Signal Processing and Machine Learning

1. Statistical Learning

1.1. 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: $G_T(\theta_0) \leq \alpha$

Two Error Types:
False Alarm: $\theta = \theta_0, T(x) = 1$
$G_T(\theta_0) = P(\{T(X) = 1\}; \theta_0)$
Detection Error: $\theta = \theta_1, T(x) = 0$
$1 - G_T(\theta_1) = P(\{T(X) = 0\}; \theta_1)$

1.2. 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 $R(x) \geq c(\alpha)$ 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

1.3. Neyman-Pearson-Test

minimizes the detection error, while fulfilling a predefined error level α
$\operatorname{argmax}_{d_{\mathrm{NP}}} \mathrm{E}[d_{\mathrm{NP}}(x)|\theta = \theta_1]$ s.t. $E[d_{\mathrm{NP}}(x)|\theta = \theta_0] \leq \alpha$

NP-Test to the error level α:
x_α is chosen as: $x_\alpha = (1-\alpha)$-quantile of $f_x(x;\theta_0)$

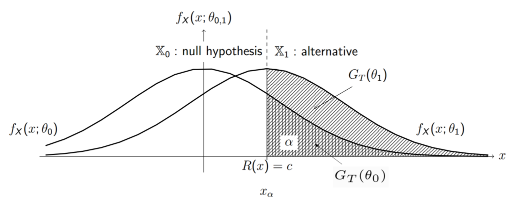
If $P(\{R(x) = c; \theta_0\}) = 0 \leftrightarrow$ (if x is continous):

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

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 $G_T(\theta_0)$

1.4. 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 $P(\theta_0) = P(\theta_1) \rightarrow T_{\mathrm{Bayes}} = T_{\mathrm{ML}}$

Multiple Hypothesis $\{\theta_0, \ldots, \theta_k\}; \mathbb{X}_0, \ldots, \mathbb{X}_k \in \mathbb{X}$:
$T_{\mathrm{Bayes}} = \operatorname*{argmin}_{k \in 1,\ldots,K} \{P(\theta_k|x)\}$

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

1.5. Linear Alternative Tests

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

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

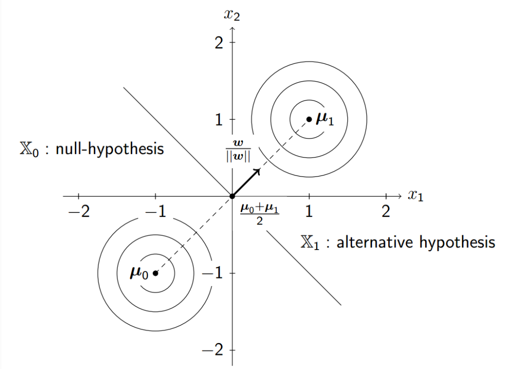
For Gaussian $f_X(x; \mu_k, C_k)$ with θ_0 and θ_1 corresponding to $\{\mu_0, C_0\}$ and $\{\mu_1, C_1\}$, it follows that

- 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.,

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

  - case 1: $\underline{\boldsymbol{C}}_0 = \underline{\boldsymbol{C}}_1 = \sigma^2 \underline{\boldsymbol{I}}_N$
    $\underline{\boldsymbol{w}}^\top = (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^\top$,
    $w_0 = \frac{1}{2}(\underline{\boldsymbol{\mu}}_1^\top \underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0^\top \underline{\boldsymbol{\mu}}_0) - \sigma^2 \ln\!\left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)}\right)$
    $\underline{\boldsymbol{w}}$ colinear with $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$
    $\rightarrow$ hyperplane orthogonal to $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$
  - case 2: $\underline{\boldsymbol{C}}_0 = \underline{\boldsymbol{C}}_1 = \underline{\boldsymbol{C}}$
    $\underline{\boldsymbol{w}}^\top = (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^\top \underline{\boldsymbol{C}}^{-1}$,
    $w_0 = \frac{1}{2}(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^\top \underline{\boldsymbol{C}}^{-1}(\underline{\boldsymbol{\mu}}_1 + \underline{\boldsymbol{\mu}}_0) - \ln\!\left(\frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)}\right)$
    in general $\underline{\boldsymbol{w}}$ **not** colinear with $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$
    $\rightarrow$ hyperplane **not** orthogonal to $(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)$
- 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.,

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



2. Learning and Generalization

2.1. 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*

2.2. Bias-Variance Decomposition

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

3. Classification Trees and Random Forests

3.1. 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
$\Rightarrow \underline{\mathbf{x}} \mapsto n_i(\underline{\mathbf{x}})$

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: $\boxed{I_{CART}} =$

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

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

3.2. 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

4. Hypothesis Testing

making a decision based on the observations

4.1. 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 $\mathsf{E}[d(X)|\theta] \leq \alpha, \forall \theta \in \Theta_0$

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

Power: Sensitivity/Recall/Hit Rate: $\frac{\mathsf{TP}}{\mathsf{TP+FN}} = 1-\beta$
Specificity/True negative rate: $\frac{\mathsf{TN}}{\mathsf{FP+TN}} = 1-\alpha$
Precision/Positive Prediciton rate: $\frac{\mathsf{TP}}{\mathsf{TP+FP}}$
Accuracy: $\frac{\mathsf{TP+TN}}{\mathsf{P+N}} = \frac{2-\alpha-\beta}{2}$

4.1.1. Design of a test
Cost criterion $G_\varphi : \Theta \to [0,1], \theta \mapsto \mathsf{E}[d(X)|\theta]$
False Positive lower than α: $G_d(\theta)|_{\theta\in\Theta_0} \leq \alpha, \forall \theta \in \Theta_0$
False Negative small as possible: $\max\{G_d(\theta)|_{\theta\in\Theta_1}\}, \forall \theta \in \Theta_1$

4.2. 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:
$f_{X\,|\,T}(x|T(x) = t, \theta) = f_{X\,|\,T}(x|T(x) = t)$

5. Math

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

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

Folgen und Reihen
$\sum\limits_{k=1}^{n} k = \frac{n(n+1)}{2}$ Aritmetrische Summenformel $\qquad \sum\limits_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q}$ Geometrische Summenformel $\qquad \sum\limits_{n=0}^{\infty} \frac{\underline{z}^n}{n!} = e^{\underline{z}}$ Exponentialreihe

Mittelwerte $\quad$ ($\sum$ von i bis N) $\quad$ (Median: Mitte einer geordneten Liste)
$\overline{x}_{\mathsf{ar}} = \frac{1}{N}\sum x_i$ (Arithmetisches) $\geq$ $\overline{x}_{\mathsf{geo}} = \sqrt[N]{\prod x_i}$ (Geometrisches Mittel) $\geq$ $\overline{x}_{\mathsf{hm}} = \frac{N}{\sum \frac{1}{x_i}}$ (Harmonisches)

Ungleichungen: $\qquad$ Bernoulli-Ungleichung: $(1+x)^n \geq 1 + nx$
$||x| - |y|| \leq |x \pm y| \leq |x| + |y|$ Dreiecksungleichung $\qquad \left|\underline{x}^\top \cdot \underline{y}\right| \leq \|\underline{x}\| \cdot \|\underline{y}\|$ Cauchy-Schwarz-Ungleichung

