Signal Processing and Machine Learning

1. Math

1.1. Exp. und Log. $e^x := \lim\limits_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ $\quad e \approx 2,71828$

$a^x = e^{x\ln a}$ $\quad \log_a x = \frac{\ln x}{\ln a}$ $\quad \ln x \leq x-1$

$\ln(x^a) = a\ln(x)$ $\quad \ln(\frac{x}{a}) = \ln x - \ln a$ $\quad \log(1) = 0$

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)

$(\underline{A}+\underline{B})^\top = \underline{A}^\top + \underline{B}^\top$ $\quad (\underline{A}\cdot\underline{B})^\top = \underline{B}^\top\cdot\underline{A}^\top$

$(\underline{A}^\top)^{-1} = (\underline{A}^{-1})^\top$ $\quad (\underline{A}\cdot\underline{B})^{-1} = \underline{B}^{-1}\underline{A}^{-1}$

$\dim\mathbb{K} = n = \operatorname{rang}\underline{A} + \dim\ker\underline{A}$ $\quad \operatorname{rang}\underline{A} = \operatorname{rang}\underline{A}^\top$

1.2.1. Quadratische Matrizen $A \in \mathbb{K}^{n\times n}$

regulär/invertierbar/nicht-singulär $\Leftrightarrow \det(\underline{A}) \neq 0 \Leftrightarrow \operatorname{rang}\underline{A} = n$

singulär/nicht-invertierbar $\Leftrightarrow \det(\underline{A}) = 0 \Leftrightarrow \operatorname{rang}\underline{A} \neq n$

orthogonal $\Leftrightarrow \underline{A}^\top = \underline{A}^{-1} \Rightarrow \det(\underline{A}) = \pm 1$

symmetrisch: $\underline{A} = \underline{A}^\top$ $\quad$ schiefsymmetrisch: $\underline{A} = -\underline{A}^\top$

1.2.2. Determinante von $\underline{A} \in \mathbb{K}^{n\times n}$: $\det(\underline{A}) = |\underline{A}|$

$$\det\begin{bmatrix}\underline{A} & \underline{0}\\ \underline{C} & \underline{D}\end{bmatrix} = \det\begin{bmatrix}\underline{A} & \underline{B}\\ \underline{0} & \underline{D}\end{bmatrix} = \det(\underline{A})\det(\underline{D})$$

$\det(\underline{A}) = \det(\underline{A}^T)$ $\quad \det(\underline{A}^{-1}) = \det(\underline{A})^{-1}$

$\det(\underline{A}\underline{B}) = \det(\underline{A})\det(\underline{B}) = \det(\underline{B})\det(\underline{A}) = \det(\underline{B}\underline{A})$

Hat $\underline{A}$ 2 linear abhäng. Zeilen/Spalten $\Rightarrow |\underline{A}| = 0$

1.2.3. Eigenwerte (EW) λ und Eigenvektoren (EV) $\underline{v}$

$$\boxed{\underline{A}\underline{v} = \lambda\underline{v} \qquad \det\underline{A} = \prod \lambda_i \qquad \operatorname{Sp}\underline{A} = \sum a_{ii} = \sum \lambda_i}$$

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 A

$\det(\underline{A}) = ad - bc$ $\quad \begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1} = \frac{1}{\det\underline{A}}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$

$\operatorname{Sp}(\underline{A}) = a + d$

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

1.2.5. Differentiation

$\frac{\partial\underline{x}^\top\underline{y}}{\partial\underline{x}} = \frac{\partial\underline{y}^\top\underline{x}}{\partial\underline{x}} = \underline{y}$ $\quad \frac{\partial\underline{x}^\top\underline{A}\underline{x}}{\partial\underline{x}} = (\underline{A}+\underline{A}^\top)\underline{x}$

$\frac{\partial\underline{x}^\top\underline{A}\underline{y}}{\partial\underline{A}} = \underline{x}\underline{y}^\top$ $\quad \frac{\partial\det(\underline{B}\underline{A}\underline{C})}{\partial\underline{A}} = \det(\underline{B}\underline{A}\underline{C})\left(\underline{A}^{-1}\right)^\top$

1.3. Integrale $\int e^x\,\mathrm{d}x = e^x = (e^x)'$

Partielle Integration: $\int uw' = uw - \int u'w$

Substitution: $\int f(g(x))g'(x)\,\mathrm{d}x = \int f(t)\,\mathrm{d}t$

$$\int e^{at}\sin(bt)\,\mathrm{d}t = e^{at}\frac{a\sin(bt)+b\cos(bt)}{a^2+b^2}$$

$\int\frac{\mathrm{d}t}{\sqrt{at+b}} = \frac{2\sqrt{at+b}}{a}$ $\quad \int t^2e^{at}\,\mathrm{d}t = \frac{(ax-1)^2+1}{a^3}e^{at}$

