Introduction to Machine Learning
NPFL 054

`http://ufal.mff.cuni.cz/course/npfl054`

Barbora Hladká
hladka@ufal.mff.cuni.cz

Martin Holub
holub@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Lecture 4 — Evaluation and Statistical tests

Outline

I. Remarks on evaluation

- Sample error and generalization error – a brief recap
- Cross-validation, leave-one-out, bootstrap heuristic
- Binary classifier evaluation metrics

II. Statistical hypothesis testing

- General principles of hypothesis testing
  - classical examples, fair die, classifier accuracy
  - null hypothesis, test statistic, p-values, significance and confidence levels
  - confidence intervals
- Testing the mean of normal population
  - t-test and confidence interval for the mean
- Chi-square tests
  - goodness-of-fit test, independence test

Fundamentals of classifier evaluation

Definition (Empirical and sample error)

Given a sample set S, the **empirical error** (aka *observed error*) of classifier h is the observed number of errors that h does on S.

The **sample error** of hypothesis h with respect to target function f and data sample S is the proportion of examples that h misclassifies

$$\text{error}_S = \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

where

- $n = |S|$ is the sample size
- $f(x)$ is the true classification of example x
- $h(x)$ is the predicted class of example x
- $\delta(f(x) \neq h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise.

Sample error and generalization error

Definition (Generalization error)

The **generalization error** (aka *true error*) of hypothesis h with respect to target function f and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn randomly according to $\mathcal{D}$.

$$\text{error}_{\mathcal{D}} = \Pr_{x \in \mathcal{D}} \{\delta(f(x) \neq h(x)\}$$
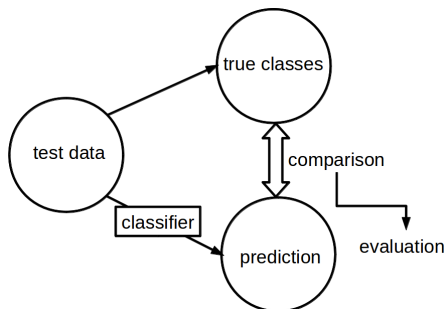
Generalization error – how to estimate?

Typically, the generalization error is not an observable quantity because the distribution $\mathcal{D}$ is usually unknown.

$\longrightarrow$ **The question is**

How well does error_S estimate $\text{error}_{\mathcal{D}}$?
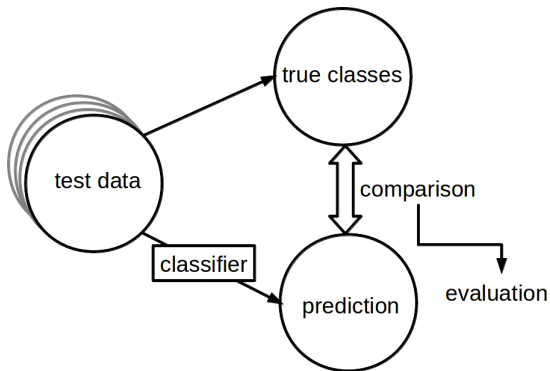
Classifier evaluation

The evaluation process

Is it enough to test your classifier on one test set?
You can get a good/bad result by chance!

The ideal evaluation

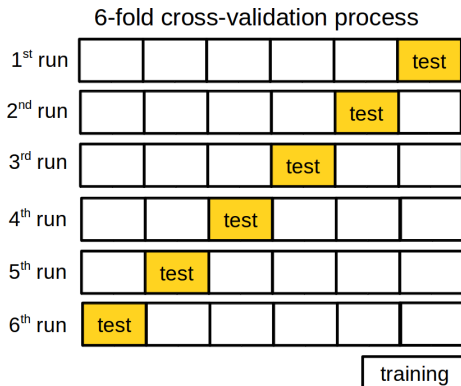**The more test data, the more confident evaluation . . .**

k-fold cross-validation

Development working data is partitioned into k subsets
of equal (or approximately equal) size. Then you do k iterations.

In the i-th step of the iteration, the i-th subset is used as a test set, while the remaining parts form a training set.

Example

6-fold cross-validation process

1st run						test
2nd run					test	
3rd run				test		
4th run			test			
5th run		test				
6th run	test					

training

Using k-fold cross-validation: Which k is the best?

