Machine Learning
Performance Measures and Statistical Tests

Prof. Matthias Hein

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Lecture 15, 13.12.2013

Program of today and next time

How to do model evaluation and model selection (for classification)

- Different performance measures
  (confusion matrix, ROC-curve and AUC),
- Statistical tests and confidence intervals,
- Test error as an estimator of the true error
  - confidence intervals
  - sample complexity
- Comparison of two/multiple classifiers - Which one is better ?
  Application of statistical tests.
- Model selection
  - using validation sets,
  - using cross-validation.

Performance Measures

Up to now

- 0-1-loss (count of errors),
- convex surrogates of the 0-1-loss $\Rightarrow$ hinge loss, logistic loss, etc.

But

- error count is not always meaningful (unbalanced classes),
- even if we use a convex surrogate for training, evaluation is done using the desired 0-1-loss.

WARNING !

IMPORTANT:	The performance of a learning method has to be evaluated using **independent data** which has **not** been used in the construction of the learning method.

It makes no sense to use the training error because we have optimized it !

Performance Measures II

The basis of all performance measures - the confusion matrix

$$\text{true pos. } tp = \sum_{i=1}^{m} \mathbb{1}_{\hat{f}(X_i)=1} \mathbb{1}_{Y_i=1}, \qquad \text{false neg. } fn = \sum_{i=1}^{m} \mathbb{1}_{\hat{f}(X_i)=-1} \mathbb{1}_{Y_i=1},$$

$$\text{false pos. } fp = \sum_{i=1}^{m} \mathbb{1}_{\hat{f}(X_i)=1} \mathbb{1}_{Y_i=-1}, \qquad \text{true neg. } tn = \sum_{i=1}^{m} \mathbb{1}_{\hat{f}(X_i)=-1} \mathbb{1}_{Y_i=-1}.$$

	positive Prediction	negative Prediction	total cases
positive cases	true positives (tp)	false negatives (fn)	$P = tp + fn$
negative cases	false positives (fp)	true negatives (tn)	$N = fp + tn$
total pred.	$\hat{P} = tp + fp$	$\hat{N} = fn + tn$	$m = N + P = \hat{P} + \hat{N}$

The confusion matrix for a binary classification problem.

Performance Measures III

The basis of all performance measures - the confusion matrix

- all other performance measures can be derived from the confusion matrix
- it makes more sense to report the confusion matrix since then everybody can build its favorite performance measure on its own
- includes all information about class probabilities $\Rightarrow$ important for cost-sensitive learning

Performance Measures IV

The standard performance measures:

Accuracy:	accuracy = (tp+tn)/(tp+fp+fn+tn)= (P*tpr + N*tnr)/(P+N), percentage of examples which are **correctly classified,** the accuracy is an estimator of $\mathrm{P}(f(X) = Y)$
Error:	error = (fp+fn)/(tp+fp+fn+tn) =(fp+fn)/(N+P), percentage of examples which are **wrongly classified,** the error is an estimator of $\mathrm{P}(f(X) \neq Y)$

True/False positive/negative rates

true positive rate: (or **sensitivity**)	tpr $=$ tp/(tp+fn)=tp/P, ratio of **pos. ex. correctly classified as pos.** the tpr is an estimator of $\mathrm{P}(f(X) = 1 \mid Y = 1)$.
false positive rate:	fpr $=$ fp/(fp+tn)=fp/N, ratio of **neg. ex. wrongly classified as pos.** the fpr is an estimator of $\mathrm{P}(f(X) = 1 \mid Y = -1)$.
false negative rate:	fnr $=$ fn/(tp+fn)=fn/P, ratio of **pos. ex. wrongly classified as neg.** the fnr is an estimator of $\mathrm{P}(f(X) = -1 \mid Y = 1)$.
true negative rate: (or **specificity**)	tnr $=$ tn/(fp+tn)=tn/N, ratio of **neg. ex. correctly classified as neg.** the tnr is an estimator of $\mathrm{P}(f(X) = -1 \mid Y = -1)$.

Performance Measures V

Useful measures in case of highly unbalanced classes:

positive predictive value	ppv $=$ tp$/$(tp+fp), ratio of **true pos. ex. classified as pos.,** ppv is an estimator of $\mathrm{P}(Y = 1 \mid f(X) = 1)$.
negative predictive value	npv $=$ tn$/$(tn+fn) ratio of **true neg. ex. classified as neg.,** npv is an estimator of $\mathrm{P}(Y = -1 \mid f(X) = -1)$.

- Measure for how often is the classifier correct when it predicts a certain class.
- In medical applications (diagnosis of diseases), in particular where the class probabilities are highly unbalanced, one uses alternatively to sensitivity and specificity, the **negative and positive predictive value**.