$\int te^{at}\,\mathrm{d}t = \frac{at-1}{a^2}e^{at}$ $\quad \int xe^{ax^2}\,\mathrm{d}x = \frac{1}{2a}e^{ax^2}$

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: $\frac{n!}{k_1!\cdot k_2!\cdot\ldots}$

Binomialkoeffizient $\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!\cdot(n-k)!}$

$\binom{n}{0} = 1$ $\quad \binom{n}{1} = n$ $\quad \binom{4}{2} = 6$ $\quad \binom{5}{2} = 10$ $\quad \binom{6}{2} = 15$

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

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

2.3. Wahrscheinlichkeitsmaß P

$\mathsf{P}(A) = \frac{|A|}{|\Omega|}$ $\quad \mathsf{P}(A\cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A\cap B)$

2.3.1. Axiome von Kolmogorow

Nichtnegativität: $\mathsf{P}(A) \geq 0 \Rightarrow \mathsf{P}: \mathbb{F} \mapsto [0,1]$

Normiertheit: $\mathsf{P}(\Omega) = 1$

Additivität: $\mathsf{P}\left(\bigcup\limits_{i=1}^{\infty} A_i\right) = \sum\limits_{i=1}^{\infty}\mathsf{P}(A_i)$,

wenn $A_i \cap A_j = \emptyset, \forall i \neq j$

2.4. Bedingte Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit für A falls B bereits eingetreten ist:

$\mathsf{P}_B(A) = \mathsf{P}(A|B) = \frac{\mathsf{P}(A\cap B)}{\mathsf{P}(B)}$

2.4.1. Totale Wahrscheinlichkeit und Satz von Bayes

Es muss gelten: $\bigcup\limits_{i\in I} B_i = \Omega$ für $B_i \cap B_j = \emptyset, \forall i \neq j$

Totale Wahrscheinlichkeit: $\mathsf{P}(A) = \sum\limits_{i\in I}\mathsf{P}(A|B_i)\,\mathsf{P}(B_i)$

Satz von Bayes: $\mathsf{P}(B_k|A) = \frac{\mathsf{P}(A|B_k)\,\mathsf{P}(B_k)}{\sum\limits_{i\in I}\mathsf{P}(A|B_i)\,\mathsf{P}(B_i)}$

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

2.5. Distribution

Bezeichnung	Abk.	Zusammenhang
Wahrscheinlichkeitsdichte	pdf	$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$
Kumulative Verteilungsfkt.	cdf	$F_X(x) = \int\limits_{-\infty}^{x} f_X(\xi)\,\mathrm{d}\xi$

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

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

$$\underbrace{f_{X,Y}(x,y)}_{\text{Joint PDF}} = f_{X\,|\,Y}(x,y)f_Y(y) = f_{Y\,|\,X}(y,x)f_X(x)$$

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

2.7. Bedingte Zufallsvariablen

Ereignis A gegeben: $F_{X\,|\,A}(x|A) = \mathsf{P}(\{X \leq x\}|A)$

ZV Y gegeben: $F_{X\,|\,Y}(x|y) = \mathsf{P}(\{X \leq x\}|\{Y = y\})$

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

$f_{X\,|\,Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\mathrm{d}F_{X\,|\,Y}(x|y)}{\mathrm{d}x}$

2.8. Unabhängigkeit von Zufallsvariablen

$X_1, \cdots, X_n$ sind stochastisch unabhängig, wenn für jedes $\underline{x} \in \mathbb{R}^n$ gilt:

$F_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = \prod\limits_{i=1}^{n} F_{X_i}(x_i)$

$p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = \prod\limits_{i=1}^{n} p_{X_i}(x_i)$

$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = \prod\limits_{i=1}^{n} f_{X_i}(x_i)$

3. Gaussian Stuff

3.1. Gaussian Channel

Channel: $Y = hs_i + N$ with $h \sim \mathcal{N}, N \sim \mathcal{N}$

$L(y_1, \ldots, y_N) = \prod\limits_{i=1}^{n} f_{Y_i}(y_i, h)$

$f_{Y_i}(y_i, h) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y_i - hs_i)^2\right)$

