Machine Learning

Performance Measures and Statistical Tests II

Prof. Matthias Hein

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

Lecture 16, 18.12.2013

Program of last time and today

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.

A nonparametric test - the permutation test

- **Nonparametric test:** for testing whether two independent samples

  $$X_1, \dots, X_m \sim F_X, \quad \text{and} \quad Y_1, \dots, Y_n \sim F_Y,$$

  come from the same distribution.
- **Null hypothesis:** $H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y.$
- No specific **test statistic**, but test statistic determines test properties.
  Example: test if means are different

  $$T(X_1, \dots, X_m, Y_1, \dots, Y_n) = |\hat{\mu}_X - \hat{\mu}_Y|,$$

- All permutations of the data: $Z = (X_1, \dots, X_m, Y_1, \dots, Y_n)$, where $|Z| = n + m =: N$, are **equally likely** to be obtained as a sample from F_X and F_Y **under** H_0.
- **Idea:** Compute test statistic for all permutations and compare the value with the original test statistic.

A nonparametric test - the permutation test II

- The distribution P_0 that puts $\frac{1}{N!}$ mass on each value x of the test statistic is called the **permutation distribution** of T.

$$F_0(x) = \mathrm{P}_0\big(T \leq x\big) = \frac{1}{N!}\sum_{j=1}^{N!} \mathbb{1}_{T_j \leq x}.$$

- We define the **rejection region** R for the chosen **significance level** α,

$$\mathrm{P}_0\big(T \in R\big) \leq \alpha.$$

- If the rejection region R is symmetric around zero, then the p-value of the observed value $t_{\mathrm{org}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is

$$\begin{aligned} \text{p-value} &= \mathrm{P}_0\big(|T| > |t_{\mathrm{org}}|\big) = F_0(-t_{\mathrm{org}}) + 1 - F_0(t_{\mathrm{org}}) \\ &= \frac{1}{N!}\sum_{j=1}^{N!} \mathbb{1}_{T_j < -|t_{\mathrm{org}}|} + \mathbb{1}_{T_j > |t_{\mathrm{org}}|}. \end{aligned}$$

A nonparametric test - the permutation test III

Example:

	Permutations	Test statistic T
original data	$(2,5,6)$	2.5
	$(2,6,5)$	1
	$(5,6,2)$	3.5
	$(5,2,6)$	2.5
	$(6,2,5)$	1
	$(6,5,2)$	3.5

Table : Example: Suppose our data is $X_1 = 2, X_2 = 5$ and $Y_1 = 6$. We compute $T(X_1, X_2, Y_1) = |\hat{\mu}_X - \hat{\mu}_Y| = 2.5$. The table shows all 6 permutations together with the values of T. The p-value is given as $\mathrm{P}_0(T > 2.5) = \frac{1}{3}$.

A nonparametric test - the permutation test III

Problem: All permutations not feasible for reasonable N !
Solution: Monte Carlo sampling - sample randomly permutations and average.
Approximate permutation distribution:

$$F_0(x) = \mathrm{P}_0\big(T < x\big) \approx \frac{1}{S} \sum_{j=1}^{S} \mathbb{1}_{T_j < x}.$$

One can derive bounds how many random permutations S one has to draw for a certain accuracy of the p-value (small p, large S).

Usefulness:

- small sample size (parametric test do not apply),
- large sample size - often parametric tests easier to compute,
- partially still useful if data is not independent

Multiple tests

Goal: We want to test several hypothesis e.g. our new classifier is better than SVM, LDA, Logistic Regression,... $\Longrightarrow$ **multiple tests**

Problem: For each individual test the probability of false rejection is α $\Longrightarrow$ but the probability of at least one false rejection is much higher !

Solution: correct the test and/or signficance level.

- huge literature on that topic !
- we discuss only the most simple and very conservative method ($\Longrightarrow$ with low power) $\Longrightarrow$ **Bonferroni correction**

Multiple tests II - Bonferroni correction

Suppose we have m different hypothesis tests

$$H_0^i \quad \text{versus} \quad H_1^i, \quad i = 1, \ldots, m,$$

and let $p_1, \ldots, p_m$ be the computed p-values of the tests.
Then **reject** the null hypothesis H_0^i if $p_i \leq \frac{\alpha}{m}$.