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

5.1. Exp. und Log. $\quad e^x := \lim\limits_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$ $\qquad e \approx 2{,}71828$
$a^x = e^{x \ln a}$ $\qquad \log_a x = \frac{\ln x}{\ln a}$ $\qquad \ln x \leq x - 1$
$\ln(x^a) = a\ln(x)$ $\qquad \ln(\frac{x}{a}) = \ln x - \ln a$ $\qquad \log(1) = 0$

5.2. 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)
$(\underline{A} + \underline{B})^\top = \underline{A}^\top + \underline{B}^\top$ $\qquad (\underline{A}\cdot\underline{B})^\top = \underline{B}^\top \cdot \underline{A}^\top$
$(\underline{A}^\top)^{-1} = (\underline{A}^{-1})^\top$ $\qquad (\underline{A}\cdot\underline{B})^{-1} = \underline{B}^{-1}\underline{A}^{-1}$
$\dim \mathbb{K} = n = \operatorname{rang} \underline{A} + \dim\ker \underline{A}$ $\qquad \operatorname{rang} \underline{A} = \operatorname{rang} \underline{A}^\top$

5.2.1. Quadratische Matrizen $\underline{A} \in \mathbb{K}^{n\times n}$
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: $\underline{A} = \underline{A}^\top$ $\qquad$ schiefsymmetrisch: $\underline{A} = -\underline{A}^\top$

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

5.2.3. 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.

5.2.4. Spezialfall 2×2 Matrix A
$\det(\underline{A}) = ad - bc$ $\qquad \begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1} = \frac{1}{\det \underline{A}}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$
$\operatorname{Sp}(\underline{A}) = a + d$
$\lambda_{1/2} = \frac{\operatorname{Sp}\underline{A}}{2} \pm \sqrt{\left(\frac{\operatorname{sp}\underline{A}}{2}\right)^2 - \det \underline{A}}$

5.2.5. Differentiation
$\frac{\partial \underline{x}^\top\underline{y}}{\partial \underline{x}} = \frac{\partial \underline{y}^\top\underline{x}}{\partial \underline{x}} = \underline{y}$ $\qquad \frac{\partial \underline{x}^\top\underline{A}\underline{x}}{\partial \underline{x}} = (\underline{A} + \underline{A}^\top)\underline{x}$
$\frac{\partial \underline{x}^\top\underline{A}\underline{y}}{\partial \underline{A}} = \underline{x}\underline{y}^\top$ $\qquad \frac{\partial \det(\underline{B}\underline{A}\underline{C})}{\partial \underline{A}} = \det(\underline{B}\underline{A}\underline{C})\left(\underline{A}^{-1}\right)^\top$

5.2.6. Ableitungsregeln ($\forall \lambda, \mu \in \mathbb{R}$)
Linearität: $(\lambda f + \mu g)'(x) = \lambda f'(x) + \mu g'(x_0)$
Produkt: $(f \cdot g)'(x) = f'(x)g(x) + f(x)g'(x)$
Quotient: $\left(\frac{f}{g}\right)'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$ $\quad \left(\frac{\mathsf{NAZ-ZAN}}{\mathsf{N}^2}\right)$
Kettenregel $(f(g(x)))' = f'(g(x))g'(x)$

5.3. Integrale $\int e^x\,\mathrm{d}x = e^x = (e^x)'$
Partielle Integration: $\int uw' = uw - \int u'w$
Substitution: $\int f(g(x))g'(x)\,\mathrm{d}x = \int f(t)\,\mathrm{d}t$

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

$\int e^{at}\sin(bt)\,\mathrm{d}t = e^{at}\frac{a\sin(bt) + b\cos(bt)}{a^2+b^2}$
$\int \frac{\mathrm{d}t}{\sqrt{at+b}} = \frac{2\sqrt{at+b}}{a}$ $\qquad \int t^2e^{at}\,\mathrm{d}t = \frac{(ax-1)^2+1}{a^3}e^{at}$
$\int te^{at}\,\mathrm{d}t = \frac{at-1}{a^2}e^{at}$ $\qquad \int xe^{ax^2}\,\mathrm{d}x = \frac{1}{2a}e^{ax^2}$

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

6. Support Vector Machines
Motivation and Background

6.1. Kernel Methods
Kernel Methods is non-parametic estimation, these make no assumption on statistical model $\to$ purely Data-Based.

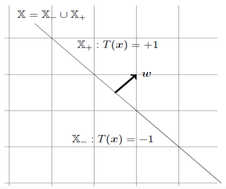
Test Statistic $\boxed{\mathbb{X} \to \mathbb{R}, \mathbf{x} \mapsto S(\mathbf{x}) = \sum\limits_{k=1}^{M}\lambda_k g(\mathbf{x}, \mu_\mathbf{k})}$

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

μ_k: representative for Sample Set $\mathbb{S} = \{x_1, \ldots, x_M\}$
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**

6.2. 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($\mathbb{X} = \mathbb{X}_0 \cup \mathbb{X}_1$)
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
$\boxed{\mathbb{S} = \{(x_1, y_1), \ldots, (x_M, y_M)\}}$
$x_i \in \mathbb{R}^N$, $y_i \in \{\Theta_0, \Theta_1\}$
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: $\boxed{M \geq dim(\mathbb{X})}$
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

6.3. Linear Kernels
Test Statistic for linear test
$\boxed{S(x) = \sum\limits_{i=1}^{M}\lambda_i {\mathbf{x_i}}^T\mathbf{x} + wo = \mathbf{w}^T\mathbf{x} + wo}$ $\quad \mathbf{w} = \sum_{i=1}^{M}\lambda_i x_i$
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):
$$\boxed{T(\mathbf{x}) = sign(S(\mathbf{x})) = \begin{cases} +1 & ; \quad \mathbf{w}^T\mathbf{x} + wo \geq 0\\ -1 & ; \quad otherwise\end{cases}}$$
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:
$\forall k \in \{1, \ldots M\} : T(\mathbf{x}_k) = y_k$

$\Rightarrow$ **Support Vector Methods**: $\boxed{y_k(\mathbf{w}^T\mathbf{x_k} + wo) \geq \epsilon, \forall k}$
Robust solution: maximize margin ϵ for constant norm of $\mathbf{w}$

Application

6.4. Support Vector Methods
only feasible for normalized weight vectors
$\max\limits_{w} \epsilon$ s.t. $y_k\frac{\mathbf{w}^T}{\|\mathbf{w}\|_2}\mathbf{x}_k \geq \epsilon, \forall k$, $w_0 = 0$
$\Leftrightarrow \min\limits_{w} \frac{1}{2}\|\mathbf{w}\|_2^2$ s.t. $y_k\mathbf{w}^T\mathbf{x}_k \geq 1, \forall k$
Optimization Problem convex $\to$ **Langragian Method**

Dual Problem: $\max\limits_{\mathbf{u}}\min\limits_{\mathbf{w}} \Phi(\mathbf{w}, \mathbf{u})$ 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:
$\Phi(\mathbf{w}(\mathbf{u}), \mathbf{u}) = \Phi(\sum_{k=1}^{M}u_k y_k\mathbf{x}_k, u_1 \ldots, u_M)$
$= -\frac{1}{2}\sum_{k=1}^{M}\sum_{l=1}^{M}u_k u_l y_k y_l \mathbf{x}_k^T\mathbf{x}_l + \sum_{k=1}^{M}u_k$
$= -\frac{1}{2}\mathbf{u^T YXX^T Yu} + \mathbf{1^T u}$
$\mathbf{X} = \begin{bmatrix}\mathbf{x_1^T}\\ \vdots\\ \mathbf{x_M^T}\end{bmatrix}, \mathbf{Y} = \begin{bmatrix}y_1 & & \\ & \ddots & \\ & & y_M\end{bmatrix}, \mathbf{1} = \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$
Alternativ 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:
$\mathbf{1} \in \operatorname{conce}[\mathbf{YXX^TY}]$