$\hat{h}_{ML} = \operatorname*{argmin}\limits_h\{\left\|\underline{y} - h\underline{s}\right\|^2\} = \frac{\underline{s}^\top\underline{y}}{\underline{s}^\top\underline{s}}$

If multidimensional channel: $\underline{y} = \underline{S}\underline{h} + \underline{n}$:

$L(\underline{y}, \underline{h}) = \frac{1}{\sqrt{\det(2\pi\underline{C})}}\exp\left(-\frac{1}{2}(\underline{y} - \underline{S}\underline{h})^\top\underline{C}^{-1}(\underline{y} - \underline{S}\underline{h})\right)$

$l(\underline{y}, \underline{h}) = \frac{1}{2}\left(\log(\det(2\pi\underline{C}) - (\underline{y} - \underline{S}\underline{h})^\top\underline{C}^{-1}(\underline{y} - \underline{S}\underline{h})\right)$

$\frac{\mathrm{d}}{\mathrm{d}h}(\underline{y} - \underline{S}\underline{h})^\top\underline{C}^{-1}(\underline{y} - \underline{S}\underline{h}) = -2\underline{S}^\top\underline{C}^{-1}(\underline{y} - \underline{S}\underline{h})$

Gaussian Covariance: if $Y \sim \mathcal{N}(0, \sigma^2)$, $N \sim \mathcal{N}(0, \sigma^2)$:

$\underline{C}_Y = \mathrm{Cov}[Y, Y] = \mathsf{E}[(Y - \mu)(Y - \mu)^\top] = \mathsf{E}[Y\,Y^\top]$

For Channel $Y = Sh + N$: $\mathsf{E}[Y\,Y^\top] = S\,\mathsf{E}[hh^\top]S^\top + \mathsf{E}[NN^\top]$

3.2. Multivariate Gaussian Distributions

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

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

Affine transformations $\underline{\mathbf{y}} = \underline{A}\underline{\mathbf{x}} + \underline{b}$ are jointly Gaussian with

$\underline{\mathbf{y}} \sim \mathcal{N}(\underline{A}\underline{\mu}_{\underline{\mathbf{x}}} + \underline{b}, \underline{A}\underline{C}_{\underline{\mathbf{x}}}\underline{A}^\top)$

All marginal PDFs are Gaussian as well

Contour Lines

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

3.3. Conditional Gaussian

$\underline{A} \sim \mathcal{N}(\underline{\mu}_{\underline{A}}, \underline{C}_{\underline{A}})$, $\underline{B} \sim \mathcal{N}(\underline{\mu}_{\underline{B}}, \underline{C}_{\underline{B}})$

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

Conditional Mean:

$\mathsf{E}[\underline{A}|\underline{B} = \underline{b}] = \underline{\mu}_{\underline{A}|\underline{B}=\underline{b}} = \underline{\mu}_{\underline{A}} + \underline{C}_{\underline{AB}}\,\underline{C}_{\underline{BB}}^{-1}\left(\underline{b} - \underline{\mu}_{\underline{B}}\right)$

Conditional Variance:

$\underline{C}_{\underline{A}|\underline{B}} = \underline{C}_{\underline{AA}} - \underline{C}_{\underline{AB}}\,\underline{C}_{\underline{BB}}^{-1}\,\underline{C}_{\underline{BA}}$

3.4. Misc

If CDF of gaussian distribution given $\Phi(z) \sim \mathcal{N}(0,1)$ then for $X \sim \mathcal{N}(\mu, \sigma^2)$ the CDF is given as $\Phi(\frac{x-\mu}{\sigma})$

Gaussian: $\Phi(-z) = 1 - \Phi(z)$

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 $\quad k$: Anzahl der Erfolge

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

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

4.2. Normalverteilung

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

4.3. Sonstiges

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

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

5. Wichtige Parameter

5.1. Erwartungswert (1. zentrales Moment)

gibt den mittleren Wert einer Zufallsvariablen an

$$\boxed{\mu_X = \mathsf{E}[X] = \underbrace{\sum_{x\in\Omega'} x\cdot\mathsf{P}_X(x)}_{\text{diskrete } X:\Omega\to\Omega'} \;\hat{=}\; \underbrace{\int_{\mathbb{R}} x\cdot f_X(x)\,\mathrm{d}x}_{\text{stetige } X:\Omega\to\mathbb{R}}}$$

$\mathsf{E}[\alpha X + \beta Y] = \alpha\,\mathsf{E}[X] + \beta\,\mathsf{E}[Y]$ $\quad X \leq Y \Rightarrow \mathsf{E}[X] \leq \mathsf{E}[Y]$

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

$\mathsf{E}[X\,Y] = \mathsf{E}[X]\,\mathsf{E}[Y]$, falls X und Y stochastisch unabhängig

Umkehrung nicht möglichich: Unkorreliertheit $\not\Rightarrow$ Stoch. Unabhängig!