Performance Measures VI

Measures used in information retrieval (one class problem):

recall:	recall = tp/(tp+fn)=tp/P, (same as true positive rate) ratio of **true pos. ex. correctly classified as pos.** recall is an estimator of $\mathrm{P}(f(X) = 1 \mid Y = 1)$.
precision:	precision = tp/(tp+fp), (same as positive predictive value) ratio of **true pos. examples of all ex. classified as pos.** precision is an estimator of $\mathrm{P}(Y = 1 \mid f(X) = 1)$.

- recall: ratio of the detected relevant documents
- precision: ratio of relevant documents in the result list,

Definition

The **F-score** F is the harmonic mean of precision and recall,

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}} = \frac{2p\,r}{p + r}.$$

Performance Measures VII

Performance measures:

- different weighting of false positive and false negatives,
- differences most prominent when data is unbalanced,
- solution: assign cost to false positives and false negatives and then optimize this cost.

$\Rightarrow$ In a ROC curve one integrates all possible class probabilities in one plot.

ROC-curve and AUC

Motivation:

- find best threshold for a desired true positive/false positive rate,
- corresponds to a certain ratio of class probabilities,
- ROC= Receiver Operating Characteristic,
- developed in the 2nd world war to access the optimal threshold for radar detection systems,

ROC curve:

- change decision threshold for classifier $f : \mathcal{X} \to \mathbb{R}$,

$$\hat{f}_\theta(x) = \operatorname{sign}(f(x) + \theta),$$

- ROC-curve: plot the true positive rate versus the false positive rate by varying the discrimination threshold θ.
- naive way: compute true positive/false positive rate for all thresholds.

Formal definition:

Definition

Let $tpr(\theta)$ and $fpr(\theta)$ the true and false positive error rate of $\hat{f}_\theta$, then the ROC-function $ROC : [0,1] \to [0,1]$ is defined as

$$\mathrm{ROC}(x) = tpr\Big(fpr^{-1}(x)\Big).$$

Properties:

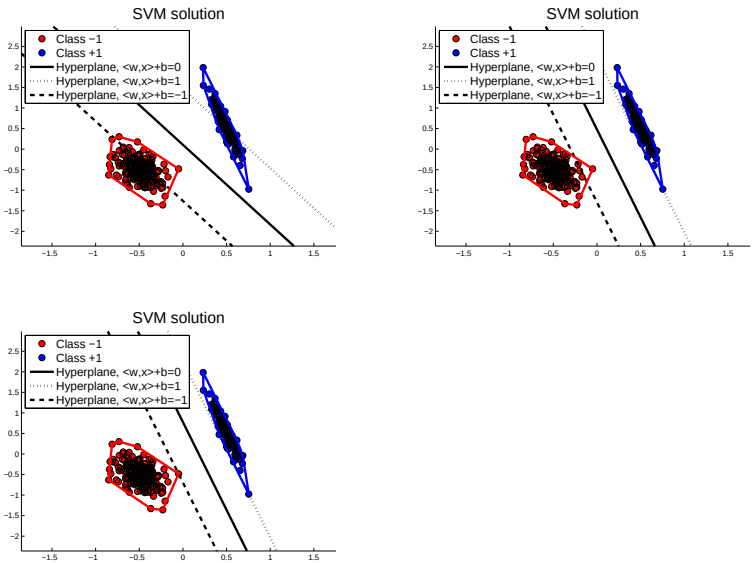
- $\mathrm{ROC}(0) = 0$ and $\mathrm{ROC}(1) = 1$,
- the ROC function is monotonically increasing,
- the ROC function can be multi-valued.

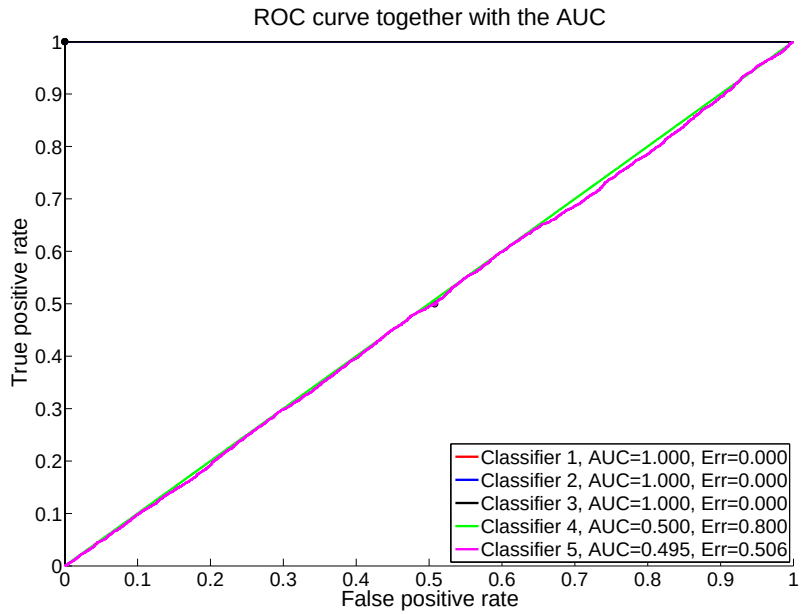
Practical computation:

1. For m test instances there are only $m+1$ possible thresholds.

- sort Y_{test} in **decreasing** order of $f(X_{\text{test}})$, and then start with $(0,0)$,
- for i=1:m
  - if $Y_{test}(i) = 1$, move up by $1/P$,
  - if $Y_{\text{test}}(i) = -1$, move right by $1/N$,
  - ties in $f(X_{\text{test}})$ require special handling.
    - ★ move up by n_+/P, where n_+ is the number of positive test points which have the same output $f(X_{\text{test}})$,
    - ★ move right by n_-/N, where n_- is the number of negative test points which have the same output $f(X_{\text{test}})$.

Ideal ROC-curve:





Left: Three linear soft margin SVM's with $C = 1, 10, 10000$, **Right:** The ROC-curve of the three SVM's (red, blue, black) together with the baseline of the (interpolated) constant classifier (green) and a random classifier (magenta). We have perfect separation of the classes - ideal ROC curve. What is the ROC curve of the constant classifier ?

Linear Interpolation of points on the ROC-curve:

- two different thresholds θ_1, θ_2,
- tpr_1, tpr_2 and fpr_1, fpr_2 are the true and false positive error rates associated to the thresholds θ_1, θ_2,
- Then the linearly interpolated true/false positive error rates

$$tpr(\lambda) = \lambda\, tpr_1 + (1-\lambda)\, tpr_2, \qquad fpr(\lambda) = \lambda\, fpr_1 + (1-\lambda)\, fpr_2,$$

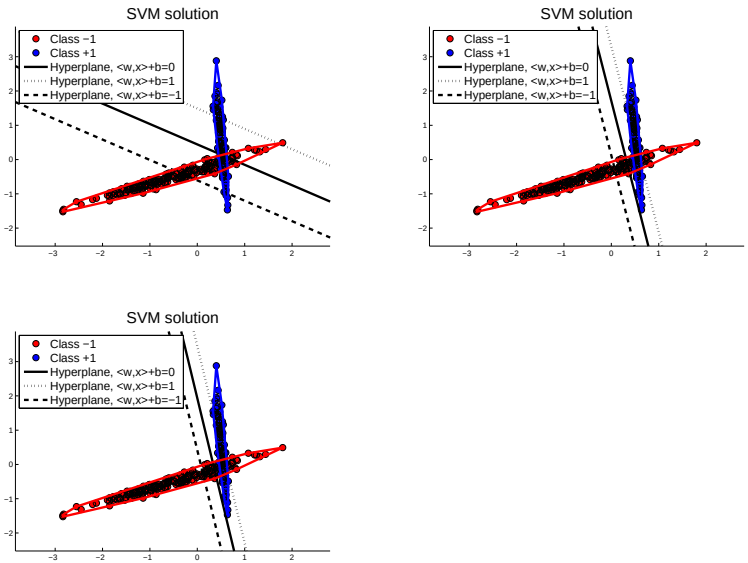
can be generated by combining $\hat{f}_{\theta_1}$ and $\hat{f}_{\theta_2}$ using a weighted coin flip,

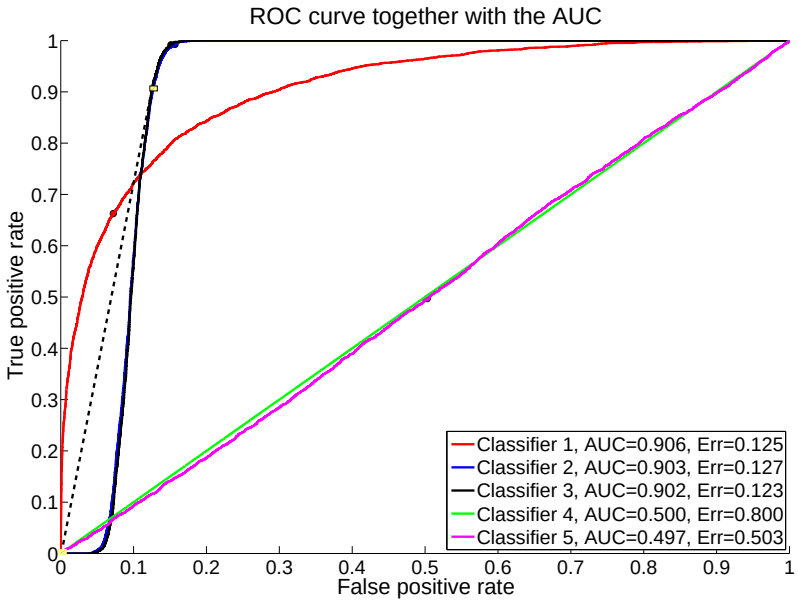
$$\forall\, \lambda \in [0,1], \qquad \hat{f}_{\lambda\theta_1+(1-\lambda)\theta_2}(x) = \begin{cases} \hat{f}_{\theta_1}(x) & \text{with probability } \lambda, \\ \hat{f}_{\theta_2}(x) & \text{with probability } 1-\lambda. \end{cases}$$