Theorem

Using the Bonferroni method, the probability of falsely rejecting **any** *null hypothesis is less than or equal to α and greater than $\frac{\alpha}{m}$.*

Proof: Let R be the event of at least one false rejection and let R_i be the event that the i-th null hypothesis is falsely rejected which has using Bonferroni the probability $\mathrm{P}(R_i) = \frac{\alpha}{m}$. Then,

$$\mathrm{P}(R) = \mathrm{P}\Big(\cup_{i=1}^m R_i \Big) \leq \sum_{i=1}^m \mathrm{P}(R_i) \leq \sum_{i=1}^m \frac{\alpha}{m} = \alpha.$$

On the other hand, $\mathrm{P}(R) = \mathrm{P}(\bigcup_{i=1}^m R_i) \geq \min_{i=1,\ldots,m} \mathrm{P}(R_i) = \frac{\alpha}{m}$.

Evaluation of a classifier

Setting: We have computed the test error of a classifier f for m test samples.

Questions:

- Can we make any assertions if the **true error** of f is close to the **test error** ?
- For a given confidence level and sample size can we give a **confidence interval for the true error** of f given the error on an independent test set ?
- Given a confidence interval and confidence level **how many test samples do we need ?**

Evaluation of a classifier - Setting

Setting:

- a classifier f_n trained on T_n, the training set of n i.i.d. samples (X_i, Y_i) (note that f_n is random due to its dependence on the training set)
- a test (evaluation) set $E_m = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ of size m drawn i.i.d. from the data generating measure,
- the test error $\hat{R}_m(f_n) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{f_n(X_i) \neq Y_i}$.
- the true error (expected loss) $R(f_n)$ of f_n, $R(f_n) = \mathbb{E}[\mathbb{1}_{f_n(X) \neq Y} \mid T_n]$.

Important note: For simplicity the dependence of the learned classifier f_n on the training sample is dropped and we denote the classifier as f. However, every statement made in this chapter is **conditional on the training sample** T_n. The dependence of f_n on the training sample is the topic of statistical learning theory. **The randomness in this section enters through the test sample.**

Evaluation of a classifier - Main idea

Main insight:

The test error $\hat{R}_m(f)$ as an estimator of the true error $R(f)$ is *unbiased*:

$$\mathbb{E}[\hat{R}_m(f)] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathbb{1}_{f(X_i)\neq Y_i}] = \mathbb{E}[\mathbb{1}_{f(X)\neq Y}] = R(f),$$

since the test sample is drawn i.i.d..

Given an i.i.d. test sample of m samples, the number of errors made on this sample will be distributed like a binomial distribution with parameters m and $R(f)$.

$$\sum_{i=1}^{m}\mathbb{1}_{f(X_i)\neq Y_i} = m\,\hat{R}_m(f) \sim \mathrm{Bin}(m, R(f)).$$

Obviously, we do not know the true error $R(f_n)$ but we know the test error $\hat{R}_m(f) \Longrightarrow$ classical problem - estimation of the parameter of a binomial distribution.

Confidence interval for the test error

Classical Clopper-Pearson confidence interval for the binomial distribution.

Theorem

Let $F(k, m, p) = \mathrm{P}(X \leq k)$ *for* $X \sim \mathrm{Bin}(m, p)$ *and determine for* $0 < \delta < 1$ *the coefficients* a, b *by*

$$F(m\hat{R}_m(f), m, b) = \frac{\delta}{2}, \qquad 1 - F(m\hat{R}_m(f) - 1, m, a) = \frac{\delta}{2},$$

where $\hat{R}_m(f)$ *is the test error computed on* m *i.i.d test samples. If* $\hat{R}_m(f) = 0$, $a = 0$ *and if* $\hat{R}_m(f) = 1$, $b = 1$. *Then we have with probability* $1 - \delta$,

$$a < R(f) < b.$$

How to read this: for all $a < R(f) < b$, there exist $k_1(R(f)), k_2(R(f))$ such that

$$\mathrm{P}\Big(k_1(R(f)) < m\hat{R}_m(f) < k_2(R(f))\Big) \geq 1 - \delta.$$

Confidence interval continued

Problem: CP-confidence interval is complicated.
$\Longrightarrow$ not easily resolvable for the sample size.