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

$\mathsf{E}[g(X)] = \sum\limits_{x\in\Omega'} g(x)\,\mathsf{P}_X(x) \;\hat{=}\; \int\limits_{\mathbb{R}} g(x)f_X(x)\,\mathrm{d}x$

5.2. Varianz (2. zentrales Moment)

ist ein Maß für die Stärke der Abweichung vom Erwartungswert

$$\boxed{\sigma_X^2 = \mathrm{Var}[X] = \mathsf{E}\left[(X - \mathsf{E}[X])^2\right] = \mathsf{E}[X^2] - \mathsf{E}[X]^2}$$

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

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

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

5.3. Kovarianz

Maß für den linearen Zusammenhang zweier Variablen

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

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

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

5.3.1. Korrelation = standardisierte Kovarianz

$\rho(X,Y) = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\cdot\mathrm{Var}[Y]}} = \frac{C_{x,y}}{\sigma_x \cdot \sigma_y} \qquad \rho(X,Y) \in [-1;1]$

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

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

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

6. Statistical Learning

6.1. Definition

Statistical Model

Statistical Model:	$\{\mathbb{X}, \mathbb{F}, \mathrm{P}_\theta; \theta \in \Theta\}$
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:

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

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

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

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

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

6.2. Maximum Likelihood Test

ML Ratio Test Statistic (Likelihood Ratio):

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

ML Test:

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

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

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

6.3. Neyman-Pearson-Test

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

$\operatorname{argmax}_{d_{\mathrm{NP}}} \mathrm{E}[d_{\mathrm{NP}}(x)|\theta = \theta_1] \quad$ s.t. $\quad E[d_{\mathrm{NP}}(x)|\theta = \theta_0] \le \alpha$

NP-Test to the error level α:

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

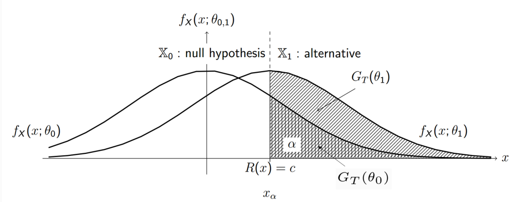
If $P(\{R(x) = c; \theta_0\}) = 0 \leftrightarrow$ (if x is continous):

$$T_{\mathrm{NP}}(x) = \begin{cases} 1 & R(x) > c \\ 0 & R(x) < c \end{cases} \qquad \text{Likelihood-Ratio: } R(x) = \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)}$$

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

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

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



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

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:

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

$$T_{\mathrm{Bayes}} = \operatorname*{argmin}_T\{P_\epsilon\} = \begin{cases} 1 & ; \quad \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > c = \frac{P(\theta_0)}{P(\theta_1)} \\ 0 & ; \quad \text{otherwise} \end{cases}$$

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

with :

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

if $P(\theta_0) = P(\theta_1) \to T_{\mathrm{Bayes}} = T_{\mathrm{ML}}$

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

$T_{\mathrm{Bayes}} = \operatorname*{argmin}_{k \in 1,\dots,K} \{P(\theta_k|x)\}$

Loss Function:

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

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

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

6.5. Linear Alternative Tests

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

$\log R(\underline{x}) = -\frac{1}{2}\ln(\frac{\det(\underline{C}_1)}{\det(\underline{C}_0)}) - \frac{1}{2}(\underline{x} - \underline{\mu}_1)^\top \underline{C}_1^{-1}(\underline{x} - \underline{\mu}_1) +$

$+ \frac{1}{2}(\underline{x} - \underline{\mu}_0)^\top \underline{C}_0^{-1}(\underline{x} - \underline{\mu}_0) = \ln(\frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)})$ (seperating 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 $C_0 \neq C_1$, $logR(x) = 0$ is non-linear and the separating surfaces are surfaces of second order: parabolic, hyperbolic, or elliptic surfaces.
- if $C_0 = C_1$, $logR(x) = 0$ is affine and thus defines a hyperplane in $\mathbb{X}$ which decomposes $\mathbb{X}$ into $\mathbb{X}_0$ and $\mathbb{X}_1$, i.e.,

