichkeit
Bedingte Wahrscheinlichkeit für A falls B bereits eingetreten ist:
 $\mathbb{P}_B(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

2.4.1. Totale Wahrscheinlichkeit und Satz von Bayes
Es muss gelten: $\bigcup_{i \in I} B_i = \Omega$ für $B_i \cap B_j = \emptyset, \forall i \neq j$

Totale Wahrscheinlichkeit: $\mathbb{P}(A) = \sum_{i \in I} \mathbb{P}(A|B_i) \mathbb{P}(B_i)$

Satz von Bayes: $\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k) \mathbb{P}(B_k)}{\sum_{i \in I} \mathbb{P}(A|B_i) \mathbb{P}(B_i)}$

Multiplikationssatz: $\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \mathbb{P}(B) = \mathbb{P}(B|A) \mathbb{P}(A)$

2.5. Zufallsvariable
 $X : \Omega \mapsto \Omega'$ ist Zufallsvariable, wenn für jedes Ereignis $A' \in \mathbb{F}'$ im Bildraum ein Ereignis A im Urbildraum \mathbb{F} existiert, sodass $\{\omega \in \Omega | X(\omega) \in A'\} \in \mathbb{F}$

2.6. Distribution

Bezeichnung	Abk.	Zusammenhang
Wahrscheinlichkeitsdichte	pdf	$f_X(x) = \frac{dF_X(x)}{dx}$
Kumulative Verteilungsfkt.	cdf	$F_X(x) = \int_{-\infty}^x f_X(\xi) d\xi$

Joint CDF: $F_{X,Y}(x,y) = \mathbb{P}(\{X \leq x, Y \leq y\})$

2.7. Relations between $f_X(x), f_{X,Y}(x,y), f_{X|Y}(x|y)$

$$f_{X,Y}(x,y) = f_X|Y(x,y) f_Y(y) = f_Y|X(y,x) f_X(x)$$

Joint PDF

$$\underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,\xi) d\xi}_{\text{Marginalization}} = \underbrace{\int_{-\infty}^{\infty} f_{X|Y}(x,\xi) f_Y(\xi) d\xi}_{\text{Total Probability}} = f_X(x)$$

2.8. Bedingte Zufallsvariablen

Ereignis A gegeben: $F_{X|A}(x|A) = \mathbb{P}(\{X \leq x\} | A)$
ZV Y gegeben: $F_{X|Y}(x|y) = \mathbb{P}(\{X \leq x\} | \{Y = y\})$

$p_{X,Y}(x,y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$
 $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{dF_{X|Y}(x|y)}{dx}$

2.9. Unabhängigkeit von Zufallsvariablen
 X_1, \dots, X_n sind stochastisch unabhängig, wenn für jedes $\underline{x} \in \mathbb{R}^n$ gilt:

$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i)$
 $p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i)$
 $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$

3. Gaussian Stuff

3.1. Gaussian Channel
Channel: $Y = h s_i + N$ with $h \sim \mathcal{N}, N \sim \mathcal{N}$
 $L(y_1, \dots, y_N) = \prod_{i=1}^N f_{Y_i}(y_i, h)$
 $f_{Y_i}(y_i, h) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y_i - h s_i)^2\right)$
 $\hat{h}_{ML} = \underset{h}{\text{argmin}} \left\{ \left\| \underline{y} - h \underline{s} \right\|^2 \right\} = \frac{\underline{s}^T \underline{y}}{\underline{s}^T \underline{s}}$

If multidimensional channel: $\underline{y} = \underline{S} \underline{h} + \underline{n}$:
 $L(\underline{y}, \underline{h}) = \frac{1}{\sqrt{\det(2\pi \underline{C})}} \exp\left(-\frac{1}{2}(\underline{y} - \underline{S} \underline{h})^T \underline{C}^{-1}(\underline{y} - \underline{S} \underline{h})\right)$
 $l(\underline{y}, \underline{h}) = \frac{1}{2} \left(\log(\det(2\pi \underline{C})) - (\underline{y} - \underline{S} \underline{h})^T \underline{C}^{-1}(\underline{y} - \underline{S} \underline{h}) \right)$
 $\frac{d}{d\underline{h}} (\underline{y} - \underline{S} \underline{h})^T \underline{C}^{-1}(\underline{y} - \underline{S} \underline{h}) = -2 \underline{S}^T \underline{C}^{-1}(\underline{y} - \underline{S} \underline{h})$

Gaussian Covariance: if $Y \sim \mathcal{N}(0, \sigma^2), N \sim \mathcal{N}(0, \sigma^2)$:
 $\underline{C}_Y = \text{Cov}[Y, Y] = \mathbb{E}[(Y - \mu)(Y - \mu)^T] = \mathbb{E}[Y Y^T]$
For Channel $Y = S h + N$: $\mathbb{E}[Y Y^T] = S \mathbb{E}[h h^T] S^T + \mathbb{E}[N N^T]$

3.2. Multivariate Gaussian Distributions
A vector \underline{x} of n independent Gaussian random variables x_i is jointly Gaussian. If $\underline{x} \sim \mathcal{N}(\underline{\mu}_{\underline{x}}, \underline{C}_{\underline{x}})$:

$$f_{\underline{x}}(\underline{x}) = f_{x_1, \dots, x_n}(x_1, \dots, x_n) = \frac{1}{\sqrt{\det(2\pi \underline{C}_{\underline{x}})}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu}_{\underline{x}})^T \underline{C}_{\underline{x}}^{-1}(\underline{x} - \underline{\mu}_{\underline{x}})\right)$$