Solution: coarser confidence interval which can be handled more easily.
We need an upper bound for:

$$\mathrm{P}\Big(|\hat{R}_m(f) - R(f)| > \varepsilon\Big).$$

Hoeffding's inequality

Theorem (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Then for any $\varepsilon > 0$ we have

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i\Big| \geq \varepsilon\Big) \leq 2\exp\Big(-\frac{2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

In particular if additionally $X_1, \ldots, X_n$ are identically distributed, we have

$$\mathrm{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X\Big| \geq \varepsilon\Big) \leq 2\,\exp\Big(-\frac{2n\varepsilon^2}{(b-a)^2}\Big).$$

- Such inequalities are called **concentration inequalities** (quantify the probability that the empirical mean deviates from its expectation).
- for binomial distribution: $a = 0$ and $b = 1$.

Confidence interval

Theorem

For a test set of size m we have with probability $1-\delta$,

$$|R(f)-\hat{R}_m(f)| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2}{\delta}\right)}.$$

In order to ensure that

$$|R(f)-\hat{R}_m(f)| \leq \varepsilon,$$

with probability $1-\delta$ *the required number of test samples m is given by,*

$$m \geq \frac{1}{2\varepsilon^2}\log\left(\frac{2}{\delta}\right).$$

Some numbers:

- for $m = 1000$ and $\delta = 0.05$, $|R(f)-\hat{R}_m(f)| \leq 0.0429$.
- for $\delta = 0.05$ and $\varepsilon = 0.01$, $m \geq 18444$.

Confidence interval

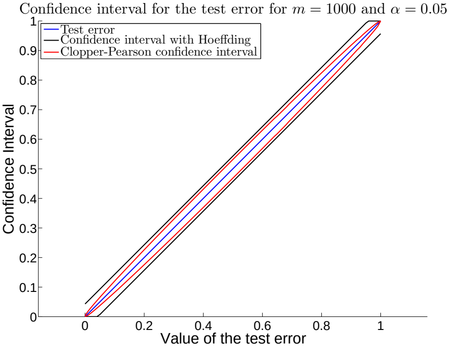
Proof: We have by Hoeffding's inequality

$$\mathrm{P}\Big(|R(f) - \hat{R}_m(f)| > \varepsilon\Big) \leq 2e^{-2m\varepsilon^2} := \delta.$$

Thus, $\varepsilon = \sqrt{\frac{1}{2m}\log\left(\frac{2}{\delta}\right)}$. The complementary event $|R(f) - \hat{R}_m(f)| \leq \varepsilon$ holds at least with probability $1 - \delta$ and with the derived form of ε we get the first result. For the other result note that for $m \geq \frac{1}{2\varepsilon^2}\log\left(\frac{2}{\delta}\right)$ we have

$$\mathrm{P}\Big(|R(f) - \hat{R}_m(f)| > \varepsilon\Big) \leq \delta.$$

Is it much worse than Clopper-Pearson ?



The confidence intervals generated by the exact Clopper-Pearson interval and the ones generated by Hoeffding for $n = 1000$ and $\alpha = 0.05$. Note, that in particular for very small (and very large) test errors the exact confidence interval is much tighter.

How to compare classifiers

Setting:
Given two classifiers f_1 and f_2 trained on a common training set T_n.

Problem:
How can we determine that one classifier is really better than the other ? Or in other words: how can we determine if their difference in test error is not just a result of statistical fluctuations but is really due to a difference in the true error.

Solution: Do a statistical test !