The goal: get a good estimate of generalization error
⇒ low bias: not to underestimate, nor overestimate
⇒ low variance: low sensitivity to the data sample

Small k (k close to 2)
⟶ small training sets, error rate tends to be overestimated

Large k (up to the data set size)
⟶ could be computationally demanding = main practical problem
⟶ small test sets, training sets are almost identical
⟶ low bias, but high variance

Heuristic recommendation: cca $5 \leq k \leq 10$
⟶ moderate bias, moderate variance, moderate computational cost
⟶ has been empirically shown to yield good eror rate estimates

Stratified cross-validation – each class should be represented in roughly the same proportion as in the entire data set
⟶ if data sets are small, the risk of purely ramdom split should be avoided

Leave-one-out cross-validation (LOOCV)

Extreme case: $k = n$, where n is the size of the development data set
  $\longrightarrow$ leave-one-out method (LOOCV)

Advantages

- using maximum training sets $\rightarrow$ low bias
- no randomness in data splitting

Disadvantages

- training sets are almost identical(!) $\rightarrow$ high bias
  – LOOCV averages the outputs of n models that are highly positively correlated with each other
  – high bias is the reason why k-fold CV with moderate k often gives more accurate estimates of the test error than does LOOCV
- may be (typically) too time-consuming
- similar class distribution in training and test data is not guaranteed
  – The extreme case: 50 % class A, 50 % class B. Then the trivial MFC classifier has true error 50 %, but LOOCV estimate is 100 % accuracy!

Recommended evaluation procedure

Typically, use k-fold cross-validation for $k = 5$ or $k = 10$ for estimating the performance (accuracy, etc.)

Then compute

- the mean value of performance estimate
- standard deviation
- confidence intervals

Report mean values of performance estimates and their standard deviations, or (better) 95 % confidence intervals around the mean.

A simple sampling method

Motivation: When the total number of examples is very small (≤ 50), even the leave-one-out method becomes unreliable.

- repeat 2-fold cross-validation (e.g. 100 times)

- it has been shown that the average quality of estimation is better than the leave-one-out method

Bootstrapping principle

Bootstrap sampling – generating different training subsets

- New data sets $D_1, \ldots, D_K$ are drawn from an original data set D
  with replacement, each of the same size as the original $|D| = n$.

- Then in the i-th step of the iteration, D_i is used as a training set, while all
  the other examples $\mathbf{x} \in D \setminus D_i$ form the actual test set.

How many examples will appear in the bootstrap samples?

- The probability that we pick an instance is $1/n$, and the probability that we
  do not pick an instance is $1 - 1/n$. The probability that we do not pick an
  instance after n draws is $(1 - 1/n)^n \approx e^{-1} \doteq 0.368$.

- It means that for training the system will not use 36.8 % of the data, and
  thus the error estimate will be rather pessimistic.

A simple bootstrap heuristic

- Suppose a development data set of n examples
- An optimistic error rate e_ℓ of the model is obtained by building and testing on all available examples
  - Train a model using all n examples
  - Get training error $=$ optimistic estimate e_ℓ.
- A pesimistic error rate e_0 is obtained by make 200 bootstrap samples
  - Randomly select n examples with replacement and train a model
    - on average, it will be 63.2 % of the original set
  - Test the model on the examples not used in the training
    - on average, it will be 36.8 % of the original set
  - Get the test error
  - Get mean test error as an average quality $=$ pesimistic estimate e_0.
- Finally, the error ".632 estimator" is defined as a linear combination

$$e = 0.368 \cdot e_\ell + 0.632 \cdot e_0$$

Notes

The .632 estimator can break down in overfitting situations (when e_ℓ is close to 0).

The error estimation obtained by a hundred 2-fold CV runs may be used instead of e_0.