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

  - case 1: $\underline{C}_0 = \underline{C}_1 = \sigma^2 \underline{I}_N$
    $\underline{w}^\top = (\underline{\mu}_1 - \underline{\mu}_0)^\top,$
    $w_0 = \frac{1}{2}(\underline{\mu}_1^\top \underline{\mu}_1 - \underline{\mu}_0^\top \underline{\mu}_0) - \sigma^2 \ln(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)})$
    $\underline{w}$ colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$
    $\to$ hyperplane orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$
  - case 2: $\underline{C}_0 = \underline{C}_1 = \underline{C}$
    $\underline{w}^\top = (\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}^{-1},$
    $w_0 = \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_0)^\top \underline{C}^{-1}(\underline{\mu}_1 + \underline{\mu}_0) - \ln(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)})$
    in general $\underline{w}$ **not** colinear with $(\underline{\mu}_1 - \underline{\mu}_0)$
    $\to$ hyperplane **not** orthogonal to $(\underline{\mu}_1 - \underline{\mu}_0)$
- if $C_0 = C_1$ and $\mu_0 = -\mu_1$, $logR(x) = 0$ is linear and defines a separating hyperplane in $\mathbb{X}$ which contains the origin, i.e.,

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

7. Hypothesis Testing

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)

Descision rule $\varphi: \mathbb{X} \to [0,1]$ with

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

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

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

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

Precision/Positive Prediciton rate: $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

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

7.1.1. Design of a test

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

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

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 $\to$ purely Data-Based.

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

linear combination of Kernel Function $g(., \mu_k)$, g() 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

Infernce 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)$\ge$c) decomposes sample space $\mathbb{X}$ into two disjoint subsets($\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1$)

Seperating surface between $\mathbb{X}_0$ and $\mathbb{X}_1$ given by:

$\{\mathbf{x}|R(\mathbf{x}) = c\}$ The relative postion of a sample x_j to the seperating surface determines choice of hypothesis

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

$x_i \in \mathbb{R}^N$, $y_i \in \{\Theta_0, \Theta_1\}$

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 statisfy: $\boxed{M \ge dim(\mathbb{X})}$

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

8.3. Linear Kernels

Test Statistic for linear test

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

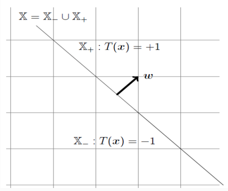
Hyperplane defined by $\mathbf{w}$(normal vector or weight vector) and w_o approximates seperating surface between $\mathbb{X}_-$ and $\mathbb{X}_+$

$\to$Decistion rule T(x):

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

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

(Orientation of w chosen such that w points into direction of θ_1("+1"hypothesis))



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

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

$\Rightarrow$ **Support Vector Methods**: $\boxed{y_k(\mathbf{w}^T \mathbf{x_k} + wo) \ge \epsilon, \forall k}$

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

8.4. Support Vector Methods

only feasible for normalized weight vectors

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

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

Optimization Problem convex $\rightarrow$ **Langragian Method**

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

Langragian Multiplier: $u_k \geq 0$

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

$\frac{\partial \Phi(\mathbf{w},\mathbf{u})}{\partial \mathbf{w}}|_{\mathbf{w}=\mathbf{w}(\mathbf{u})} \cdot = 0 \leftrightarrow \mathbf{w}(\mathbf{u}) = \sum_{k=1}^{M} \underbrace{u_k y_k}_{\lambda_k} \mathbf{x_k}$

Evaluate dual function:

$\Phi(\mathbf{w}(\mathbf{u}), \mathbf{u}) = \Phi(\sum_{k=1}^{M} u_k y_k \mathbf{x}_k, u_1 \ldots, u_M)$

$= -\frac{1}{2}\sum_{k=1}^{M}\sum_{l=1}^{M} u_k u_l y_k y_l \mathbf{x}_k^T \mathbf{x}_l + \sum_{k=1}^{M} u_k$

$= -\frac{1}{2}\mathbf{u^T Y X X^T Y u} + \mathbf{1^T u}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1^T} \\ \vdots \\ \mathbf{x_M^T} \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} y_1 & & \\ & \ddots & \\ & & y_M \end{bmatrix}, \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Alternative to approach above:

Iterative Solution:

Choose one element $\mathbf{x}_k$ out of sample set $\mathbb{S} = \{\mathbf{x_1}, \ldots, \mathbf{x_M}\}$ and randomly set:

$u_k \leftarrow u_k + \max_{\{} \eta \frac{\partial \phi(\mathbf{u})}{\partial u_k}, -u_k\}, \forall k$

Necessary and sufficient condition for existence of solution given by:

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

8.5. Suport Vectors

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

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

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

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

Elements with inactive dual variables dont contribute to Kernel Test

Optimal Weight Vektor $\mathbf{w}^* = \mathbf{w}(\mathbf{u}^*)$ of Kernel Test constructed by Support Vectors only: $\boxed{\mathbf{w}^* = \sum_{\mathbf{x}_k \in \mathbb{S}_{SV}} u_k^* y_k \mathbf{x}_k}$

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

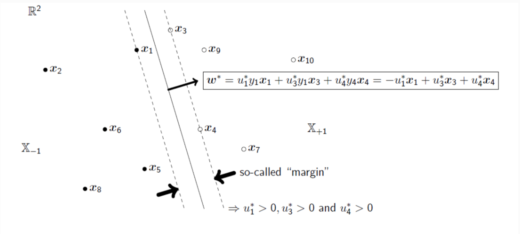


Fig. 2.2: The elements $x_k \in \mathbb{S}$ with ACTIVE DUAL VARIABLES $u_k^* > 0$ are called SUPPORT VECTORS.

Discussion

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

$\rightarrow$ instead of $\min_{\mathbf{w}} \frac{1}{2}\mathbf{w^T w}$ s.t. $y_k \mathbf{w^T x}_k \geq 1$ we get

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

8.6. Kernel Trick

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

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

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

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

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

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

In Optimization Problem and resulting Dual Function and Variables replace $\mathbf{x}$ by $\varphi(\mathbf{x}_k)$ $\rightarrow$ Dual OP: $\max_{\mathbf{u} \geq 0}\{-\mathbf{u^T Y G Y u} + \mathbf{1^T u}\}$

$$\text{Kernel Matrix G} = \begin{bmatrix} g(\mathbf{x_1}, \mathbf{x_2}) & \cdots & g(\mathbf{x_1}, \mathbf{x_M}) \\ \vdots & & \vdots \\ g(\mathbf{x_M}, \mathbf{x_1}) & \cdots & g(\mathbf{x_M}, \mathbf{x_M}) \end{bmatrix} \in \mathbb{R}^{MxM}$$

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

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

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

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

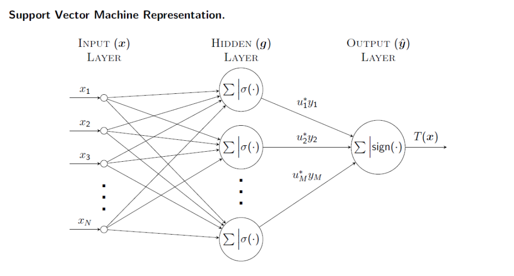


Fig. 2.4: The interpretation of a SUPPORT VECTOR MACHINE as a NEURAL NETWORK with three layers and a non-linear function σ. For POLYNOMIAL KERNELS each SINGLE HIDDEN LAYER UNIT is described by $g_{\text{poly}}(x, x_k) = \sigma(z_k)$, with $\sigma(z_k) = z_k^d$ and $z_k = x_k^T x + 1$.

9. Learning and Generalization

9.1. Empirical Risk Function and Generalization Error

ML scenarios (unknown Stochastical Model) base learning on:

$Risk_{emp}(T; \mathbb{S}) = \frac{1}{M}\sum_{i=1}^{M} L(T(\underline{\mathbf{x}}_i), y_i), \quad (\underline{\mathbf{x}}_i, y_i) \in \mathbb{S}$

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

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

$Risk = E_{S,X,Y}[L(T(X; S), Y)] = E_X[1 - P_{Y|X}(Y = T_B(X)) + \boxed{(1 - P_{S|X}(T(X; S) = T_B(X)))}(2P_{Y|X}(Y = T_B(X)) - 1)]$, $T_B(X)$ is the unknown *Bayes Test*

If the potential set $\mathbb{S}$ would be selected from a distribution such that the derived Test $T(\underline{\mathbf{x}}; \mathbb{S})$ and the corresponding Bayes Test $T_B(\underline{\mathbf{x}})$ 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 $\underline{\mathbf{x}}$ the Bayes Test $T_B(X)$ decides for the false label y).