Comparison of two classifiers on a test set - t-Test

- $(\hat{Y}_1^1, \ldots, \hat{Y}_m^1)$ and $(\hat{Y}_1^2, \ldots, \hat{Y}_m^2)$: predictions of classifier f^1 and f^2,
- **Null hypothesis:** The true error of both classifiers is equal,
- We define $W_i = \mathbb{1}_{\hat{Y}_i^1 \neq Y_i} - \mathbb{1}_{\hat{Y}_i^2 \neq Y_i} \Rightarrow \frac{1}{m}\sum_{i=1}^m W_i = \hat{R}_m(f^1) - \hat{R}_m(f^2)$.
- **Test statistic**, with $\sigma_W^2 = \frac{1}{m-1}\sum_{i=1}^m (W_i - \frac{1}{m}\sum_{i=1}^m W_i)^2$,

$$T(\hat{Y}^1, \hat{Y}^2) = \sqrt{m}\frac{\hat{R}_m(f^1) - \hat{R}_m(f^2)}{\sigma_W}$$

and the region of rejection is given by $|T(\hat{Y}^1, \hat{Y}^2)| \geq q_{1-\frac{\alpha}{2}}$, where q_α is the α-Quantile of the student t- distribution with $m-1$ degrees of freedom which has the density

$$p(x) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi}\,\Gamma\left(\frac{m-1}{2}\right)}\left(1 + \frac{x^2}{m-1}\right)^{-\frac{m}{2}}.$$

Comparison of two classifiers - permutation test

- The test statistic becomes with $W_i = \mathbb{1}_{\hat{Y}_i^1 \neq Y_i} - \mathbb{1}_{\hat{Y}_i^2 \neq Y_i}$,
$$T(\hat{Y}^1, \hat{Y}^2) = \sum_{i=1}^{m} W_i.$$
- Under $\mathcal{H}_0$: true error of both classifiers is equal $\Rightarrow$ exchanging predictions (flipping W_i) does not change their error. Compute:
$$T(\pi) = \sum_{i=1}^{m} \pi_i W_i,$$
where $\pi \in \{-1, 1\}^m$ and thus there are 2^m possible values.
- For $m > 20$ infeasible $\Rightarrow$ random sampling,
$$\pi_i = \begin{cases} 1, & \text{with prob. } \frac{1}{2}, \\ -1, & \text{with prob. } \frac{1}{2}. \end{cases}.$$
- The (approximate) p-value can be computed as
$$p - value = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{|T(\pi)| > |T(\hat{Y}^1, \hat{Y}^2)|}.$$

Relation between test and true error

Discussion

- paired t-test can be done fast. However, the parametric assumption (t-distribution) would only be true given that $W_i = \mathbb{1}_{\hat{Y}_i^1 \neq Y_i} - \mathbb{1}_{\hat{Y}_i^2 \neq Y_i}$ is Gaussian distributed (which it is not).
- the permutation test is slow and suffers from the discrete distribution problem, however there is no assumption made about the data distribution

In practice: paired t-test is usually done, but for small sample sizes one should consider the permutation test.

Test against multiple classifiers: use Bonferroni correction.

Model selection

Questions:

How to determine the optimal parameters of a classifier (regularization parameter λ,...) ?
How to select among a set of classifiers the best one ?

Solution:

- ideal way - training/validation/test set.
- practical way - cross validation.

Model selection II

Model selection based on validation set

- partition the data into: training, validation and test set,
- train the different classifiers (with different parameters),
- compute error of all classifiers/parameters on the validation set,
- select the best classifier,
- train on training and validation set and estimate the true error by computing the error on the test set.

The validation set is often also called **hold-out** sample since it is not used during training.

Remark: If one is not interested in the error estimate, one can discard the testing partion.

Hold out selection - Discussion

Pro:

- Clear separation between training and validation set $\Longrightarrow$ no interdependencies which could lead to overfitting.
- The procedure is less computationally expensive then cross validation ($\Longrightarrow$ for large data set method of choice)
- rigorous analysis possible - consistency of model selection.

Contra:

- a lot of data is wasted for the validation and test set. In particular, if one expects small true errors than one needs large validation sets in order to discriminate between different parameters/classifiers,
- choice of parameters depends on the number of samples $\Longrightarrow$ a parameter which is optimal for the classification of the validation set, need not be optimal anymore if one trains on the training **plus** the validation set.

Consistency of model selection

We select the classifier which has smallest validation error,

$$f' = \arg\min_{f \in \mathcal{F}} \hat{R}_m(f),$$

where $\mathcal{F}$ is our set of parameters/classifiers.