Affine transformations $\underline{y} = \underline{A} \underline{x} + \underline{b}$ are jointly Gaussian with $\underline{y} \sim \mathcal{N}(\underline{A} \underline{\mu}_{\underline{x}} + \underline{b}, \underline{A} \underline{C}_{\underline{x}} \underline{A}^T)$
All marginal PDFs are Gaussian as well

Contour Lines
Ellipsoid with central point $\mathbb{E}[\underline{y}]$ and main axis are the eigenvectors of $\underline{C}_{\underline{y}}^{-1}$

3.3. Conditional Gaussian
 $\underline{A} \sim \mathcal{N}(\underline{\mu}_{\underline{A}}, \underline{C}_{\underline{A}}), \underline{B} \sim \mathcal{N}(\underline{\mu}_{\underline{B}}, \underline{C}_{\underline{B}})$
 $\Rightarrow (\underline{A}|\underline{B}=\underline{b}) \sim \mathcal{N}(\underline{\mu}_{\underline{A}|\underline{B}}, \underline{C}_{\underline{A}|\underline{B}})$

Conditional Mean:
 $\mathbb{E}[\underline{A}|\underline{B}=\underline{b}] = \underline{\mu}_{\underline{A}|\underline{B}=\underline{b}} = \underline{\mu}_{\underline{A}} + \underline{C}_{\underline{A}\underline{B}} \underline{C}_{\underline{B}\underline{B}}^{-1}(\underline{b} - \underline{\mu}_{\underline{B}})$

Conditional Variance:
 $\underline{C}_{\underline{A}|\underline{B}} = \underline{C}_{\underline{A}\underline{A}} - \underline{C}_{\underline{A}\underline{B}} \underline{C}_{\underline{B}\underline{B}}^{-1} \underline{C}_{\underline{B}\underline{A}}$

3.4. Misc
If CDF of gaussian distribution given $\Phi(z) \sim \mathcal{N}(0, 1)$ then for $X \sim \mathcal{N}(1, 1)$ the CDF is given as $\Phi(x - \mu_x)$

4. Common Distributions

4.1. Binomialverteilung $\mathcal{B}(n, p)$ mit $p \in [0, 1], n \in \mathbb{N}$
Folge von n Bernoulli-Experimenten
 p : Wahrscheinlichkeit für Erfolg k : Anzahl der Erfolge

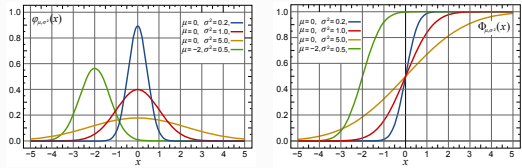
$$p_X(k) = B_{n,p}(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0, \dots, n\} \\ 0 & \text{sonst} \end{cases}$$

$\mathbb{E}[X] = np$	$\text{Var}[X] = np(1-p)$	$G_X(z) = (pz + 1 - p)^n$
Erwartungswert	Varianz	Wahrscheinlichkeitserz. Funktion

4.2. Normalverteilung

WDF/PDF:

KVF/CDF:



$$\text{WDF: } f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R} \quad \mu \in \mathbb{R} \quad \sigma > 0$$

$$\begin{array}{lll} E(X) = \mu & \text{Var}(X) = \sigma^2 & \varphi_X(\omega) = e^{j\omega\mu - \frac{\omega^2\sigma^2}{2}} \\ \text{Erwartungswert} & \text{Varianz} & \text{Charakt. Funktion} \end{array}$$

4.3. Sonstiges

Gammadistribution $\Gamma(\alpha, \beta)$: $E[X] = \frac{\alpha}{\beta}$

Exponential: $f(x, \lambda) = \lambda e^{-\lambda x}$ $E[X] = \lambda^{-1}$ $\text{Var}[X] = \lambda^{-2}$

5. Wichtige Parameter

5.1. Erwartungswert (1. zentrales Moment)

gibt den mittleren Wert einer Zufallsvariablen an

$$\mu_X = E[X] = \sum_{x \in \Omega'} x \cdot P_X(x) \triangleq \int_{\mathbb{R}} x \cdot f_X(x) dx$$

diskrete $X: \Omega \rightarrow \Omega'$ stetige $X: \Omega \rightarrow \mathbb{R}$

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y] \quad X \leq Y \Rightarrow E[X] \leq E[Y]$$

$$E[X^2] = \text{Var}[X] + E[X]^2$$

$E[X Y] = E[X] E[Y]$, falls X und Y stochastisch unabhängig
Umkehrung nicht möglich: Unkorreliertheit \nRightarrow Stoch. Unabhängig!