10. Classification Trees and Random Forests

10.1. CART Algorithms

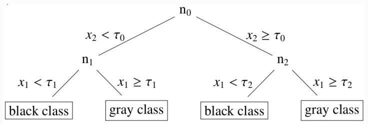
Generate Binary Trees by splitting $\mathbb{X}$ at each (internal/root) node: $\mathbb{X}_{i,left} = \{\underline{\mathbf{x}} \in \mathbb{X}_i | x_{j_i} < \tau_i\} \quad \mathbb{X}_{i,right} = \mathbb{X}_i \backslash \mathbb{X}_{i,left}$

Root/Internal node: Binary decision based on chosen threshold $\tau_i \in \mathbb{R}$, feature $x_{j_i} = [\underline{\mathbf{x}}]_{j_i}$ with $j_i \in \mathbb{J} = \{1, \ldots, dim[\mathbb{X}]\}$ 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

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

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

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

$\Rightarrow \{j_i, \tau_i\} = \operatorname{argmin}_{j \in \mathbb{J}, \tau \in \mathbb{R}}\left\{\sum_{k=1}^{K}\left(1 - \frac{M_k(\mathbb{S}_{i,left})}{M(\mathbb{S}_{i,left})}\right)\frac{M_k(\mathbb{S}_{i,left})}{M(\mathbb{S}_i)} + \left(1 - \frac{M_k(\mathbb{S}_{i,right})}{M(\mathbb{S}_{i,right})}\right)\frac{M_k(\mathbb{S}_{i,right})}{M(\mathbb{S}_i)}\right\}$

Overfitting(comes with high purity) can be controlled by a *Test Set* $\mathbb{S}_{Test}$.

Decision Rule: At terminal node n_i, input $\underline{\mathbf{x}}$ is assigned to $T_{CART}(\underline{\mathbf{x}}; \mathbb{S}) : \mathbb{X} \mapsto \{1, \ldots, K\}, \underline{\mathbf{x}} \mapsto \operatorname{argmax}_k\{M_k(\mathbb{S}_i)\}$

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

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

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

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

11. Deep Neural Networks

11.1. From Kernel to Neural Networks (NN)

NN: methodology by which KERNELS are determined by chosen learning method based on the available training data $\rightarrow$ KERNELS are composed by a concatenation of multiple VECTOR VALUED functions

$$\boxed{g(x) = f^{(L)}(f^{(L-1)}(\ldots f^{(2)}(f^{(1)}(x; W^{(1)}, v^{(1)}); W^{(2)}, v^{(2)})\ldots; W^{(L-1)}, v^{(L-1)}); W^{(L)}, v^{(L)})}$$

$f^{(l)}(*; W^{(l)}, v^{(l)}) \in \mathbb{R}^{N_t}$ represents the l-th layer of NN

NN consist of L+2 layers (INPUT Layer $x \in \mathbb{R}^N$ and LAYER OF OUTPUTS $f^{(NN)} \in \mathbb{R}^{N_{L+1}}$

HIDDEN LAYER (L=1) often enough

If $L > 1$ NN is called **DEEP**

Mapping between NN layers consists typically of AFFINE TRANSFORMATION of the output of the preceding layer;

$\mathbb{R}^{N_{t-1}} \rightarrow \mathbb{R}^{N_t} : f^{(l-1)} \rightarrow^{(l)} = W^{(l),T} f^{(l-1)} + v(l)$,

and the elementwise NONLINEAR TRANSFORMATION of the resulting INTERNAL STATE VECTOR $z^{(l)}$ by means of a NONLINEAR FUNCTION $\sigma^{(l)}$

$$\boxed{f^{(l)}(f^{(l-1)}; W^l), v^{(l)}) = \sigma^{(l)}(W^{(l),T} f^{(l-1)} + v^{(l)})}$$

Elements of $^{(l)}$ and $v^{(l)}$ are called weights of the lth NN layer

- INPUT LAYER (l=0) of NN equals INPUT VECTOR $x \in \mathbb{R}^N$
- OUTPUT LAYER (l=L+1) of NN equals OUTPUT VECTOR $f^{(NN)} \in \mathbb{R}^{N_{L+1}}$
- NONLINEAR FUNCTION $\sigma_i^{(l)}$ of the HIDDEN LAYERS is different from the OUTPUT FUNCTION of the OUTPUT LAYER
- latter depends on LOSS FUNCTION and the chosen LEARNING ALGORITHM

Single nonlinear function of the output vector of the previous layer composed by the i-th LINEAR FUNCTIONAL $w_i^{(l)}$, the CONSTANT $v_i^{(l)}$ and the i-th nonlinear function $\sigma^{(l)}$ of the next layer = NEURON. WEIGHTS represent the SYNAPTIC STRENGHTS and the nonlinear function $\sigma_i^{(l)}$ = ACTIVATION FUNCTION

$$\boxed{\sigma_i^{(l)}(\sum_{j=1}^{N(l-1)} w_{i,j}^{(l)} f_j^{(l-1)} + v_i^{(l)})}$$

Single Neuron Representation.

