Lecture 9: Test error evaluation and model selection

Prof. Dr. Mario Fritz

2022-05-22

https://fritz.cispa.saarland
Trustworthy AI
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

Outline

Bibliography

Test Error Estimation

Statistical tests

Summary

Main references

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

Outline

Estimating test error

Refer to slides: “L09-TestError.pdf” (Credits to Prof. Vreeken)

Outline

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) α (where α 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 different**.
  - Otherwise, we cannot make any statement about the relation between RL and LDA.

Definition

A (parametric) 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.
3. A test statistic T is a function of the n samples X_n, and thus a random variable. The distribution of T, given H_0 is true, is known. A region of rejection B_n is chosen, such that if the null hypothesis is true

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

4. H_0 is rejected (one assumes that H_1 holds) if $T(X_n) \in B_n$.

Test can be **parametric** or **nonparametric**. The test can be an equality $H_0 : \theta = \theta_0$ (**two-sided test**) or inequality $\Theta = \mathbb{R}$, $H_0 : \theta \gtreqless \theta_0$ (**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, α 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.

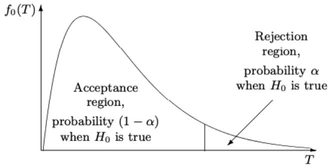
Let P_θ be the probability measure with parameter θ, then the **power function** of a test is $\beta(\theta) = \mathrm{P}_\theta(T(X_n) \in B_n)$. The rejection region B_n has been chosen such that, $\beta(\theta) \leq \alpha, \text{ for all } \theta \in \Theta_0$.

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

Goal: high power!

Level α-test

- Compute a test statistic T.
- Consider the null distribution of T, given H_0 is true, and find the portion that corresponds to α, i.e. the region where we reject H_0.
- 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

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

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 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 - 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

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

- 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:**

$$\beta(\mu) = \mathrm{P}_\mu\Big(|T(X)| > q_{1-\frac{\alpha}{2}}\Big) = 1 - \Phi\Big(q_{1-\frac{\alpha}{2}} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\Big) + \Phi\Big(-q_{1-\frac{\alpha}{2}} - \sqrt{n}\frac{\mu-\mu_0}{\sigma}\Big),$$

here Φ denotes the cumulative distribution of the standard normal, i.e.,
$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{x^2}{2}} dx = P(X \leq x)$ with $X \sim \mathcal{N}(0,1)$

Example - Z-test II

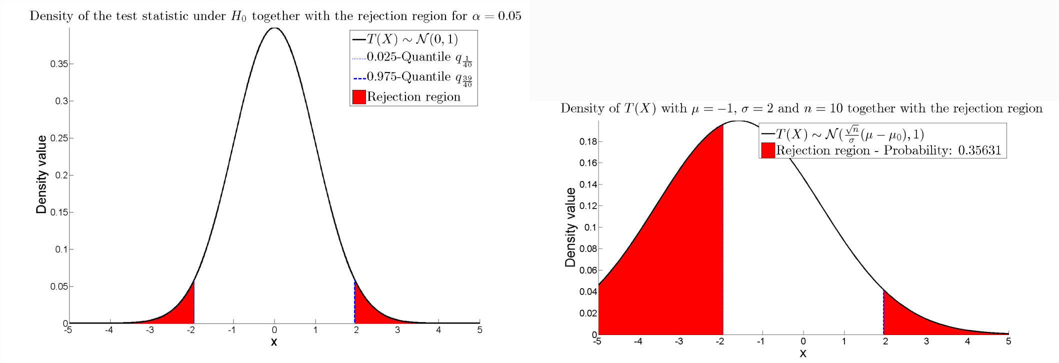


Figure 1: Left: The distribution of the test-statistic under the null hypothesis together with the rejection region for the significance level $\alpha = 0.05$. Right: The computation of the power of the test for $\mu = -1$, $\sigma = 2$ and $n = 10$.

Example - Z-test III

Numerical example:

- 10 samples from Gaussians with $\sigma = 2$.
- Test $H_0 : \mu = 0$ with $\alpha = 0.05 \Longrightarrow$ 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}	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 $T = -2.03$
  $\Longrightarrow$ reject null hypothesis (true: $\mu = -1$),
- test statistic for sample 2 is $T = 0.08$
  we do not reject the null hypothesis (true: $\mu = 0$).

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.

L09 - Example Statistical Test in Linear Regression.pdf

Outline

Ideal way of doing model selection

- Partition the data into: training, validation and test set.
- Train the different models/methods (with different parameters and complexities).
- Compute error of all classifiers/parameters on the validation set.
- Select the best method (statistical test can be run here to analyze statistical significance).
- Train on training and validation set and estimate the true error of the chosen classifier by computing its test error on the test (hold-out) set.