

# Lecture 8: Test error evaluation and model selection

Isabel Valera

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

16.05.2023

# Outline

- 1 Bibliography
- 2 Test Error Estimation
- 3 Statistical tests
- 4 Summary

# Main references

- ESL - Chapter 7
- Probability theory recap by Prof. Wolf - Chapter 8

# Outline

- 1 Bibliography
- 2 Test Error Estimation
- 3 Statistical tests
- 4 Summary

# Estimating test error

Refer to slides on "Test Error Estimation" (Credits to Prof. Vreeken)

# Outline

- 1 Bibliography
- 2 Test Error Estimation
- 3 Statistical tests**
- 4 Summary

# Relation between test and true error

So far we have focused on getting an estimation of the error of a trained supervised learning model (for either classification or regression) using a test data set with  $m$  samples.

However, the key question we did not address is if our estimator agrees with the true error, i.e., **Test error = True error?**

## Key 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?
- In the case of classification, can we test if the classifier is significantly better than random guessing?

# Idea

## A statistical test:

- formulate a (null) hypothesis  $H_0$  and an alternative hypothesis  $H_1$ , which should be mutually exclusive.
- tries to falsify a given null hypothesis  $H_0$  (e.g. LR and LDA lead to same classification error), in favor of  $H_1$ .
- to this end, it defines a region of rejection which, if  $H_0$  is true, has probability (less than)  $\alpha$  (where  $\alpha$  is the **significance level**),
- computes a test statistic  $T$  (e.g. difference of the test errors of LR and LDA),
- rejects the null hypothesis if  $T$  attains a value in the region of rejection otherwise we keep the null hypothesis, e.g.,
  - If we reject the null hypothesis, we say that the difference between LR and LDA is **statistically significant**.
  - Otherwise, we cannot make any statement about the relation between RL and LDA.



# Definition

## A (parametric) statistical test

- 1 Let  $\Theta$  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  $\theta \in \Theta \setminus \Theta_0$ ,
- 2 A significance level  $\alpha$  is chosen.
- 3 A test statistic  $T$  is a function of the  $m$  samples  $D_m$  (e.g., the test data set), and thus a random variable. The distribution of  $T$ , given  $H_0$  is true, is known. A region of rejection  $B_m$  is chosen, such that if the null hypothesis is true

$$\forall \theta \in \Theta_0, \quad P_\theta(T(D_m) \in B_m) \leq \alpha.$$

- 4  $H_0$  is rejected (one assumes that  $H_1$  holds) if  $T(D_m) \in B_m$ .

The test can be an equality  $H_0 : \theta = \theta_0$  (**two-sided test**) or inequality  $\Theta = \mathbb{R}$ ,  $H_0 : \theta \geq \theta_0$  (**one-sided test**)

# 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

The **type I error** is  $\alpha = P\{\text{reject } H_0 | \theta; H_0 \text{ is true}\}$ . Typically,  $\alpha$  is chosen very small, e.g.,  $\alpha \in \{0.01, 0.05, 0.10\}$  such that the type I error is kept small and we only reject  $H_0$  with a lot of confidence.

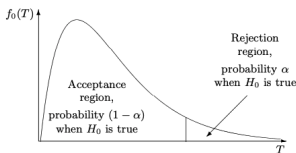
Let  $P_\theta$  be the probability measure with parameter  $\theta$ , then the **power function** of a test is  $\beta(\theta) = P_\theta(T(D_m) \in B_m)$ . The rejection region  $B_m$  is chosen such that,  $\beta(\theta) \leq \alpha$ , for all  $\theta \in \Theta_0$ .

The **type II error** is  $1 - \beta(\theta)$  for  $\theta \in \Theta \setminus \Theta_0$  and corresponds to  $P\{\text{reject } H_0 | \theta; H_1 \text{ is true}\}$

**Goal:** high power!

# Level $\alpha$ -test

- 1 Compute a test statistic  $T$ .
- 2 Consider the null distribution of  $T$ , given  $H_0$  is true, and find the portion that corresponds to  $\alpha$ , i.e. the region where we reject  $H_0$ .
- 3 Accept  $H_0$  if  $T$  lies in the acceptance region and reject it otherwise.