5.1.1. Für Funktionen von Zufallsvariablen $g(x)$

$$E[g(X)] = \sum_{x \in \Omega'} g(x) P_X(x) \triangleq \int_{\mathbb{R}} g(x) f_X(x) dx$$

5.2. Varianz (2. zentrales Moment)

ist ein Maß für die Stärke der Abweichung vom Erwartungswert

$$\sigma_X^2 = \text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

$$\text{Var}[\alpha X + \beta] = \alpha^2 \text{Var}[X] \quad \text{Var}[X] = \text{Cov}[X, X]$$

$$\text{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \text{Var}[X_i] + \sum_{i \neq j} \text{Cov}[X_i, X_j]$$

Standard Abweichung: $\sigma = \sqrt{\text{Var}[X]}$

5.3. Kovarianz

Maß für den linearen Zusammenhang zweier Variablen

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])^T] = \\ &= E[X Y^T] - E[X] E[Y]^T = \text{Cov}[Y, X] \end{aligned}$$

$$\text{Cov}[\alpha X + \beta, \gamma Y + \delta] = \alpha \gamma \text{Cov}[X, Y]$$

$$\text{Cov}[X + U, Y + V] = \text{Cov}[X, Y] + \text{Cov}[X, V] + \text{Cov}[U, Y] + \text{Cov}[U, V]$$

5.3.1. Korrelation = standardisierte Kovarianz

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \cdot \text{Var}[Y]}} = \frac{C_{xy}}{\sigma_x \sigma_y} \quad \rho(X, Y) \in [-1; 1]$$

5.3.2. Kovarianzmatrix für $\underline{z} = (\underline{x}, \underline{y})^T$

$$\text{Cov}[\underline{z}] = \underline{C}_{\underline{z}} = \begin{bmatrix} C_X & C_{XY} \\ C_{XY}^T & C_Y \end{bmatrix} = \begin{bmatrix} \text{Cov}[X, X] & \text{Cov}[X, Y] \\ \text{Cov}[Y, X] & \text{Cov}[Y, Y] \end{bmatrix}$$

Immer symmetrisch: $C_{xy} = C_{yx}!$ Für Matrizen: $\underline{C}_{\underline{x}\underline{y}} = \underline{C}_{\underline{y}\underline{x}}^T$

6. Statistical Learning

6.1. Definition

Statistical Model

Statistical Model: $\{\mathbb{X}, \mathbb{F}, P_\theta; \theta \in \Theta\}$
Sample Space: Ω
Observation Space: \mathbb{X}
Sigma Algebra: \mathbb{F}
Probability: P_θ
Test (decision rule): $T: \mathbb{X} \mapsto \{\theta_0, \theta_1\}, x \mapsto T(x)$
Null Hypothesis: $H_0: \theta \in \Theta_0$
Alternative Hypothesis: $H_1: \theta \in \Theta_1$

Cost Criterion G_T :

$$G_T: \{\theta_0, \theta_1\} \mapsto [0, 1], \theta \mapsto P(\{T(X) = 1\}; \theta)$$

$$= E[T(X); \theta] = \int_{\mathbb{X}} T(x) f_X(x; \theta) dx$$

Error Level α : $G_T(\theta_0) \leq \alpha$

Two Error Types:

False Alarm: $\theta = \theta_0, T(x) = 1$

$G_T(\theta_0) = P(\{T(X) = 1\}; \theta_0)$

Detection Error: $\theta = \theta_1, T(x) = 0$

$1 - G_T(\theta_1) = P(\{T(X) = 0\}; \theta_1)$

6.2. Maximum Likelihood Test

ML Ratio Test Statistic (Likelihood Ratio):

$$R(x) = \begin{cases} \frac{f_X(x; \theta_1)}{f_X(x; \theta_0)} & ; f_X(x; \theta_0) > 0 \\ \infty & ; f_X(x; \theta_0) = 0 \text{ and } f_X(x; \theta_1) > 0 \end{cases}$$

ML Test:

$$T_{ML}: \mathbb{X} \mapsto \{0, 1\}, x \mapsto \begin{cases} 1 & ; R(x) > c = 1 \\ 0 & ; \text{otherwise} \end{cases}$$

if $R(x)$ is monotonous then it is possible to make a decision by directly comparing x to a threshold x_α and every $R(x) \geq c(\alpha)$ will lead to a unique threshold for $x_\alpha < x$

if $c \neq 1$ False Alarm Error Probability can be adjusted \rightarrow Neyman Pearson Test

6.3. Neyman-Pearson-Test

minimizes the detection error, while fulfilling a predefined error level α

$$\arg\max_{d_{NP}} E[d_{NP}(x)|\theta = \theta_1] \quad \text{s.t.} \quad E[d_{NP}(x)|\theta = \theta_0] \leq \alpha$$

NP-Test to the error level α :

x_α is chosen as: $x_\alpha = (1 - \alpha)$ -quantile of $f_X(x; \theta_0)$

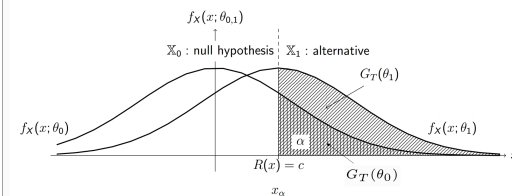
If $P(\{R(x) = c; \theta_0\}) = 0 \Leftrightarrow$ (if x is continuous):

$$T_{NP}(x) = \begin{cases} 1 & R(x) > c \\ 0 & R(x) < c \end{cases} \quad R(x) = \frac{f_X(x; \theta_1)}{f_X(x; \theta_0)}$$

If $P(\{R(x) = c; \theta_0\}) > 0$:

$$T_{NP}(x) = \begin{cases} 1 & R(x) > c \\ \gamma & R(x) = c, \quad (\text{randomized decision}) \\ 0 & R(x) < c \end{cases}$$

with $\gamma = \frac{\alpha - P(\{R(x) > c; \theta_0\})}{P(\{R(x) = c; \theta_0\})}$ error level α



Maximum Likelihood Detector: $T_{ML}(x) = \begin{cases} 1 & R(x) > 1 \\ 0 & \text{otherwise} \end{cases}$