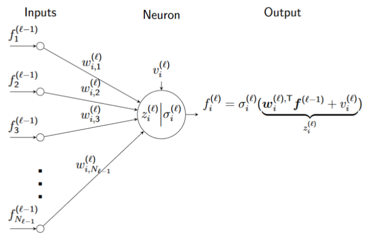


Fig. 5.1: A single neuron representation of the i-th output element of the ℓ-th network layer

Neural Network.

Representation of a FEEDFORWARD NEURAL NETWORK – aka MULTILAYER PERCEPTRON (MLP)

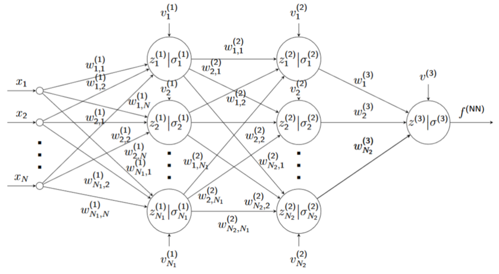


Fig. 5.2: DEEP NN with an input layer, two hidden layers, and one output function.

11.2. Activation Functions

ReLU Activation Functions

most popular chose for the activation function$\sigma_i^{(l)}$ $\rightarrow$ RECTIFIED LINEAR UNIT FUNCTION (RELU)

$$\sigma(z_i^{(l)}) = max(0, z_i^{(l)}) \in \mathbb{R}_+$$
with $z_i(l) = \sum_{j=1}^{N_{l-1}} w_{i,j}^{(l)} f_j^{(l-1)} + v_i^{(l)}$

- PIECEWISE LINEAR FUNCTION which is zero for a negative state variable
- efficient for the training of network weights, since its gradient with respect to the weight parameters does not experience any saturation for large positive values of the state variable, i.e.

$$\frac{\partial\sigma(z_i^{(l)})}{\partial w_{i,j}^{(l)}} = \frac{\partial\sigma(z_i^{(l)})}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial w_{i,j}^{(l)}} = unit(z_i^{(l)} f_j^{(l-1)}) \text{ and}$$
$$\frac{\partial\sigma(z_i^{(l)})}{\partial v_{i,j}^{(l)}} = \frac{\partial\sigma(z_i^{(l)})}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial v_{i,j}^{(l)}} = unit(z_i^{(l)})$$
with the UNIT STEP FUNCTION unit(z)$\in 0, 1$

Hyperbolic Tangent Activation Functions

Used to be standard before RELU

$$\sigma(z_i^{(l)}) = tanh(z_i^{(l)}) = \frac{e^{z_i^{(l)}} - e^{-z_i^{(l)}}}{e^{z_i^{(l)}} + e^{-z_i^{(l)}}} \in [-1, +1]$$
with $z_i^{(l)} = \sum_{j=1}^{N_{l-1}} w_{i,j}^{(l)} f_j^{(l-1)} + v_i^{(l)}$

The HYPERBOLIC TANGENT FUNCTION suffers from a saturation of its gradient with respect to weight parameters for large absolute values of the state variable, i.e.

$$\frac{\partial\omega(z_i^{(l)})}{\partial w_{i,j}^{(l)}} = \frac{\partial\omega(z_i^{(l)})}{\partial z_i^{(l)}} \frac{\partial z_i^{(l)}}{\partial w_{i,j}^{(l)}} = (1 - tanh^2(z_i^{(l)}))f_j^{(l-1)} \text{and}$$
$$\frac{\partial\omega(z_i^{(l)})}{\partial v_i^{(l)}} = (1 - tanh^2(z_i^{(l)}))$$

Advantage: for small values of the state variable near $z_i^{(l)} = 0$ the HYPERBOLIC TANGENT FUNCTION resembles a LINEAR MODEL

HYPERBOLIC TANGENT FUNCTION is very similiar to s.c. SIGMOID FUNCTION $\omega_{SIGMOID}(z_i^{(l)}) = \frac{1}{1+e^{-z_i^{(l)}}}$

$\rightarrow tanh(z_i^{(l)}) = 2\sigma_{SIGMOID}(2z_i^{(l)}) - 1$