In the illustration on the left,  $T$  is expected to be large if  $H_A$  is true. Therefore, the rejection region is at the right tail of the distribution. In a two-sided test, we consider portions of  $\alpha/2$  at both tails of the distribution.

We always have that

$$\begin{aligned} P\{T \in \text{acceptance region} \mid H_0 \text{ is true}\} &= 1 - \alpha \\ \text{and} \\ P\{T \in \text{rejection region} \mid H_0 \text{ is true}\} &= \alpha. \end{aligned}$$

*Example:* The Standard Normal Null Distribution (Z-test) assumes that the null distribution of  $T$  is a standard normal (i.e., zero-mean and unit variance).

# p-Value

## Definition

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

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

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

If we have

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

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

$$\text{p-value} = P_{\theta_0}(T(D_m) \geq t_{\text{obs}}).$$

# Example - Z-test

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

$$T(D_m) = \sqrt{m} \frac{\frac{1}{m} \sum_{i=1}^m x_i - \mu_0}{\sigma}.$$

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

$$P\left(|T(D_m)| > q_{1-\frac{\alpha}{2}}\right) = \alpha.$$

- **Power function:**

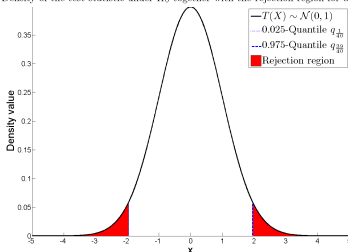
$$\begin{aligned} \beta(\mu) &= P_\mu\left(|T(D_m)| > q_{1-\frac{\alpha}{2}}\right) \\ &= 1 - \Phi\left(q_{1-\frac{\alpha}{2}} - \sqrt{m} \frac{\mu - \mu_0}{\sigma}\right) + \Phi\left(-q_{1-\frac{\alpha}{2}} - \sqrt{m} \frac{\mu - \mu_0}{\sigma}\right), \end{aligned}$$

here  $\Phi$  denotes the cumulative distribution of the standard normal, i.e.,

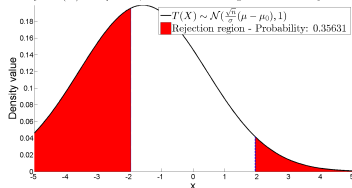
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx = P(X \leq x) \text{ with } X \sim \mathcal{N}(0, 1)$$

# Example - Z-test II

Density of the test statistic under  $H_0$  together with the rejection region for  $\alpha = 0.05$



Density of  $T(X)$  with  $\mu = -1$ ,  $\sigma = 2$  and  $n = 10$  together with the rejection region



**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  $\mu = -1$ ,  $\sigma = 2$  and  $n = 10$ .

# Example - Z-test III

## Numerical example:

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

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$	$T$
Sample 1	-2.80	-0.62	-0.37	-0.58	0.58	-0.66	0.38	-4.40	-2.04	-2.31	-2.03
Sample 2	0.59	-2.67	1.43	3.25	-1.38	1.72	2.51	-3.19	-2.88	1.14	0.08

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

# Applications in ML

- Model comparison and selection:
  - Check if a new ML model leads to an improvement over another method that is statistically significant (the null hypothesis is that the new method has smaller error than the other). This approach is often used for feature selection.
  - Compare several classification methods with the chosen one to know if the latter is better than all the other ones. In this case the null hypothesis is that all classification methods perform similarly.

**Check Example Statistical Test in Linear Regression**



# Outline

- 1 Bibliography
- 2 Test Error Estimation
- 3 Statistical tests
- 4 Summary**

# Ideal way of doing model selection

- 1 Partition the data into: training, validation and test set.
- 2 Train the different models/methods (with different parameters and complexities) on the training set.
- 3 Evaluate performance of all classifiers/parameters on the validation set.
- 4 Select the best method on the validation set. Here, a statistical test can be run here to analyze statistical significance of our choice.
- 5 (*Optional*) Re-train selected model using training and validation sets (useful when dataset is small).
- 6 Estimate the true performance metric (e.g., error) of the chosen model in the test (held-out) set.

*Note:* Steps 1-4 can be replaced by any cross-validation approach, e.g., k-fold.

*More about cross-validation techniques [here](#)*