Lemma

Let f' be the minimizer of the validation error $\hat{R}_m(f)$ and denote by $R(f)$ the true error of f, then

$$|R(f') - \inf_{f \in \mathcal{F}} R(f)| \leq 2 \sup_{f \in \mathcal{F}} |\hat{R}_m(f) - R(f)|.$$

Proof:

$$\begin{aligned} R(f') - \inf_{f \in \mathcal{F}} R(f) =& R(f') - \hat{R}_m(f') + \hat{R}_m(f') - \inf_{f \in \mathcal{F}} R(f) \\ \leq& R(f') - \hat{R}_m(f') + \sup_{f \in \mathcal{F}} \big(\hat{R}_m(f) - R(f)\big) \\ \leq& 2 \sup_{f \in \mathcal{F}} |\hat{R}_m(f) - R(f)|, \end{aligned}$$

where in the first inequality we used $\hat{R}_m(f) \geq \hat{R}_m(f'), \quad \forall f \in \mathcal{F}$.

Consistency of model selection II

Let $\mathcal{F}$ be the *finite* set of classifiers/parameters one is using with $N = |\mathcal{F}|$. Using a union bound together with Hoeffding's inequality one gets

$$\mathrm{P}\Big(\sup_{f \in \mathcal{F}} |R(f) - \hat{R}_m(f)| > \varepsilon \Big) = \mathrm{P}\Big(\cup_{i=1}^N \big\{ |R(f_i) - \hat{R}_m(f_i)| > \varepsilon \big\} \Big)$$

$$\leq \sum_{i=1}^N \mathrm{P}\Big(|R(f_i) - \hat{R}_m(f_i)| > \varepsilon \Big) \leq 2\,N\,e^{-2m\varepsilon^2}.$$

Theorem

Let f' be the minimizer of the validation error $\hat{R}_m(f)$. Then with probability $1-\delta$, we have

$$R(f') \leq \inf_{f \in \mathcal{F}} R(f) + \sqrt{\frac{2}{m} \log\Big(\frac{2N}{\delta}\Big)}.$$

$\Longrightarrow$ for sufficiently large m the chosen classifier f' is close to the optimal one in $\mathcal{F}$.

Model selection based on cross validation

K-fold Cross validation

- partition the N data-points into K folds
$$\{(X_1^1, Y_1^1), \ldots, (X_{N_1}^1, Y_{N_1}^1), \ldots, (X_1^K, Y_1^K), \ldots, (X_{N_K}^K, Y_{N_K}^K)\},$$
of size $N_1, \ldots, N_K$ (where the N_k are roughly equal),
- select the k-the fold for testing and train on the rest of the data which yields a classifier f^k (for all sets of parameters),
- compute test error on the k-th fold, $R^k(f^k) = \sum_{i=1}^{N_k} \mathbb{1}_{f^k(X_i^k) \neq Y_i^k}$,
- do that for all K folds
- compute the cross-validation error $R^{CV} = \frac{1}{N} \sum_{k=1}^{K} R^k(f^k)$,
- select the parameter with the smallest cross-validation error.
- train on the whole dataset using the optimal parameter.

Model selection based on cross validation II

Leave one-out cross validation:
Number of folds $K = N$ - train on all but one datapoint and test on the left out one.

In practice:

- 5-fold or 10-fold cross validation

Model selection based on cross validation III

Advantages of cross validation:

- All parts of the data are used for finding the optimal parameter ⇒ can have less variance then the holdout method (depends on the size of the folds).

Disadvantages of cross validation:

- Often computationally very expensive.
- It is quite difficult to upper bound the true error in terms of the cross-validation error.

Model selection - Overfitting

Overfitting:

- for small sample sizes and a large number of parameters/classifiers both model selection techniques can overfit.
- if one has many parameters just by chance one of them can work well also on the test set.

Sanity check:

- Use permutation test with null hypothesis that inputs and labels are independent (there is nothing to learn in this case because there is no dependency between input and output)
- repeat cross-validation with randomly permuted labels and compute the cross-validation error.
- compute the p-value based on the distribution of cross-validation error.