6.5. Suport Vectors
Dual OP.:$\max\limits_{\mathbf{u}}\sum_{k=1}^{M}(-\frac{1}{2}\sum_{l=1}^{M}u_k u_l y_k y_l \mathbf{x_k^T}\mathbf{x_l} + u_k)$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: $\boxed{\mathbf{w}^* = \sum\limits_{\mathbf{x}_k\in\mathbb{S}_{SV}} u_k^* y_k\mathbf{x}_k}$

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

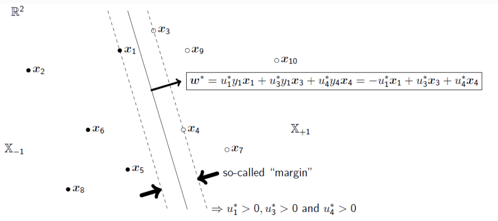


Fig. 2.2: The elements $x_k \in \mathbb{S}$ with Active Dual Variables $u_k^* > 0$ are called Support Vectors.

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

$\to$ instead of $\min\limits_{\mathbf{w}}\frac{1}{2}\mathbf{w^T w}$ s.t. $y_k\mathbf{w^T x_k} \geq 1$ we get
$\min\limits_{\mathbf{w},\epsilon}\frac{1}{2}\mathbf{w^T w} + \rho\sum_{k=1}^{M}\epsilon_k$ s.t.$y_k\mathbf{w^T x_k} \geq 1 - \epsilon_k, \forall k, \underline{\epsilon}, \rho \geq 0$

6.6. 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** $\varphi : \mathbb{S} \rightarrow \mathbb{S}_\varphi$

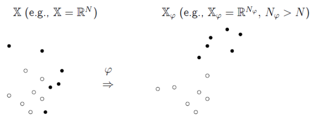


Fig. 2.3: Transfer the problem statement into a higher-dimensional (inner product) space without introducing additional degrees of freedom by means of a so-called FEATURE MAP $\varphi : \mathbb{S} \rightarrow \mathbb{S}_\varphi$.

Construction of Linear Test in $\mathbb{R}^3$ 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: $\max_{\mathbf{u} \geq 0} \{-\mathbf{u}^\mathbf{T}\mathbf{Y}\mathbf{G}\mathbf{Y}\mathbf{u} + \mathbf{1}^\mathbf{T}\mathbf{u}\}$

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

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

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.

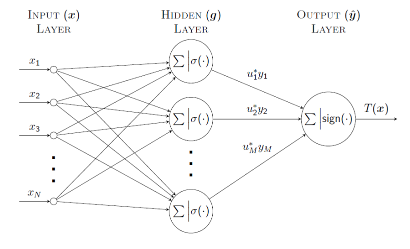


Fig. 2.4: The interpretation of a SUPPORT VECTOR MACHINE as a NEURAL NETWORK with three layers and a non-linear function σ. For POLYNOMIAL KERNELS each SINGLE HIDDEN LAYER UNIT is described by $g_{poly}(x, x_k) = \sigma(z_k)$, with $\sigma(z_k) = z_k^d$ and $z_k = x_k^T x + 1$.

7. Probability Theory Basics

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

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

7.2. 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	$\mathsf{P} : \mathbb{F} \rightarrow [0, 1]$	$\mathsf{P}(A) = \frac{\|A\|}{\|\Omega\|}$

7.3. Wahrscheinlichkeitsmaß P

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

7.3.1. Axiome von Kolmogorow

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

Normiertheit: $P(\Omega) = 1$

Additivität: $\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$, wenn $A_i \cap A_j = \emptyset$, $\forall i \neq j$

7.4. Bedingte Wahrscheinlichkeit

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

7.4.1. 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: $\mathsf{P}(A) = \sum_{i \in I} \mathsf{P}(A|B_i)\,\mathsf{P}(B_i)$

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

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

7.5. 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 $\{\omega \in \Omega | X(\omega) \in A'\} \in \mathbb{F}$

7.6. Distribution

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

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

7.7. 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

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

7.8. Bedingte Zufallsvariablen

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

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

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

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

7.9. Unabhängigkeit von Zufallsvariablen

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

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

8. Common Distributions

8.1. Binomialverteilung $\mathcal{B}(n, p)$ mit $p \in [0, 1], n \in \mathbb{N}$

Folge von n Bernoulli-Experimenten

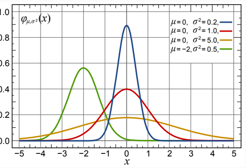
p: Wahrscheinlichkeit für Erfolg $\quad k$: Anzahl der Erfolge

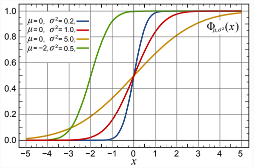
$$p_X(k) = B_{n,p}(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0, \ldots, n\} \\ 0 & \text{sonst} \end{cases}$$

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

8.2. Normalverteilung

WDF/PDF: KVF/CDF:





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

$E(X) = \mu$ Erwartungswert	$\mathsf{Var}(X) = \sigma^2$ Varianz	$\varphi_X(\omega) = e^{j\omega\mu - \frac{\omega^2\sigma^2}{2}}$ Charakt. Funktion

8.3. Sonstiges

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

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

9. Wichtige Parameter

9.1. Erwartungswert (1. zentrales Moment)

gibt den mittleren Wert einer Zufallsvariablen an

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

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

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

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

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

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

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

9.2. Varianz (2. zentrales Moment)

ist ein Maß für die Stärke der Abweichung vom Erwartungswert

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

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

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

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

9.3. Kovarianz

Maß für den linearen Zusammenhang zweier Variablen

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

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

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

9.3.1. Korrelation = standardisierte Kovarianz

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

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

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

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

10. Estimation

10.1. Estimation

Statistic Estimation treats the problem of inferring underlying characteristics of unknown random variables on the basis of observations of outputs of those random variables.

Sample Space Ω	nonempty set of outputs of experiment
Sigma Algebra $\mathbb{F} \subseteq 2^\Omega$	set of subsets of outputs (events)
Probability $\mathsf{P} : \mathbb{F} \mapsto [0, 1]$	
Random Variable $X : \Omega \mapsto \mathbb{X}$	mapped subsets of Ω
Observations: $x_1, \ldots, x_N$	single values of X
Observation Space $\mathbb{X}$	possible observations of X
Unknown parameter $\theta \in \Theta$	parameter of propability function
Estimator $T : \mathbb{X} \mapsto \Theta$	$T(X) = \hat{\theta}$, finds $\hat{\theta}$ from X

unknown parm. θ	estimation of param. $\hat{\theta}$
R.V. of param. Θ	estim. of R.V. of parm $T(X) = \hat{\Theta}$

10.2. Quality Properties of Estimators

Consistent: If $\lim_{N\to\infty} T(x_1,\dots,x_N) = \theta$

Bias $\mathrm{Bias}(T) := \mathsf{E}[T(X_1,\dots,X_N)] - \theta$

unbiased if $\mathrm{Bias}(T) = 0$ (biased estimators can provide better estimates than unbiased estimators.)