ROC Graphs: plot $G_T(\theta_1)$ as a function of $G_T(\theta_0)$

6.4. Bayes Test (MAP Test)

Prior knowledge about possible hypotheses:

$$P(\{\theta \in \Theta_0\}) + P(\{\theta \in \Theta_1\}) = 1$$

$$T_{\text{Bayes}} = \arg\min_T \{P_\epsilon\} = \begin{cases} 1 & ; \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > c = \frac{P(\theta_0)}{P(\theta_1)} \\ 0 & ; \text{otherwise} \end{cases}$$

$$= \begin{cases} 1 & ; P(\theta_1|x) > P(\theta_0|x) \\ 0 & ; \text{otherwise} \end{cases}$$

with:

$$P_\epsilon = P(\theta_0)G_T(\theta_0) + P(\theta_1)(1 - G_T(\theta_1))$$

if $P(\theta_0) = P(\theta_1) \rightarrow T_{\text{Bayes}} = T_{ML}$

Multiple Hypothesis $\{\theta_0, \dots, \theta_k\}; \mathbb{X}_0, \dots, \mathbb{X}_k \in \mathbb{X}$:

$$T_{\text{Bayes}} = \arg\min_{k \in 1, \dots, K} \{P(\theta_k|x)\}$$

Loss Function:

$$L(T(x), \theta) = \begin{cases} L_0 & ; T(x) = 1, \text{ but } \theta = \theta_0 \quad (\text{FALSE ALARM}) \\ L_1 & ; T(x) = 0, \text{ but } \theta = \theta_1 \quad (\text{DETEC. ERROR}) \\ 0 & ; \text{otherwise} \end{cases}$$

L_i denotes the Loss Value in cases where the correct decision parameter θ_i is missed.

$$\text{Risk}(T) = E[L(T(X), \theta)] = E[E[L(T(x), \theta)|x = X]]$$

6.5. Linear Alternative Tests

Estimate normal vector \underline{w}^T and w_0 , which separate \mathbb{X} into \mathbb{X}_0 and \mathbb{X}_1

$$\log R(\underline{x}) = -\frac{1}{2} \ln \left(\frac{\det(\underline{C}_1)}{\det(\underline{C}_0)} \right) - \frac{1}{2} (\underline{x} - \underline{\mu}_1)^T \underline{C}_1^{-1} (\underline{x} - \underline{\mu}_1) + \frac{1}{2} (\underline{x} - \underline{\mu}_0)^T \underline{C}_0^{-1} (\underline{x} - \underline{\mu}_0) = \ln \left(\frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)} \right) \quad (\text{separating surface})$$

For Gaussian $f_X(x; \mu_k, C_k)$ with θ_0 and θ_1 corresponding to $\{\mu_0, C_0\}$ and $\{\mu_1, C_1\}$, it follows that

• if $C_0 \neq C_1$, $\log R(x) = 0$ is non-linear and the separating surfaces are surfaces of second order: parabolic, hyperbolic, or elliptic surfaces.

• if $C_0 = C_1$, $\log R(x) = 0$ is affine and thus defines a hyperplane in \mathbb{X} which decomposes \mathbb{X} into \mathbb{X}_0 and \mathbb{X}_1 , i.e.,

$$T: \mathbb{X} \mapsto \mathbb{R}, \underline{x} \mapsto \begin{cases} 1 & \underline{w}^T \underline{x} > w_0 \\ 0 & \text{otherwise} \end{cases}$$

– case 1: $\underline{C}_0 = \underline{C}_1 = \sigma^2 \underline{I}_N$

$$\underline{w}^T = (\underline{\mu}_1 - \underline{\mu}_0)^T,$$

$$w_0 = \frac{1}{2} (\underline{\mu}_1^T \underline{\mu}_1 - \underline{\mu}_0^T \underline{\mu}_0) - \sigma^2 \ln \left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)} \right)$$

\underline{w} colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$
 \rightarrow hyperplane orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$

– case 2: $\underline{C}_0 = \underline{C}_1 = \underline{C}$

$$\underline{w}^T = (\underline{\mu}_1 - \underline{\mu}_0)^T \underline{C}^{-1},$$

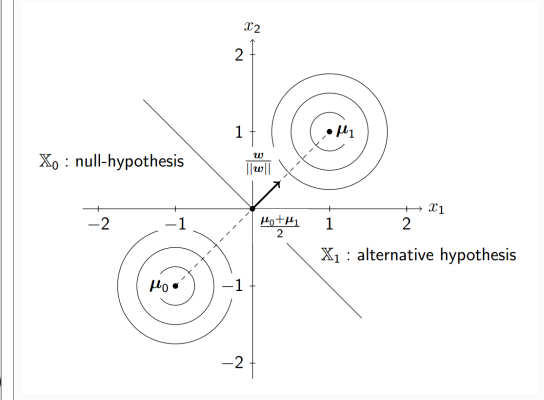
$$w_0 = \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_0)^T \underline{C}^{-1} (\underline{\mu}_1 + \underline{\mu}_0) - \ln \left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)} \right)$$

in general \underline{w} not colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$

\rightarrow hyperplane **not** orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$

• if $C_0 = C_1$ and $\mu_0 = -\mu_1$, $\log R(x) = 0$ is linear and defines a separating hyperplane in \mathbb{X} which contains the origin, i.e.,

$$T: \mathbb{X} \mapsto \mathbb{R}, \underline{x} \mapsto \begin{cases} 1 & \underline{w}^T \underline{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$