Combination of classifier

- given a set of classifiers and their ROC-curves construct the convex hull of the ROC curves by combining different classifiers.

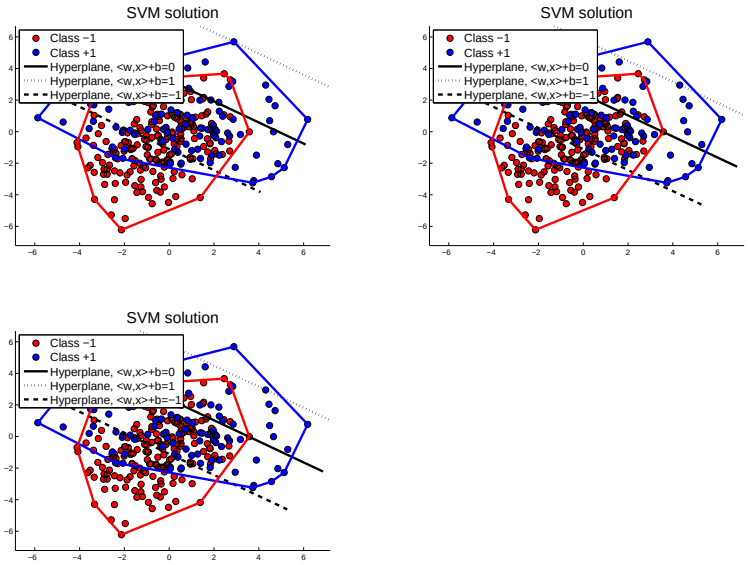
Reasonable ROC-curve:

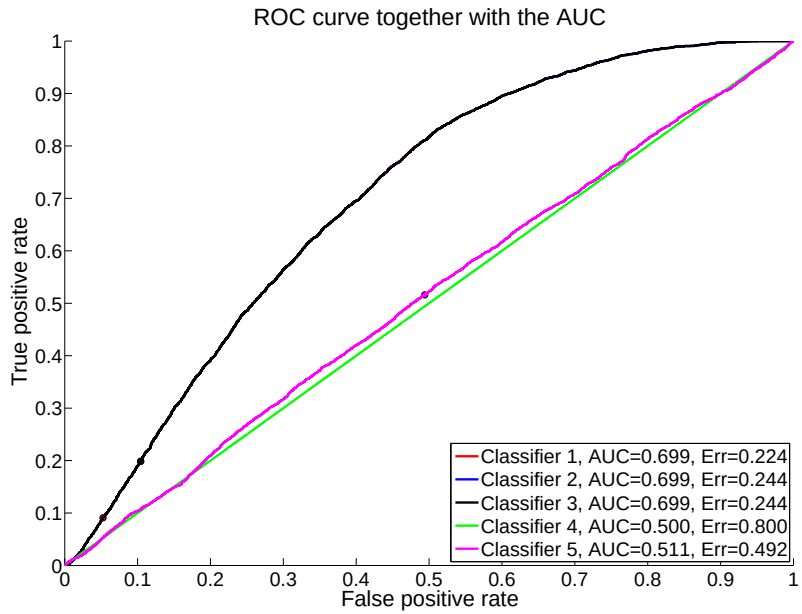




Problem is not separable anymore - ROC-curves come closer to the random baseline For small false positive error rates two of the SVM's perform worse than random. Combination of classifiers - indicated by the dotted line in the ROC plot and the yellow rectangles indicate the classifiers which are combined.

Bad ROC-curve:





A very difficult problem, where the classes are largely overlapping. The ROC-curves of the three SVM's are still above random.

ROC-curve and AUC IV

How to compare different ROC curves ?

- A ROC curve can only be said to be better than another curve if the curve always lies above the second one,
- alternative: summarize the curve with a single number,
- measure the **Area under the Curve (AUC)**,

Interpretation of the AUC:

The AUC measures basically the **quality of the ranking** of a classifier.

- all positive ex. are ranked higher than all negative ones $\Longrightarrow$ AUC$= 1$,
- all negative ex. are ranked higher than all positive ones $\Longrightarrow$ AUC$= 0$,
- If the ordering is random, then the expectation of the AUC is 0.5.