Variance $\mathsf{Var}[T] := \mathsf{E}\left[(T - \mathsf{E}[T])^2\right]$

10.3. Mean Square Error (MSE)

The MSE is an extension of the Variance $\mathsf{Var}[T] := \mathsf{E}\left[(T - \mathsf{E}[T])^2\right]$:

$$\text{MSE: } \varepsilon[T] = \underbrace{\mathsf{E}\left[(T-\theta)^2\right]}_{=\mathsf{E}[(\hat\theta-\theta)^2]} = \mathsf{Var}(T) + (\mathrm{Bias}[T])^2$$

If Θ is also r.v. $\Rightarrow$ mean over both (e.g. Bayes est.):

Mean MSE: $\mathsf{E}[(T(X)-\Theta)^2] = \mathsf{E}\left[\mathsf{E}\left[(T(X)-\Theta)^2|\Theta=\theta\right]\right]$

10.3.1. Minimum Mean Square Error (MMSE)

Minimizes mean square error: $\arg\min_{\hat\theta} \mathsf{E}\left[(\hat\theta-\theta)^2\right]$

$\mathsf{E}\left[(\hat\theta-\theta)^2\right] = \mathsf{E}[\theta^2] - 2\hat\theta\,\mathsf{E}[\theta] + \hat\theta^2$

Solution: $\frac{\mathrm{d}}{\mathrm{d}\hat\theta}\mathsf{E}\left[(\hat\theta-\theta)^2\right] \overset{!}{=} 0 = -2\,\mathsf{E}[\theta] + 2\hat\theta \quad \Rightarrow \hat\theta_{\mathsf{MMSE}} = \mathsf{E}[\theta]$

10.4. Maximum Likelihood

Given model $\{\mathbb{X}, \mathbb{F}, \mathsf{P}_\theta; \theta \in \Theta\}$, assume $\mathsf{P}_\theta(\underline{\boldsymbol{x}})$ or $f_{\underline{X}}(\underline{\boldsymbol{x}}, \theta)$ for observed data $\underline{\boldsymbol{x}}$. Estimate parameter θ so that the likelihood $L(\underline{\boldsymbol{x}}, \theta)$ or $L(\theta|\,X = \underline{\boldsymbol{x}})$ to obtain $\underline{\boldsymbol{x}}$ is maximized.

Likelihood Function: (Prob. for θ given $\underline{\boldsymbol{x}}$)

Discrete: $L(x_1,\dots,x_N;\theta) = \mathsf{P}_\theta(x_1,\dots,x_N)$

Continuous: $L(x_1,\dots,x_N;\theta) = f_{X_1,\dots,X_N}(x_1,\dots,x_N,\theta)$

If N observations are Identically Independently Distributed (i.i.d.):

$$L(\underline{\boldsymbol{x}},\theta) = \prod_{i=1}^N \mathsf{P}_\theta(x_i) = \prod_{i=1}^N f_{X_i}(x_i)$$

ML Estimator (Picks θ): $T_{\mathsf{ML}} : X \mapsto \underset{\theta\in\Theta}{\operatorname{argmax}}\{L(X,\theta)\} =$

$= \underset{\theta\in\Theta}{\operatorname{argmax}}\{\log L(\underline{X},\theta)\} \overset{\text{i.i.d.}}{=} \underset{\theta\in\Theta}{\operatorname{argmax}}\{\sum \log L(x_i,\theta)\}$

Find Maximum: $\frac{\partial L(\underline{\boldsymbol{x}},\theta)}{\partial\theta} = \frac{\mathrm{d}}{\mathrm{d}\theta}\log L(x;\theta)\Big|_{\theta=\hat\theta} \overset{!}{=} 0$

Solve for θ to obtain ML estimator function $\hat\theta_{\mathsf{ML}}$

Check quality of estimator with MSE

Maximum-Likelihood Estimator is Asymptotically Efficient. However, there might be not enough samples and the likelihood function is often not known.

10.5. Uniformly Minimum Variance Unbiased (UMVU) Estimators (Best unbiased estimators)

Best unbiased estimator: Lowest Variance of all estimators.

Fisher's Information Inequality: Estimate lower bound of variance if

- $L(x,\theta) > 0, \forall x, \theta$
- $L(x,\theta)$ is diffable for θ
- $\int_{\mathbb{X}} \frac{\partial}{\partial\theta} L(x,\theta)\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int_{\mathbb{X}} L(x,\theta)\,\mathrm{d}x$

Score Function:

$g(x,\theta) = \frac{\partial}{\partial\theta}\log L(x,\theta) = \frac{\frac{\partial}{\partial\theta}L(x,\theta)}{L(x,\theta)} \qquad \mathsf{E}[g(x,\theta)] = 0$

Fischer Information:

$I_{\mathsf{F}}(\theta) := \mathsf{Var}[g(X,\theta)] = \mathsf{E}[g(x,\theta)^2] = -\,\mathsf{E}\left[\frac{\partial^2}{\partial\theta^2}\log L(X,\theta)\right]$

Cramér-Rao Lower Bound (CRB): (if T is unbiased)

$$\mathsf{Var}[T(X)] \geq \left(\frac{\partial\,\mathsf{E}[T(X)]}{\partial\theta}\right)^2 \frac{1}{I_{\mathsf{F}}(\theta)} \qquad \mathsf{Var}[T(X)] \geq \frac{1}{I_{\mathsf{F}}(\theta)}$$

For N i.i.d. observations: $I_{\mathsf{F}}^{(N)}(x,\theta) = N\cdot I_{\mathsf{F}}^{(1)}(x,\theta)$

10.5.1. Exponential Models

If $f_X(x) = \frac{h(x)\exp\left(a(\theta)t(x)\right)}{\exp(b(\theta))}$ then $I_F(\theta) = \frac{\partial a(\theta)}{\partial\theta}\frac{\partial E[t(X)]}{\partial\theta}$

Some Derivations: (check in exam)

Uniformly: Not diffable $\Rightarrow$ no $I_F(\theta)$

Normal $\mathcal{N}(\theta,\sigma^2)$: $g(x,\theta) = \frac{(x-\theta)}{\sigma^2} \qquad I_{\mathsf{F}}(\theta) = \frac{1}{\sigma^2}$

Binomial $\mathcal{B}(\theta,K)$: $g(x,\theta) = \frac{x}{\theta} - \frac{K-x}{1-\theta} \qquad I_{\mathsf{F}}(\theta) = \frac{K}{\theta(1-\theta)}$

10.6. Bayes Estimation (Conditional Mean)

A Priori information about θ is known as probability $f_\Theta(\theta;\sigma)$ with random variable Θ and parameter σ. Now the conditional pdf $f_{X|\Theta}(x,\theta)$ is used to find θ by minimizing the mean MSE instead of uniformly MSE.