7. Hypothesis Testing

making a decision based on the observations

7.1. Definition

Null hypothesis $H_0: \theta \in \Theta_0$ (Assumed first to be true)

Alternate hypothesis $H_1: \theta \in \Theta_1$ (The one to proof)

Decision rule $\varphi: \mathbb{X} \rightarrow [0, 1]$ with $\varphi(x) = 1$: decide for H_1 , $\varphi(x) = 0$: decide for H_0 Error level α with $E[d(X)|\theta] \leq \alpha, \forall \theta \in \Theta_0$

Error Type	Decision \ Reality	H_1 false (H_0 true)	H_1 true (H_0 false)
1 (FA) False Alarm	H_1 rejected	True Negative	False Positive (Type 2)
2 (DE) Detection Error	H_1 accepted	False Positive (Type 1)	True Positive
		$P = 1 - \alpha$	$P = \beta$
		$P = \alpha$	$P = 1 - \beta$

$$\text{Power: Sensitivity/Recall/Hit Rate: } \frac{TP}{TP + FN} = 1 - \beta$$

$$\text{Specificity/True negative rate: } \frac{TN}{FP + TN} = 1 - \alpha$$

$$\text{Precision/Positive Prediction rate: } \frac{TP}{TP + FP}$$

$$\text{Accuracy: } \frac{TP + TN}{P + N} = \frac{2 - \alpha - \beta}{2}$$

7.1.1. Design of a test

Cost criterion $G_\varphi: \Theta \rightarrow [0, 1], \theta \mapsto E[d(X)|\theta]$

False Positive lower than α : $G_d(\theta)|_{\theta \in \Theta_0} \leq \alpha, \forall \theta \in \Theta_0$

False Negative small as possible: $\max\{G_d(\theta)|_{\theta \in \Theta_1}\}, \forall \theta \in \Theta_1$

7.2. Sufficient Statistics

Sufficiency for a test $T(X)$ means that no other test statistic, i.e., function of the observations \underline{x} , contains additional information about the parameter θ to be estimated:

$$f_{X|T}(x|T(x) = t, \theta) = f_{X|T}(x|T(x) = t)$$

8. Support Vector Machines

Motivation and Background

8.1. Kernel Methods

Kernel Methods is non-parametric estimation, these make no assumption on statistical model \rightarrow purely Data-Based.

$$\text{Test Statistic } \mathbb{X} \rightarrow \mathbb{R}, \mathbf{x} \mapsto S(\mathbf{x}) = \sum_{k=1}^M \lambda_k g(\mathbf{x}, \mu_k)$$

linear combination of Kernel Function $g(\cdot, \mu_k)$. $g(\cdot)$ generally non-linear pos. definite

μ_k : representative for Sample Set $\mathbb{S} = \{x_1, \dots, x_M\}$
 λ_k : weight coefficient determined by learning
 Sample Set \mathbb{S} is Empirical Characterization of Unknown Statistical Model
 Inference of λ_k based on Sample Set or Training Set is called **Learning**

8.2. Kernel Tests

Statistical Hypothesis Test, where a Sufficient Test Statistic is compared to threshold (i.e. $R(x) \geq c$) decomposes sample space \mathbb{X} into two disjoint subsets ($\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1$)

Separating surface between \mathbb{X}_0 and \mathbb{X}_1 given by:
 $\{x | R(x) = c\}$ The relative position of a sample x_j to the separating surface determines choice of hypothesis

$$\mathbb{S} = \{(x_1, y_1), \dots, (x_M, y_M)\}$$

$x_i \in \mathbb{R}^N$, $y_i \in \{\Theta_0, \Theta_1\}$
 Inference of Hypothesis Test based on a Sample Set that includes Labeling y_i of the elements x_i is called **Supervised Learning**

Size M of samples has to satisfy: $M \geq \dim(\mathbb{X})$

Because underlying statistical model is unknown, true θ_0 and θ_1 irrelevant \rightarrow replace them by e.g. -1,+1 for decision between hypotheses

8.3. Linear Kernels

Test Statistic for linear test

$$S(x) = \sum_{i=1}^M \lambda_i \mathbf{x}_i^T \mathbf{x} + w_0 = \mathbf{w}^T \mathbf{x} + w_0 \quad \mathbf{w} = \sum_{i=1}^M \lambda_i \mathbf{x}_i$$

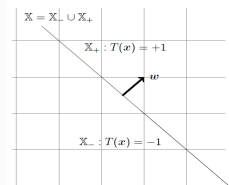
Hyperplane defined by \mathbf{w} (normal vector or weight vector) and w_0 approximates separating surface between \mathbb{X}_- and \mathbb{X}_+

\rightarrow Decision rule $T(x)$:

$$T(\mathbf{x}) = \text{sign}(S(\mathbf{x})) = \begin{cases} +1 & ; \quad \mathbf{w}^T \mathbf{x} + w_0 \geq 0 \\ -1 & ; \quad \text{otherwise} \end{cases}$$

Linear Kernel Test in sample space \mathbb{X} :

(Orientation of \mathbf{w} chosen such that \mathbf{w} points into direction of Θ_1 ("+" hypothesis))



To determine \mathbf{w} and w_0 formulate problem as constrained optimization problem with the constraints:

$\forall k \in \{1, \dots, M\} : T(\mathbf{x}_k) = y_k$

$$\Rightarrow \text{Support Vector Methods: } y_k (\mathbf{w}^T \mathbf{x}_k + w_0) \geq \epsilon, \forall k$$