Definition

Let $f : \mathcal{X} \to \mathbb{R}$ be the original classifier and denote by $X_1^+, \ldots, X_P^+$ and $X_1^-, \ldots, X_N^-$ the set of points from positive and negative class in the test set (drawn i.i.d.). Then the **AUC** is defined as,

$$\text{AUC} = \frac{1}{NP} \sum_{j=1}^{N} \sum_{i=1}^{P} \mathbb{1}_{f(X_i^+) > f(X_j^-)}.$$

(equivalent to the 2-sample Mann-Whitney U-test or Wilcoxon rank sum test).

Proposition

Let X^+ *be distributed as* $\mathrm{P}(X|Y=1)$ *and* X^- *as* $\mathrm{P}(X|Y=-1)$*, then*

$$\mathbb{E}[\,\mathrm{AUC}] = \mathrm{P}\big(f(X^+) > f(X^-)\big).$$

Proof: Decompose test sample X_{test} into the samples $X_1^+, \ldots, X_p^+$ and $X_1^-, \ldots, X_{m-p}^-$ from positive and negative class. Note, that

$$\mathbb{E}_{\{(X_1,Y_1),(X_2,Y_2),\ldots,(X_m,Y_m)\}}[\,\mathrm{AUC}] = \mathbb{E}_p[\mathbb{E}_{\{(X_1^+,\ldots,X_p^+),(X_1^-,\ldots,X_{m-p}^-)\}}[\,\mathrm{AUC}\,|\,p]],$$

where $p \sim \mathrm{Bin}(m, \mathrm{P}(Y=1))$ and X_i^+ and X_j^- are i.i.d. samples from $\mathrm{P}(X|Y=1)$ and $\mathrm{P}(X|Y=-1)$. Thus,

$$\mathbb{E}[\,\mathrm{AUC}] = \mathbb{E}_p[\frac{1}{(m-p)p}\sum_{i=1}^{p}\sum_{j=1}^{m-p}\mathbb{E}_{X_i^+,X_j^-}\mathbb{1}_{f(X_i^+)>f(X_j^-)}\,|\,p]$$

$$= \mathbb{E}_p[\frac{1}{(m-p)p}(m-p)p\,\mathrm{P}\big(f(X^+) > f(X^-)\big)\,|\,p] = \mathrm{P}\big(f(X^+) > f(X^-)\big),$$

where we have used that the pairs X_i^+ and X_j^- are i.i.d.

Multiclass performance measures

- generalization of binary measures to the multi-class case can be difficult.
- **Class balanced error:** Let X_j^k denote the samples from class k and let N_k be the cardinality of class k in the test set, then

$$\text{error}_{\text{balanced}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{j=1}^{N_k} \mathbb{1}_{f(X_j^k) \neq k}.$$

	prediction class 1	...	prediction class K	total cases
pred. class 1	$\lvert\{f(X_i) = 1, Y_i = 1\}\rvert$	...	$\lvert\{f(X_i) = K, Y_i = 1\}\rvert$	$\lvert\{Y_i = 1\}\rvert$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
pred. class K	$\lvert\{f(X_i) = 1, Y_i = K\}\rvert$	...	$\lvert\{f(X_i) = K, Y_i = K\}\rvert$	$\lvert\{Y_i = K\}\rvert$
total pred.	$\lvert\{f(X_i) = 1\}\rvert$	...	$\lvert\{f(X_i) = K\}\rvert$	

Table : The confusion matrix for a multi-class problem.

Golden rule for your machine learning problem

Golden rule for solving problems:

Think about what is the correct performance measure/cost matrix for your problem and then optimize that and not something else !

Relation between test and true error

Test error = True error ?

Given the test error of the classifier on m test samples.

Questions:

- Can we make any assertions if the true error is close to the test error ?
- For a given confidence level and sample size can we give a confidence interval for the true error given the error on an independent test set ?
- For a given confidence interval and confidence level how many test samples do we need ?
- Can we test if the classifier is significantly better than random guessing ?

Statistical Test

What is a statistical test ?

- try to falsify a given null hypothesis H_0 (e.g. SVM and LDA perform equally on data set XYZ),
- define a region of rejection which if H_0 is true has probability (less than) α (where α is the significance level),
- compute a test statistic T (e.g. difference of the test errors of SVM and LDA),
- we reject the null hypothesis if T attains a value in the region of rejection otherwise we keep the null hypothesis,
  - If we reject the null hypothesis, we say that the difference between SVM and LDA is **statistically different,**
  - In the other case we can make no statement about the relation between SVM and LDA,

Formal definition of a statistical test

1. Let Θ be a set of values, then the null hypothesis H_0 is an assertion that $\theta \in \Theta_0 \subset \Theta$ whereas the alternative hypothesis H_1 is that $\theta \in \Theta \backslash \Theta_0$,
2. A significance level α is chosen (common values are 0.05 or 0.01),
3. A test statistic $T : \mathbb{R}^n \to \mathbb{R}^d$ and a region of rejection B_n is chosen, such that if the null hypothesis is true

$$\forall\, \theta \in \Theta_0, \quad \mathrm{P}_\theta\Big(T(X_n) \in B_n\Big) \leq \alpha,$$

   where X_n is a set of n sample points,
4. An experiment is done which gives X_n,
5. H_0 is rejected (one assumes that H_1 holds) if $T(X_n) \in B_n$.