Mean MSE for Θ: $\mathsf{E}\left[\mathsf{E}[(T(X)-\Theta)^2|\Theta=\theta]\right]$

Conditional Mean Estimator:

$T_{\mathsf{CM}} : x \mapsto \mathsf{E}[\Theta|\,X = x] = \int_\Theta \theta\cdot f_{\Theta|X}(\theta|x)\,\mathrm{d}\theta$

Posterior $f_{\Theta|\underline{X}}(\theta|\underline{\boldsymbol{x}}) = \frac{f_{\underline{X}|\Theta}(\underline{\boldsymbol{x}})f_\theta(\theta)}{\int_\Theta f_{\underline{X},\xi}(\underline{\boldsymbol{x}},\xi)\,\mathrm{d}\xi} = \frac{f_{\underline{X}|\theta}(\underline{\boldsymbol{x}})f_\theta(\theta)}{f_X(x)}$

Hint: to calculate $f_{\Theta|\underline{X}}(\theta|\underline{\boldsymbol{x}})$: Replace every factor not containing θ, such as $\frac{1}{f_X(x)}$ with a factor γ and determine γ at the end such that $\int_\Theta f_{\Theta|\underline{X}}(\theta|\underline{\boldsymbol{x}})\,\mathrm{d}\theta = 1$

MMSE: $\mathsf{E}[\mathsf{Var}[X\,|\Theta=\theta]]$

Multivariate Gaussian: $X,\Theta \sim \mathcal{N} \quad \Rightarrow \sigma_X^2 = \sigma_{X\,|\Theta=\theta}^2 + \sigma_\Theta$

$T_{\mathsf{CM}} : x \mapsto \mathsf{E}[\Theta|\,X = x] = \underline{\boldsymbol{\mu}}_\Theta + \underset{\sim}{\boldsymbol{C}}_{\Theta,X}\underset{\sim}{\boldsymbol{C}}_X^{-1}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_X)$

MMSE:

$\mathsf{E}\left[\|T_{\mathsf{CM}} - \Theta\|_2^2\right] = \mathrm{tr}(\underset{\sim}{\boldsymbol{C}}_{\theta|\,X}) = \mathrm{tr}(\underset{\sim}{\boldsymbol{C}}_\Theta - \underset{\sim}{\boldsymbol{C}}_{\Theta,X}\underset{\sim}{\boldsymbol{C}}_X^{-1}\underset{\sim}{\boldsymbol{C}}_{X,\Theta})$

Orthogonality Principle:

$T_{\mathsf{CM}}(\underline{X}) - \Theta \perp h(\underline{X}) \quad \Rightarrow \quad \mathsf{E}[(T_{\mathsf{CM}}(\underline{X}) - \Theta)h(\underline{X})] = 0$

MMSE Estimator: $\hat\theta_{\mathsf{MMSE}} = \underset{\theta\in\Theta}{\arg\min}\ \mathrm{MSE}$

minimizes the MSE for all estimators

10.7. Example:

Estimate mean θ of X with prior knowledge $\theta \in \Theta \sim \mathcal{N}$:

$X \sim \mathcal{N}(\theta, \sigma_{X\,|\Theta=\theta}^2)$ and $\Theta \sim \mathcal{N}(m, \sigma_\Theta^2)$

$$\hat\theta_{\mathsf{CM}} = \mathsf{E}[\Theta|\underline{X} = \underline{\boldsymbol{x}}] = \frac{N\sigma_\Theta^2}{\sigma_{X\,|\Theta=\theta}^2 + N\sigma_\Theta^2}\hat\theta_{\mathsf{ML}} + \frac{\sigma_{X\,|\Theta=\theta}^2}{\sigma_{X\,|\Theta=\theta}^2 + N\sigma_\Theta^2}m$$

For N independent observations x_i: $\hat\theta_{\mathsf{ML}} = \frac{1}{N}\sum x_i$

Large $N \Rightarrow$ ML better, small $N \Rightarrow$ CM better

11. Linear Estimation

t is now the unknown parameter θ, we want to estimate y and $\underline{\boldsymbol{x}}$ is the input vector... review regression problem $\underline{\boldsymbol{y}} = \underset{\sim}{\boldsymbol{A}}\underline{\boldsymbol{x}}$ (we solve for $\underline{\boldsymbol{x}}$), here we solve for $\underline{\boldsymbol{t}}$, because $\underline{\boldsymbol{x}}$ is known (measured)! Confusing...

1. Training $\rightarrow$ 2. Estimation

Training: We observe y and $\underline{\boldsymbol{x}}$ (knowing both) and then based on that we try to estimate y given $\underline{\boldsymbol{x}}$ (only observe $\underline{\boldsymbol{x}}$) with a linear model $\hat y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}}$

$$\text{Estimation: } \hat y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}} + m \quad \text{or} \quad \hat y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}}$$

Given: N observations $(y_i, \underline{\boldsymbol{x}}_i)$, unknown parameters $\underline{\boldsymbol{t}}$, noise m

$$\underline{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \underset{\sim}{\boldsymbol{X}} = \begin{bmatrix} \underline{\boldsymbol{x}}_1^\top \\ \vdots \\ \underline{\boldsymbol{x}}_n^\top \end{bmatrix} \qquad \text{Note: } \hat y \neq y!!$$

Problem: Estimate y based on given (known) observations $\underline{\boldsymbol{x}}$ and unknown parameter $\underline{\boldsymbol{t}}$ with assumed linear Model: $\hat y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}}$

Note $y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}} + m \rightarrow y = \underline{\boldsymbol{x}}'^\top\underline{\boldsymbol{t}}'$ with $\underline{\boldsymbol{x}}' = \begin{pmatrix}\underline{\boldsymbol{x}} \\ 1\end{pmatrix}$, $t' = \begin{pmatrix}\underline{\boldsymbol{t}} \\ m\end{pmatrix}$

Sometimes in Exams: $\hat y = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}} \Leftrightarrow \hat{\underline{\boldsymbol{x}}} = \underset{\sim}{\boldsymbol{T}}^\top\underline{\boldsymbol{y}}$

estimate $\underline{\boldsymbol{x}}$ given $\underline{\boldsymbol{y}}$ and unknown $\underset{\sim}{\boldsymbol{T}}$

11.1. Least Square Estimation (LSE)

Tries to minimize the square error for linear Model: $\hat y_{\mathsf{LS}} = \underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}}_{\mathsf{LS}}$

Least Square Error: $\min\left[\sum_{i=1}^N (y_i - \underline{\boldsymbol{x}}_i^\top\underline{\boldsymbol{t}})^2\right] = \min_{\underline{\boldsymbol{t}}}\left\|\underline{\boldsymbol{y}} - \underset{\sim}{\boldsymbol{X}}\underline{\boldsymbol{t}}\right\|$

$$\underline{\boldsymbol{t}}_{\mathsf{LS}} = (\underset{\sim}{\boldsymbol{X}}^\top\underset{\sim}{\boldsymbol{X}})^{-1}\underset{\sim}{\boldsymbol{X}}^\top\underline{\boldsymbol{y}}$$

$\hat{\underline{\boldsymbol{y}}}_{\mathsf{LS}} = \underset{\sim}{\boldsymbol{X}}\underline{\boldsymbol{t}}_{\mathsf{LS}} \in span(X)$

Orthogonality Principle: N observations $\underline{\boldsymbol{x}}_i \in \mathbb{R}^d$

$\underset{\sim}{\boldsymbol{Y}} - \underset{\sim}{\boldsymbol{X}}\underset{\sim}{\boldsymbol{T}}_{\mathsf{LS}} \perp \mathrm{span}[\underset{\sim}{\boldsymbol{X}}] \Leftrightarrow \underset{\sim}{\boldsymbol{Y}} - \underset{\sim}{\boldsymbol{X}}\underset{\sim}{\boldsymbol{T}}_{\mathsf{LS}} \in \mathrm{null}[\underset{\sim}{X}^\top]$, thus