Robust solution: maximize margin ϵ for constant norm of \mathbf{w}

Application

8.4. Support Vector Methods

only feasible for normalized weight vectors

$$\max_{w_0} \epsilon \text{ s.t. } y_k \frac{\mathbf{w}^T}{\|\mathbf{w}\|_2} \mathbf{x}_k \geq \epsilon, \forall k, w_0 = 0$$

$$\Leftrightarrow \min_{w_0} \frac{1}{2} \|\mathbf{w}\|_2^2 \text{ s.t. } y_k \mathbf{w}^T \mathbf{x}_k \geq 1, \forall k$$

Optimization Problem convex \rightarrow **Langragian Method**

$$\text{Dual Problem: } \max_{\mathbf{u}} \min_{\mathbf{w}} \Phi(\mathbf{w}, \mathbf{u}) \text{ s.t. } \mathbf{u} \geq 0$$

Langragian Multiplier: $u_k \geq 0$

$$\text{Langragian Fct: } \Phi(\mathbf{w}, \mathbf{u}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{k=1}^M u_k (1 - y_k \mathbf{w}^T \mathbf{x}_k)$$

$$\frac{\partial \Phi(\mathbf{w}, \mathbf{u})}{\partial \mathbf{w}} \Big|_{\mathbf{w}=\mathbf{w}(\mathbf{u})} = 0 \Leftrightarrow \mathbf{w}(\mathbf{u}) = \sum_{k=1}^M \underbrace{u_k y_k}_{\lambda_k} \mathbf{x}_k$$

Evaluate dual function:

$$\begin{aligned} \Phi(\mathbf{w}(\mathbf{u}), \mathbf{u}) &= \Phi\left(\sum_{k=1}^M u_k y_k \mathbf{x}_k, u_1, \dots, u_M\right) \\ &= -\frac{1}{2} \sum_{k=1}^M \sum_{l=1}^M u_k u_l y_k y_l \mathbf{x}_k^T \mathbf{x}_l + \sum_{k=1}^M u_k \\ &= -\frac{1}{2} \mathbf{u}^T \mathbf{Y} \mathbf{X} \mathbf{X}^T \mathbf{Y} \mathbf{u} + \mathbf{1}^T \mathbf{u} \end{aligned}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_M^T \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_M \end{bmatrix}, \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Alternativ to approach above:

Iterative Solution:

Choose one element \mathbf{x}_k out of sample set $\mathbb{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ and randomly set:

$$u_k \leftarrow u_k + \max\{y_k \frac{\partial \Phi(\mathbf{u})}{\partial u_k}, -u_k\}, \forall k$$

Necessary and sufficient condition for existence of solution given by:

$$\mathbf{1} \in \text{conce}[\mathbf{Y} \mathbf{X} \mathbf{X}^T \mathbf{Y}]$$

8.5. Suport Vectors

$$\text{Dual OP: } \max_{\mathbf{u}} \sum_{k=1}^M u_k \left(-\frac{1}{2} \sum_{l=1}^M u_l y_l y_k \mathbf{x}_k^T \mathbf{x}_l + u_k\right) \text{ s.t. } u_k \geq 0$$

Optimal Dual Variables u_1^*, \dots, u_M^* either active $u_k > 0$ or inactive $u_k = 0$

Elements of \mathbb{S} with active dual variables = **Support Vectors**

$$\mathbb{S}_{SV} = \{\mathbf{x}_k \in \mathbb{S} | u_k^* > 0\}$$

Elements with inactive dual variables dont contribute to Kernel Test

Optimal Weight Vektor $\mathbf{w}^* = \mathbf{w}(\mathbf{u}^*)$ of Kernel Test constructed by

$$\text{Support Vectors only: } \mathbf{w}^* = \sum_{\mathbf{x}_k \in \mathbb{S}_{SV}} u_k^* y_k \mathbf{x}_k$$

Number of Support Vectors approx. size of $\dim[\mathbb{X}] \rightarrow$ selection of Support Vectors reduces computational complexity of Kernel Test

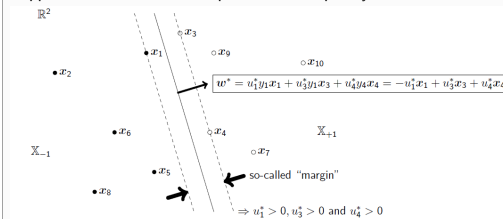


Fig. 2.2: The elements $x_k \in \mathbb{S}$ with ACTIVE DUAL VARIABLES $u_k^* > 0$ are called SUPPORT VECTORS.

Discussion

- Exists only if \mathbb{S} **Linearly Separable**
- $w_0 \neq 0$ no (straightforward) iterative solution available
- if **Linearly Inseparable** method generalized by slack variables for controlled violation of constraints

\rightarrow instead of $\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w}$ s.t. $y_k \mathbf{w}^T \mathbf{x}_k \geq 1$ we get
 $\min_{\mathbf{w}, \epsilon} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \rho \sum_{k=1}^M \epsilon_k$ s.t. $y_k \mathbf{w}^T \mathbf{x}_k \geq 1 - \epsilon_k, \forall k, \epsilon_k, \rho \geq 0$

8.6. Kernel Trick

Linear Hypothesis Test often not sufficient \rightarrow **Kernel Trick**: Generalize linear methods to non-linear approximation of separating surfaces ($\{x | \log R(x) = c\}$)

Basic Idea: Transfer problem statement into higher-dimensional space (without introducing additional degrees of freedom) by **Feature Map** $\varphi : \mathbb{S} \rightarrow \mathbb{S}_\varphi$

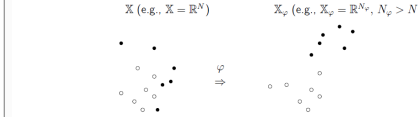


Fig. 2.3: Transfer the problem statement into a higher-dimensional (inner product) space without introducing additional degrees of freedom by means of a so-called FEATURE MAP $\varphi : \mathbb{S} \rightarrow \mathbb{S}_\varphi$.