Test can be **parametric** or **nonparametric**. If $\Theta = \mathbb{R}$, $H_0 : \theta \gtreqless \theta_0$ is a **one-sided test** and $H_0 : \theta = \theta_0$ is a **two-sided test**.

Statistical Test III

Confusion matrix of a statistical test:

decision\ reality	H_0 is correct	H_1 is correct
H_0 is not rejected	correct decision	type II error with prob. $1-\beta(\theta)$
H_0 is rejected	type I error (prob. $\leq \alpha$)	correct decision

Definition

Let P_θ be the probability measure with parameter θ, then the **power function** of a test is

$$\beta(\theta) = \mathrm{P}_\theta(T(X_n) \in B_n).$$

The rejection region B_n has been chosen such that,

$$\beta(\theta) \leq \alpha, \text{ for all } \theta \in \Theta_0.$$

The **type II error** is $1-\beta(\theta)$ for $\theta \in \Theta \backslash \Theta_0 \Longrightarrow$ **Goal:** high power

p-Value

Definition

Suppose that for every $\alpha \in (0,1)$ we have a test of size α with a corresponding rejection region $B_n(\alpha)$. Then, the **p-value** is defined as

$$\text{p-value} = \inf\{\alpha \mid T(X_n) \in B_n(\alpha)\}.$$

The p-value is thus **the smallest significance level α at which the null-hypothesis would be rejected.**

If we have

- a test statistic of the form $T : \mathbb{R}^n \to [0, \infty)$,
- and the rejection region is given as $[c(\alpha), \infty)$ for $c : (0,1) \to \mathbb{R}$.

and the computed test statistic has value t_{obs}, then

$$\text{p-value} = \mathrm{P}_{\theta_0}\Big(T(X_n) \geq t_{\text{obs}}\Big).$$

Example - Gauss test

- **Parametric test:** Gaussians $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R}$ of fixed variance.
- **Null hypothesis:** $\mu = \mu_0$ and significance level α,
- The **test statistic**, is

$$T(X) = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^{n} X_i - \mu_0}{\sigma}.$$

- Reject the null hypothesis if $|T(X)| > q_{1-\frac{\alpha}{2}}$, where q_γ is the γ-Quantile of $\mathcal{N}(0,1)$. Under the null hypothesis, $T(X) \sim \mathcal{N}(0,1)$, and thus

$$\mathrm{P}\Big(|T(X)| > q_{1-\frac{\alpha}{2}}\Big) = \alpha.$$

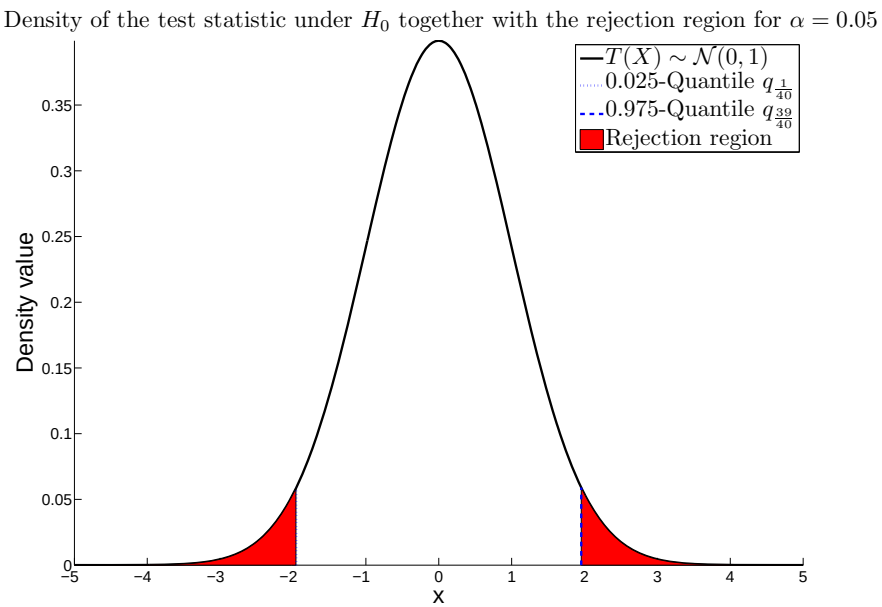
- **Power function:** $T(X) \sim \mathcal{N}(\sqrt{n}\frac{\mu-\mu_0}{\sigma}, 1)$, with