$\underset{\sim}{\boldsymbol{X}}^\top(\underset{\sim}{\boldsymbol{Y}} - \underset{\sim}{\boldsymbol{X}}\underset{\sim}{\boldsymbol{T}}_{\mathsf{LS}}) = 0$ and if $N > d \wedge \mathrm{rang}[\underset{\sim}{\boldsymbol{X}}] = d$:

$\underset{\sim}{\boldsymbol{T}}_{\mathsf{LS}} = (\underset{\sim}{\boldsymbol{X}}^\top\underset{\sim}{\boldsymbol{X}})^{-1}\underset{\sim}{\boldsymbol{X}}^\top\underset{\sim}{\boldsymbol{Y}}$

11.2. Linear Minimum Mean Square Estimator (LMMSE)

Estimate y with linear estimator $\underline{\boldsymbol{t}}$, such that $\hat y = \underline{\boldsymbol{t}}^\top\underline{\boldsymbol{x}} + m$

Note: the Model does not need to be linear! The estimator is linear!

$$\hat y_{\mathsf{LMMSE}} = \underset{t,m}{\arg\min}\,\mathsf{E}\left[\left\|\underline{\boldsymbol{y}} - (\underline{\boldsymbol{t}}^\top\underline{\boldsymbol{x}} + m)\right\|_2^2\right]$$

If Random joint variable $\underline{\boldsymbol{z}} = \begin{pmatrix}\underline{\boldsymbol{x}} \\ y\end{pmatrix}$ with

$\underline{\boldsymbol{\mu}}_{\underline{\boldsymbol{z}}} = \begin{pmatrix}\underline{\boldsymbol{\mu}}_{\underline{\boldsymbol{x}}} \\ \mu_y\end{pmatrix}$ and $\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{z}}} = \begin{bmatrix}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{x}}} & \underline{\boldsymbol{c}}_{\underline{\boldsymbol{x}}y} \\ c_{y\underline{\boldsymbol{x}}} & c_y\end{bmatrix}$ then

LMMSE Estimation of y given x is

$$\hat y = \mu_y + \underline{\boldsymbol{c}}_{y\underline{\boldsymbol{x}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{x}}}^{-1}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_{\underline{\boldsymbol{x}}}) = \underbrace{\underline{\boldsymbol{c}}_{y\underline{\boldsymbol{x}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{x}}}^{-1}}_{=\underline{\boldsymbol{t}}^\top}\underline{\boldsymbol{x}} - \underbrace{\mu_y + \underline{\boldsymbol{c}}_{y\underline{\boldsymbol{x}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{x}}}^{-1}\underline{\boldsymbol{\mu}}_{\underline{\boldsymbol{x}}}}_{=m}$$

Minimum MSE: $\mathsf{E}\left[\left\|\underline{\boldsymbol{y}} - (\underline{\boldsymbol{x}}^\top\underline{\boldsymbol{t}} + m)\right\|_2^2\right] = c_y - c_{y\underline{\boldsymbol{x}}}C_{\underline{\boldsymbol{x}}}^{-1}\underline{\boldsymbol{c}}_{\underline{\boldsymbol{x}}y}$

Hint: First calculate $\hat y$ in general and then set variables according to system equation.

Multivariate: $\hat{\underline{\boldsymbol{y}}} = \underset{\sim}{\boldsymbol{T}}_{\mathsf{LMMSE}}^\top\underline{\boldsymbol{x}} \qquad \underset{\sim}{\boldsymbol{T}}_{\mathsf{LMMSE}}^\top = \underset{\sim}{\boldsymbol{C}}_{\underline{\mathbf{y}}\underline{\mathbf{x}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\mathbf{x}}}^{-1}$

If $\underline{\boldsymbol{\mu}}_{\underline{\boldsymbol{z}}} = \underline{\boldsymbol{0}}$ then

Estimator $\hat y = \underline{\boldsymbol{c}}_{y,\underline{\boldsymbol{x}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{x}}}^{-1}\underline{\boldsymbol{x}}$

Minimum MSE: $\mathsf{E}[c_{y,\underline{\boldsymbol{x}}}] = c_y - \underline{\boldsymbol{t}}^\top\underline{\boldsymbol{c}}_{\underline{\boldsymbol{x}},y}$

11.3. Matched Filter Estimator (MF)

For channel $\underline{\boldsymbol{y}} = \underline{\boldsymbol{h}}x + \underline{\boldsymbol{v}}$, Filtered: $\underline{\boldsymbol{t}}^\top\underline{\boldsymbol{y}} = \underline{\boldsymbol{t}}^\top\underline{\boldsymbol{h}}x + \underline{\boldsymbol{t}}^\top\underline{\boldsymbol{v}}$

Find Filter $\underline{\boldsymbol{t}}^\top$ that maximizes SNR $= \frac{\|\underline{\boldsymbol{h}}x\|}{\|\underline{\boldsymbol{v}}\|}$

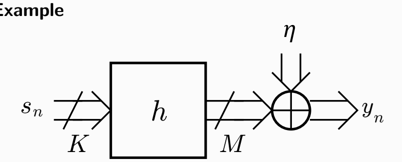
$$\underset{\sim}{\boldsymbol{t}}_{\mathsf{MF}} = \max_t\left\{\frac{\mathsf{E}\left[(\underline{\boldsymbol{t}}^\top\underline{\boldsymbol{h}}x)^2\right]}{\mathsf{E}\left[(\underline{\boldsymbol{t}}^\top\underline{\boldsymbol{v}})^2\right]}\right\}$$

In the lecture (estimate $\underline{\boldsymbol{h}}$):

$$\underset{\sim}{\boldsymbol{T}}_{\mathsf{MF}} = \max_T\left\{\frac{\left|\mathsf{E}\left[\hat{\underline{\boldsymbol{h}}}^H\underline{\boldsymbol{h}}\right]\right|^2}{\mathrm{tr}\left[\mathsf{Var}[\underset{\sim}{\boldsymbol{T}}\underline{\boldsymbol{n}}]\right]}\right\}$$

$\hat{\underline{\boldsymbol{h}}}_{\mathsf{MF}} = \underset{\sim}{\boldsymbol{T}}_{\mathsf{MF}}\underline{\boldsymbol{y}} \qquad \underset{\sim}{\boldsymbol{T}}_{\mathsf{MF}} \propto \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}}\underset{\sim}{\boldsymbol{S}}^H\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}}^{-1}$

11.4. Example



System Model: $\underline{\boldsymbol{y}}_n = \underset{\sim}{\boldsymbol{H}}\underline{\boldsymbol{s}}_n + \eta_n$

with $\underset{\sim}{\boldsymbol{H}} = (h_{m,k}) \in \mathbb{C}^{M\times K} \qquad (m \in [1,M], k \in [1,K])$

Linear Channel Model $\underline{\boldsymbol{y}} = \underset{\sim}{\boldsymbol{S}}\underline{\boldsymbol{h}} + \underline{\boldsymbol{n}}$ with

$\underline{\boldsymbol{h}} \sim \mathcal{N}(0, \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}})$ and $\underline{\boldsymbol{n}} \sim \mathcal{N}(0, \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}})$

Linear Estimator $\underset{\sim}{\boldsymbol{T}}$ estimates $\hat{\underline{\boldsymbol{h}}} = \underset{\sim}{\boldsymbol{T}}\underline{\boldsymbol{y}} \in \mathbb{C}^{MK}$