Construction of Linear Test in \mathbb{R}^3 corresponds to Non-Linear Test in \mathbb{R}^2

$$T : \mathbb{R}^3 \rightarrow \{-1, +1\}, \varphi(\mathbf{x}) \mapsto \begin{cases} +1; & \mathbf{w}_\varphi^T \varphi(\mathbf{x}) \geq 0 \\ -1; & \text{otherwise} \end{cases}$$

Linear kernel in \mathbb{X}_φ represents nonlinear kernel in $\mathbb{X} \rightarrow$ choose Kernel Funktion $g(\dots)$ directly instead of finding appropriate transformation φ

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle =: g(\mathbf{x}, \mathbf{y})$$

In Optimization Problem and resulting Dual Function and Variables

$$\text{replace } \mathbf{x} \text{ by } \varphi(\mathbf{x}_k) \rightarrow \text{Dual OP: } \max_{\mathbf{u} \geq 0} \{-\mathbf{u}^T \mathbf{Y} \mathbf{G} \mathbf{Y} \mathbf{u} + \mathbf{1}^T \mathbf{u}\}$$

$$\text{Kernel Matrix } \mathbf{G} = \begin{bmatrix} g(\mathbf{x}_1, \mathbf{x}_2) & \dots & g(\mathbf{x}_1, \mathbf{x}_M) \\ \vdots & & \vdots \\ g(\mathbf{x}_M, \mathbf{x}_1) & \dots & g(\mathbf{x}_M, \mathbf{x}_M) \end{bmatrix} \in \mathbb{R}^{M \times M}$$

After applying **Kernel Trick**: OP and Nonlinear Test T only based on Kernel Function g , transformation φ becomes obsolete

$$\text{Hypothesis Test (nonlinear): } T : \mathbf{x} \mapsto \text{sign}\left(\sum_{k=1}^M u_k^* y_k g(\mathbf{x}_k, \mathbf{x})\right)$$

Possible Kernels for Kernel Trick

Linear Kernel: $g_{lin}(\mathbf{x}, \mathbf{x}_k) = \mathbf{x}_k^T \mathbf{x}$

Polynomial Kernel: $g_{poly}(\mathbf{x}, \mathbf{x}_k) = (\mathbf{x}_k^T \mathbf{x} + 1)^d$

Sigmoid Kernel: $g_{sigm}(\mathbf{x}, \mathbf{x}_k) = \tanh(\beta(\mathbf{x}_k^T \mathbf{x}) + w_0)$

Radial Kernel: $g_{rbf}(\mathbf{x}, \mathbf{x}_k) = \exp(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}_k\|_2^2)$

Support Vector Machine Representation.

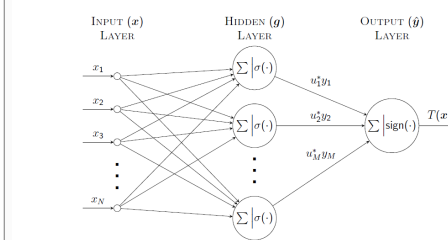


Fig. 2.4: The interpretation of a SUPPORT VECTOR MACHINE as a NEURAL NETWORK with three layers and a non-linear function σ . For POLYNOMIAL KERNELS each SINGLE HIDDEN LAYER UNIT is described by $g_{pol}(\mathbf{x}, \mathbf{x}_k) = \sigma(z_k)$, with $\sigma(z_k) = z_k^d$ and $z_k = \mathbf{x}_k^T \mathbf{x} + 1$.

9. Learning and Generalization

9.1. Empirical Risk Function and Generalization Error

ML scenarios (unknown Stochastical Model) base learning on: $Risk_{emp}(T; \mathbb{S}) = \frac{1}{M} \sum_{i=1}^M L(T(\mathbf{x}_i), y_i), (\mathbf{x}_i, y_i) \in \mathbb{S}$

$$\mathbf{x} \mapsto T(\mathbf{x}; \mathbb{S}) \quad T = \text{argmin}_{T' \in \mathbb{T}} \{Risk_{emp}(T'; \mathbb{S})\}$$

good Generalization: $Risk_{emp}(T; \mathbb{S}_{test})$ similar to $Risk_{emp}(T; \mathbb{S})$
bad Generalization:

- small \mathbb{T} that does not cover $T_{opt} \rightarrow$ cannot be selected by ML \Rightarrow strong mismatch between the desired and derived **Test** and refers to a sort of **Bias Error Term**
- too rich $\mathbb{T} \rightarrow$ fluctuating of the available data (measurement noise) is interpreted as meaningful information \Rightarrow **Overfitting**; leads to an increased **Variance Error Term**

9.2. Bias-Variance Decomposition

$$Risk = E_{S, X, Y} [L(T(X; S), Y)] = E_X [1 - P_{Y|X}(Y = T_B(X)) + (1 - P_{S|X}(T(X; S) = T_B(X))) (2P_{Y|X}(Y = T_B(X)) - 1)]$$

If the potential set \mathbb{S} would be selected from a distribution such that the derived Test $T(\mathbf{x}; \mathbb{S})$ and the corresponding Bayes Test $T_B(\mathbf{x})$ are identical almost surely, then the Risk Function achieves its minimum value which is equal to the **Irreducible Error** $E_X [1 - P_{Y|X}(Y = T_B(X))]$ (denotes the probability that for a given input \mathbf{x} the Bayes Test $T_B(X)$ decides for the false label y).