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx,$$

$$\beta(\mu) = \mathrm{P}_\mu\Big(|T(X)| > q_{1-\frac{\alpha}{2}}\Big)$$

$$= 1 - \Phi\Big(q_{1-\frac{\alpha}{2}} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\Big) + \Phi\Big(-q_{1-\frac{\alpha}{2}} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\Big).$$

Example - Gauss test II



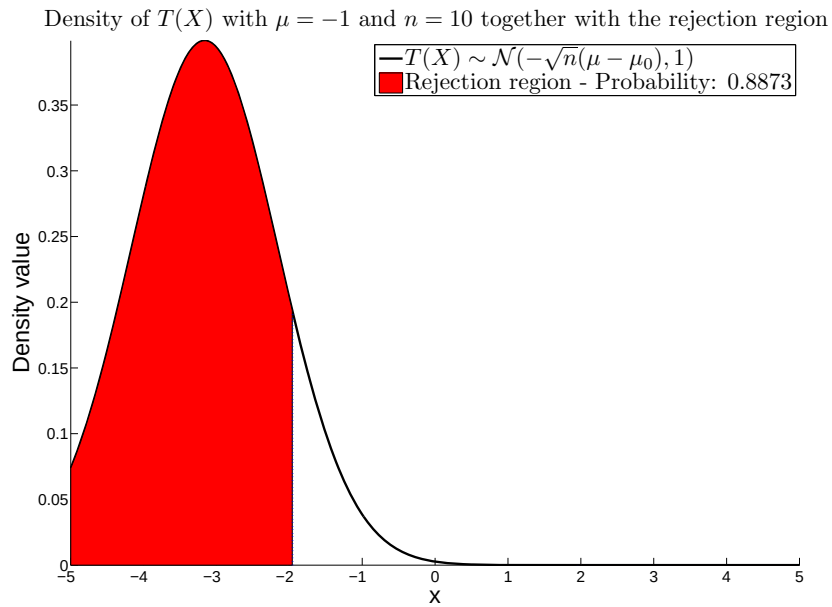


Figure : Left: The distribution of the test-statistic under the null hypothesis together with the rejection region for the significance level $\alpha = 0.05$. Right: The computation of the power of the test for $\mu = -1$ and $n = 10$.

Example - Gauss Test III

Numerical example:

- 10 samples from Gaussians with $\sigma = 2$.
- Test $H_0 : \mu = 0$ with $\alpha = 0.05 \Longrightarrow$ acceptance region: $[q_{0.025}, q_{0.975}] = [-1.96, 1.96]$.

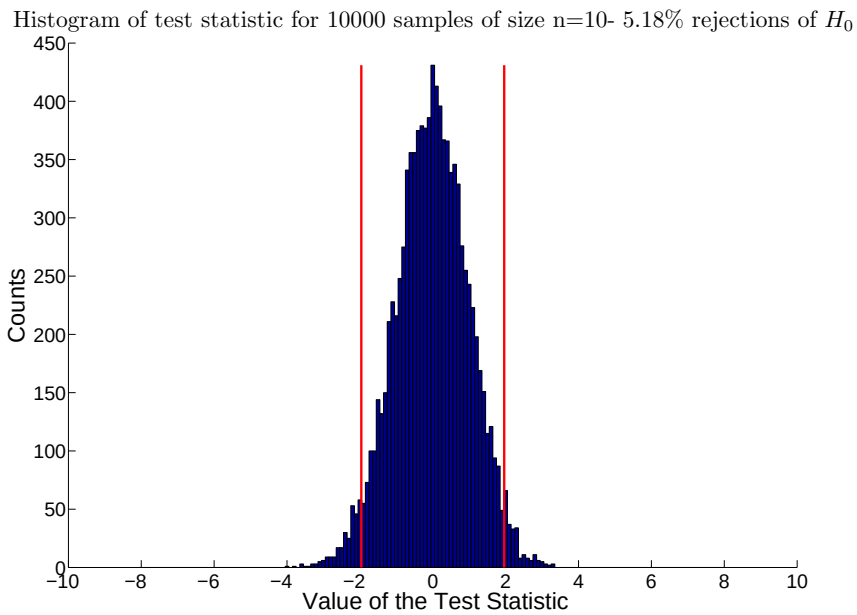
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Sample 1	-2.80	-0.62	-0.37	-0.58	0.58	-0.66	0.38	-4.40	-2.04	-2.31
Sample 2	0.59	-2.67	1.43	3.25	-1.38	1.72	2.51	-3.19	-2.88	1.14