$\underset{\sim}{\boldsymbol{T}}_{\mathsf{MMSE}} = \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}\underline{\boldsymbol{y}}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{y}}}^{-1} = \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}}\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}}(\underset{\sim}{\boldsymbol{S}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}}\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}} + \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}})^{-1}$

$\underset{\sim}{\boldsymbol{T}}_{\mathsf{ML}} = \underset{\sim}{\boldsymbol{T}}_{\mathsf{Cor}} = (\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}}^{-1}\underset{\sim}{\boldsymbol{S}})^{-1}\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}}^{-1}$

$\underset{\sim}{\boldsymbol{T}}_{\mathsf{MF}} \propto \underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{h}}}\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}}\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}}^{-1}$

For Assumption $\underset{\sim}{\boldsymbol{S}}^{\mathsf{H}}\underset{\sim}{\boldsymbol{S}} = N\sigma_s^2\underset{\sim}{\boldsymbol{1}}_{K\times M}$ and $\underset{\sim}{\boldsymbol{C}}_{\underline{\boldsymbol{n}}} = \sigma_\eta^2\underset{\sim}{\boldsymbol{1}}_{N\times M}$

Estimator	Averaged Squared Bias	Variance
ML/Correlator	0	$KM\,\frac{\sigma_\eta^2}{N\sigma_s^2}$
Matched Filter	$\sum_{i=1}^{KM}\lambda_i\left(\frac{\lambda_i}{\lambda_1} - 1\right)^2$	$\sum_{i=1}^{KM}\left(\frac{\lambda_i}{\lambda_1}\right)^2\frac{\sigma_\eta^2}{N\sigma_s^2}$
MMSE	$\sum_{i=1}^{KM}\lambda_i\left(\frac{1}{1+\frac{\sigma_\eta^2}{\lambda_i N\sigma_s^2}} - 1\right)^2$	$\sum_{i=1}^{KM}\frac{1}{\left(1+\frac{\sigma_\eta^2}{\lambda_i N\sigma_s^2}\right)^2}\frac{\sigma_\eta^2}{N\sigma_s^2}$

11.5. Estimators

Upper Bound: Uniform in $[0;\theta]$: $\hat\theta_{\mathsf{ML}} = \frac{2}{N}\sum x_i$

Probability p for $\mathcal{B}(p,N)$: $\hat p_{\mathsf{ML}} = \frac{x}{N} \qquad \hat p_{\mathsf{CM}} = \frac{x+1}{N+2}$

Mean μ for $\mathcal{N}(\mu,\sigma^2)$: $\hat\mu_{\mathsf{ML}}^2 = \frac{1}{N}\sum_{i=1}^N x_i$

Variance σ^2 for $\mathcal{N}(\mu,\sigma^2)$: $\hat\sigma_{\mathsf{ML}}^2 = \frac{1}{N}\sum_{i=1}^N (x_i - \mu)^2$

12. Gaussian Stuff

12.1. Gaussian Channel

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

$L(y_1, \dots, y_N) = \prod_{i=1}^{n} f_{Y_i}(y_i, h)$

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

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

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

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

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

12.2. Multivariate Gaussian Distributions

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

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

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

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

All marginal PDFs are Gaussian as well

Contour Lines

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

12.3. Conditional Gaussian

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

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

Conditional Mean:

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

Conditional Variance:

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

12.4. 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 $\Phi(x - \mu_x)$

13. Sequences

13.1. Random Sequences

Sequence of a random variable. Example: result of a dice is RV, roll a dice several times is a random sequence.

13.2. Markov Sequence $X_n : \Omega \to X_n$

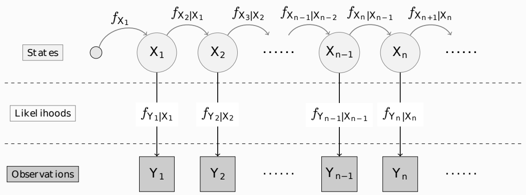
Sequence of memoryless state transitions with certain probabilities.

1. state: $f_{X_1}(x_1)$
2. state: $f_{X_2\,|\,X_1}(x_2|x_1)$

n. state: $f_{X_n\,|\,X_{n-1}}(x_n|x_{n-1})$

13.3. Hidden Markov Chains

Problem: states X_i are not visible and can only be guessed indirectly as a random variable Y_i.



Conditional pdf $f_{\underline{X}_n|\underline{Y}_n}$ Likelihood pdf $f_{Y_n\,|\,X_n}$

State-transision pdf $f_{X_n\,|\,X_{n-1}}$

Estimation:

$$f_{\underline{X}_n|\underline{Y}_n} \propto f_{\underline{Y}_n|\underline{X}_n} \cdot \int_{\mathbb{X}} f_{\underline{X}_n|\underline{X}_{n-1}} \cdot f_{\underline{X}_{n-1}|\underline{Y}_{n-1}}\,\mathrm{d}\underline{x}_{n-1}$$

14. Recursive Estimation

14.1. Kalman-Filter

recursively calculates the most likely state from previous state estimates and current observation. Shows optimum performance for Gauss-Markov Sequences.

State space:

$\underline{x}_n = \underline{G}_n\underline{x}_{n-1} + \underline{B}\underline{u}_n + \underline{v}_n$

$\underline{y}_n = \underline{H}_n\underline{x}_n + \underline{w}_n$

With gaussian process/measurement noise $\underline{v}_n/\underline{w}_n$

Short notation: $\mathrm{E}[\underline{x}_n|\underline{y}_{n-1}] = \hat{\underline{x}}_{n|n-1}$ $\quad\mathrm{E}[\underline{x}_n|\underline{y}_n] = \hat{\underline{x}}_{n|n}$

$\mathrm{E}[\underline{y}_n|\underline{y}_{n-1}] = \hat{\underline{y}}_{n|n-1}$ $\quad\mathrm{E}[\underline{y}_n|\underline{y}_n] = \hat{\underline{y}}_{n|n}$

1. step: Prediction

Mean: $\hat{\underline{x}}_{n|n-1} = \underline{G}_n\hat{\underline{x}}_{n-1|n-1}$

Covariance: $\underline{C}_{\underline{x}_{n|n-1}} = \underline{G}_n\underline{C}_{\underline{x}_{n-1|n-1}}\underline{G}_n^\top + \underline{C}_{\underline{v}}$

2. step: Update

Mean: $\hat{\underline{x}}_{n|n} = \hat{\underline{x}}_{n|n-1} + \underline{K}_n\left(\underline{y}_n - \underline{H}_n\hat{\underline{x}}_{n|n-1}\right)$

Covariance: $\underline{C}_{\underline{x}_{n|n}} = \underline{C}_{\underline{x}_{n|n-1}} + \underline{K}_n\underline{H}_n\underline{C}_{\underline{x}_{n|n-1}}$

$$\hat{\underline{x}}_{n|n} = \underbrace{\hat{\underline{x}}_{n|n-1}}_{\text{estimation } \mathrm{E}[X_n\,|\,Y_{n-1}=y_{n-1}]} + \overbrace{\underline{K}_n\underbrace{\left(\underline{y}_n - \underline{H}_n\hat{\underline{x}}_{n|n-1}\right)}_{\text{innovation: } \Delta y_n}}^{\text{correction: } \mathrm{E}[X_n\,|\,\Delta Y_n = y_n]}$$