10. Classification Trees and Random Forests

10.1. CART Algorithms

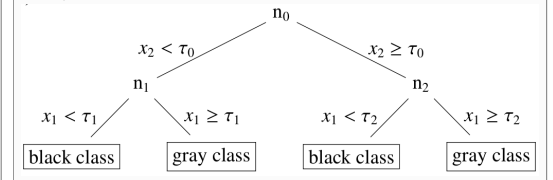
Generate Binary Trees by splitting \mathbb{X} at each (internal/root) node: $\mathbb{X}_{i, left} = \{\mathbf{x} \in \mathbb{X}_i | x_{j_i} < \tau_i\}$, $\mathbb{X}_{i, right} = \mathbb{X}_i \setminus \mathbb{X}_{i, left}$

Root/Internal node: Binary decision based on chosen threshold $\tau_i \in \mathbb{R}$, feature $x_{j_i} = [x]_{j_i}$ with $j_i \in \mathbb{J} = \{1, \dots, \dim[\mathbb{X}]\}$ aims at minimizing $Risk_{emp}(T_{CART})$

Terminal node: n_i corresponds to subset $\mathbb{X}_i \in \mathbb{X} \rightarrow$ has no more children; outputs a decision

$\Rightarrow \mathbf{x} \mapsto n_i(\mathbf{x})$

Example:



Empirical Impurity Measure: choose j_i and τ_i at n_i by:

$$I_{CART}(\mathbb{S}_i) = \sum_{k=1}^K (1 - \hat{P}_{Y|X}(Y = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}_i\}; \mathbb{S}_i)) \hat{P}_{Y|X}(Y = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}_i\}; \mathbb{S}_i)$$
with

$$\hat{P}_{Y|X}(Y = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}_i\}; \mathbb{S}_i) = \frac{M_k(\mathbb{S}_i)}{M(\mathbb{S}_i)} = \frac{|\{(\underline{\mathbf{x}}, y) \in \mathbb{S}_i | y = \theta_k\}|}{|\mathbb{S}_i|}$$

$$\Rightarrow \{j_i, \tau_i\} = \underset{j \in \mathbb{J}, \tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \left(1 - \frac{M_k(\mathbb{S}_i, left)}{M(\mathbb{S}_i, left)} \right) \frac{M_k(\mathbb{S}_i, left)}{M(\mathbb{S}_i)} + \left(1 - \frac{M_k(\mathbb{S}_i, right)}{M(\mathbb{S}_i, right)} \right) \frac{M_k(\mathbb{S}_i, right)}{M(\mathbb{S}_i)} \right\}$$
Overfitting(comes with high purity) can be controlled by a *Test Set* \mathbf{S}^{Test} .
Decision Rule: At terminal node n_i , input $\underline{\mathbf{x}}$ is assigned to

$$T_{CART}(\underline{\mathbf{x}}; \mathbb{S}) : \mathbb{X} \mapsto \{1, \dots, K\}, \underline{\mathbf{x}} \mapsto \underset{k}{\operatorname{argmax}} \{M_k(\mathbb{S}_i)\}$$

Gini Impurity Index: $\boxed{I_{CART}} =$

$$\boxed{\sum_{k=1}^K (1 - P_{Y|X}(\underline{\mathbf{y}} = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}\})) P_{Y|X}(\underline{\mathbf{y}} = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}\})} =$$

$$\sum_{k=1}^K \sum_{j=1, j \neq k}^K P_{Y|X}(\underline{\mathbf{y}} = \theta_j | \{\underline{\mathbf{x}} \in \mathbb{X}\}) P_{Y|X}(\underline{\mathbf{y}} = \theta_k | \{\underline{\mathbf{x}} \in \mathbb{X}\})$$

10.2. Random Forests

Avoid *Overfitting* (here: CART) \Rightarrow combine independent *Hypothesis Tests*: e.g. by *Majority Vote*

$$T_{maj}(\underline{\mathbf{x}}) = majority\{T_{CART}(\underline{\mathbf{x}}; \mathbb{S}^{(t)}, \nu^{(t)})\}_{t=1}^{t_{max}}$$
Randomization Parameter ν_t controls an additionally introduced Randomness between the individual Tests.
 \Rightarrow *Variance* of $T_{avg}(\underline{\mathbf{x}})$ is reduced by $1/t_{max}$ with respect to the *Variance* of the individual test.

Random Forest Method:

- $T_{RF}(\underline{\mathbf{x}}) = majority\{T_{CART}(\underline{\mathbf{x}}; \mathbb{S}^{(t)}, \mathbb{J}^{(t)})\}_{t=1}^{t_{max}}$
- Stochastic Independence by Bootstrapping of training samples (random sampling from \mathbb{S} with replacement) \Rightarrow large t_{max} guarantees excellent performance (yet Tests are still correlated)
- Overfitting not considered (maximum purity) \Rightarrow small bias of RF Method