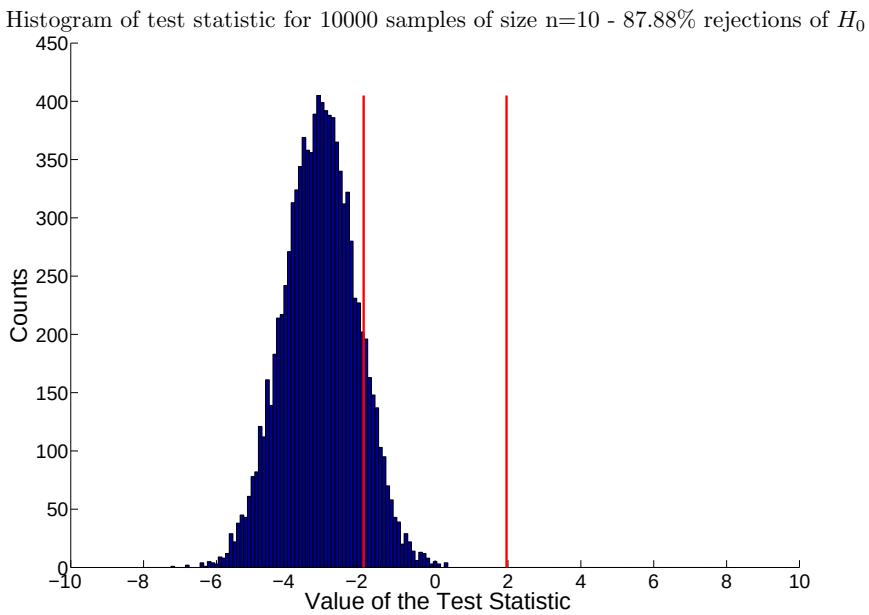
- test statistic for sample 1 is $T = -2.03$
  $\Longrightarrow$ reject null hypothesis (true: $\mu = -1$),
- test statistic for sample 2 is $T = 0.08$
  we do not reject the null hypothesis (true: $\mu = 0$).

Example - Gauss Test IV

Numerical test for α and power

- we draw 10000 samples of size 10 from the same distributions and plot histograms of the value of the test statistic
- rejections: 5.2% for the true parameter and 87.9% for the distribution with true parameter -1.





Confidence regions

Confidence regions

- Hypothesis tests are directly connected to confidence sets of estimated quantities.
- tests make sense if one wants to check some hypothesis about the data.

Definition

A **confidence interval** for a parameter $\theta \in \mathbb{R}$ is an interval $C_n = (a, b)$, where the interval borders $a = a(X_n)$ and $b = b(X_n)$ are functions of the data, such that

$$\mathrm{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta \subseteq \mathbb{R}.$$

The value $1 - \alpha$ is called the **coverage** of the confidence interval. If $\theta \in \mathbb{R}^d$ a **confidence set** (usually an ellipsoid) is a set $C_n \in \mathbb{R}^d$, which depends on the data X_n, such that

$$\mathrm{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta \subseteq \mathbb{R}^d.$$

Confidence regions II

Remark:

- the confidence set C_n is random and θ is fixed,
- it is common to use 95% intervals, that is $\alpha = 0.05$,
- there are two ways of interpreting in probabilistic terms a confidence interval.
  1. If one repeats the **same** experiment over and over again the true parameter θ will be contained in the interval in 95% of the cases. This interpretation is valid but practically usually difficult to realize, since one does an experiment once.
  2. One does k **different** experiments with data from different sources and constructs for each case a confidence interval. Then, in the limit $k \to \infty$, 95% of the constructed k confidence intervals will contain the true parameter.

Confidence regions III

- normal distributions $\mathcal{N}(\mu, \sigma^2)$ with fixed variance σ^2
- i.i.d. samples X_i from $\mathcal{N}(\mu_0, \sigma^2)$, where μ_0 denotes the true parameter.
- the empirical mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has distribution given by $\mathcal{N}(\mu_0, \frac{\sigma^2}{n}) \Longrightarrow \sqrt{n}\frac{\hat{\mu}-\mu_0}{\sigma} \sim \mathcal{N}(0,1)$.

Thus, for every true mean $\mu \in \mathbb{R}$ and every $\alpha \in (0,1)$, we have

$$\mathrm{P}_\mu\left(\left|\sqrt{n}\frac{\hat{\mu}-\mu}{\sigma}\right| \leq q_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

where q_γ is the γ-quantile of $\mathcal{N}(0,1)$. Thus,

$$a(X_n) = \hat{\mu} - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}, \qquad b(X_n) = \hat{\mu} + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}},$$

are the boundaries of the confidence interval $[a(X_n), b(X_n)]$ for the parameter μ.