With optimal **Kalman-gain** (prediction for $\underline{x}_n$ based on Δy_n):

$$\underline{K}_n = \underline{C}_{\underline{x}_{n|n-1}}\underline{H}_n^\top\underbrace{(\underline{H}_n\underline{C}_{\underline{x}_{n|n-1}}\underline{H}_n^\top + \underline{C}_{\underline{w}_n})}_{\underline{C}_{\delta y_n}}{}^{-1}$$

Innovation: closeness of the estimated mean value to the real value

$\Delta\underline{y}_n = \underline{y}_n - \hat{\underline{y}}_{n|n-1} = \underline{y}_n - \underline{H}_n\hat{\underline{x}}_{n|n-1}$

Init: $\hat{\underline{x}}_{0|-1} = \mathrm{E}[X_0]$ $\quad\sigma^2_{0|-1} = \mathrm{Var}[X_0]$

MMSE Estimator: $\hat{\underline{x}} = \int \underline{x}_n f_{X_n\,|\,Y_{(n)}}(\underline{x}_n|\underline{y}_{(n)})\,\mathrm{d}\underline{x}_n$

For non linear problems: Suboptimum nonlinear Filters: Extended KF, Unscented KF, ParticleFilter

14.2. Extended Kalman (EKF)

Linear approximation of non-linear g, h

$\underline{x}_n = g_n(\underline{x}_{n-1}, \underline{v}_n)$ $\quad\underline{v}_n \sim \mathcal{N}$

$\underline{y}_n = h_n(\underline{x}_{n-1}, \underline{w}_n)$ $\quad\underline{w}_n \sim \mathcal{N}$

14.3. Unscented Kalman (UKF)

Approximation of desired PDF $f_{X_n|Y_n}(x_n|y_n)$ by Gaussian PDF.

14.4. Particle-Filter

For non linear state space *and* non-gaussian noise

Non-linear State space:

$\underline{x}_n = g_n(\underline{x}_{n-1}, \underline{v}_n)$

$\underline{y}_n = h_n(\underline{x}_n, \underline{w}_n)$

Posterior Conditional PDF: $f_{X_n|Y_n}(x_n|y_n) \propto \overbrace{f_{Y_n|X_n}(y_n|x_n)}^{\text{likelihood}} \cdot$

$$\cdot \int_{\mathbb{X}} \underbrace{f_{X_n|X_{n-1}}(x_n|x_{n-1})}_{\text{state transition}}\,\underbrace{f_{X_{n-1}|Y_{n-1}}(x_{n-1}|y_{n-1})}_{\text{last conditional PDF}}\,\mathrm{d}x_{n-1}$$

N random Particles with particle weight w_n^i at time n

Monte-Carlo-Integration: $I = \mathrm{E}[g(X)] \approx I_N = \frac{1}{N}\sum_{i=1}^{N}\tilde{g}(x^i)$

Importance Sampling: Instead of $f_X(x)$ use **Importance Density** $q_X(x)$

$I_N = \frac{1}{N}\sum_{i=1}^{N}\tilde{w}^i g(x^i)$ with weights $\tilde{w}^i = \frac{f_X(x^i)}{q_X(x^i)}$

If $\int f_{X_n}(x)\,\mathrm{d}x \neq 1$ then $I_N = \sum_{i=1}^{N}\tilde{w}^i g(x^i)$

14.5. Conditional Stochastical Independence

$\mathrm{P}(A \cap B|E) = \mathrm{P}(A|E) \cdot \mathrm{P}(B|E)$

Given Y, X and Z are independent if

$f_{Z\,|\,Y,X}(z|y, x) = f_{Z\,|\,Y}(z|y)$ or

$f_{X,Z\,|\,Y}(x, z|y) = f_{Z\,|\,Y}(z|y) \cdot f_{X\,|\,Y}(x|y)$

$f_{Z\,|\,X,Y}(z|x, y) = f_{Z\,|\,Y}(z|y)$ or $f_{X\,|\,Z,Y}(x|z, y) = f_{X\,|\,Y}(x|y)$

15. Hypothesis Testing

making a decision based on the observations

15.1. 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 $\mathrm{E}[d(X)|\theta] \leq \alpha, \forall\theta \in \Theta_0$

Error Type	Decision \ Reality	H_1 false (H_0 true)	H_1 true (H_0 false)
1 (FA) False Alarm	H_1 rejected (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: $\frac{\mathrm{TP+TN}}{\mathrm{P+N}} = \frac{2-\alpha-\beta}{2}$

15.1.1. Design of a test

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

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

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

15.2. 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:

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

16. Tests

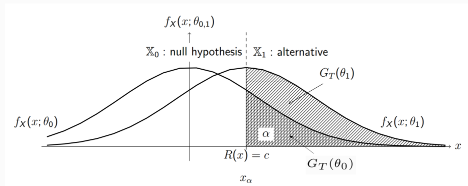
16.1. Neyman-Pearson-Test

The best test of P_0 against P_1 is

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

$\gamma = \frac{\alpha - \mathrm{P}_0(\{R>c\})}{\mathrm{P}_0(\{R=c\})}$ $\quad$ Errorlevel α

Steps: For α calculate x_α, then $c = R(x_\alpha)$



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

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

16.2. Bayes Test (MAP Detector)

Prior knowledge on possible hypotheses: $\mathrm{P}(\{\theta \in \Theta_0\}) + \mathrm{P}(\{\theta \in \Theta_1\}) = 1$, minimizes the probability of a wrong decision.

$d_{\text{Bayes}} = \begin{cases} 1 & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > \frac{c_0\,\mathrm{P}(\theta_0|x)}{c_1\,\mathrm{P}(\theta_1|x)} \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \mathrm{P}(\theta_1|x) > \mathrm{P}(\theta_0|x) \\ 0 & \text{otherwise} \end{cases}$

Risk weights c_0, c_1 are 1 by default.

If $\mathrm{P}(\theta_0) = \mathrm{P}(\theta_1)$, the Bayes test is equivalent to the ML test

Loss Function $L(d(x), \theta) = \begin{cases} c_0 & \text{type 1 } d(x) = 1, \text{ but } \theta = \theta_0 \\ c_1 & \text{type 2 } d(x) = 0, \text{ but } \theta = \theta_1 \end{cases}$

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

Multiple Hypothesis $d_{\text{Bayes}} = \begin{cases} 0 & x \in \mathbb{X}_0 \\ 1 & x \in \mathbb{X}_1 \\ 2 & x \in \mathbb{X}_2 \end{cases}$

16.3. Linear Alternative Tests

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

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

$\log R(\underline{\boldsymbol{x}}) = \frac{\ln(\det(\underset{\sim}{\boldsymbol{C}}_0))}{\ln(\det(\underset{\sim}{\boldsymbol{C}}_1))} + \frac{1}{2}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_0)^\top \underset{\sim}{\boldsymbol{C}}_0^{-1}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_0) -$
$-\frac{1}{2}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_1)^\top \underset{\sim}{\boldsymbol{C}}_1^{-1}(\underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}_1) = 0$

For 2 Gaussians, with $\underset{\sim}{\boldsymbol{C}}_0 = \underset{\sim}{\boldsymbol{C}}_1 = \underset{\sim}{\boldsymbol{C}}$: $\underline{\boldsymbol{w}}^\top = (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^\top \underset{\sim}{\boldsymbol{C}}$

and constant translation $w_0 = \frac{(\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)^\top \underset{\sim}{\boldsymbol{C}} (\underline{\boldsymbol{\mu}}_1 - \underline{\boldsymbol{\mu}}_0)}{2}$