Evaluation of binary classifiers
Confusion matrix

Binary classification $\overset{aka}{=}$ **2-class classification** $\overset{aka}{=}$ **0/1 classification**

In binary classification tasks, examples are divided into two disjoint subsets:

- **(true) positive examples** – "to be retrieved" (ones)
- **(true) negative examples** – "not to be retrieved" (zeros)

```
# Example of confusion matrix for binary classification
> table(cv.test$Class, pred.test)
        prediction
          0    1
 true  0 580   69
       1  37  144
>
```
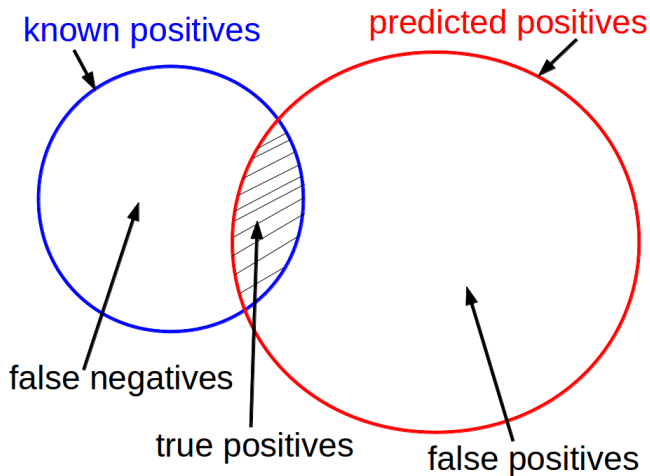
Evaluation of binary classifiers
Confusion matrix

		Predicted class	
		Positive	**Negative**
True class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

- **'Trues'** are examples correctly classified
- **'Falses'** are examples incorrectly classified
- **'Positives'** are examples predicted as positives (correctly or incorrectly)
- **'Negatives'** are examples predicted as negatives (correctly or incorrectly)

known positives

predicted positives

false negatives

true positives

false positives

Evaluation of binary classifiers
Basic performance measures

Measure	Formula
Precision	TP/(TP+FP)
Recall/Sensitivity	TP/(TP+FN)
Specificity	TN/(TN+FP)
Accuracy	(TP+TN)/(TP+FP+TN+FN)

Very often you need to **combine both good precision and good recall**. Then you usually use **balanced F-score**, so called **F-measure**

$$F = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Lady tasting tea – a famous example introduced by R. Fisher (1935)

The example is based on a real story. Fisher met a lady (Muriel Bristol) who claimed to be able to tell whether the tea or the milk was added first to a cup.

First we need to design an experiment to test her ability.
Then we need to meaningfully evaluate the result of the experiment.

Lady tasting tea – Experiment 1

The Lady is provided with 2 randomly ordered cups of tea – 1 prepared by first adding milk, the other prepared by first adding the tea. She should select the one prepared by first adding milk.

Result: The Lady selected the cup prepared by first adding milk.

What can we conclude from this experiment?

Lady tasting tea – Experiment 2

We repeat the Experiment 1 four times.

The Lady is provided with 4 pairs of 2 randomly ordered cups of tea – in each pair one cup is prepared by first adding milk, the other prepared by first adding the tea. From each pair she should select the one prepared by first adding milk.

Result: The Lady selected the 4 cups prepared by first adding milk.

What can we conclude from this experiment?

Obviously, the Experiment 2 is more convincing than the Experiment 1.

Lady tasting tea – Experiment 3 (Fisher's)

In fact, Fisher proposed to give her eight cups, four of each variety, in random order.

The Lady is provided with 8 randomly ordered cups of tea – 4 prepared by first adding milk, 4 prepared by first adding the tea. She should select the 4 cups prepared by first adding milk.

Result: The Lady selected the 4 cups prepared by first adding milk.

What can we conclude from this experiment?

Both Experiment 2 and Experiment 3 indicate that the results are probably not random. Which one is more convincing?

Fisher's experiment – random selection

Compute the probability of getting the observed result **if** the selection is random?

```
# the eight cups -- four T and four F
> cups = c(T,T,T,T,F,F,F,F)

# one million random experiments
> N = 10^6; s = numeric(N)
> for(i in 1:N) s[i] = sum(sample(cups, 4, rep=F))
> table(s)
s
    0      1      2      3      4
14433 228323 514215 228763  14266

# the probability of getting 4 T at random is
> mean(s == 4)
[1] 0.014266
```

Or, since the statistic has hypergeometric distribution, you can simply do

```
> dhyper(4,4,4,4)
[1] 0.01428571
```

Lady tasting tea
– interpretation and analysis of the experiments

How to interpret the three experiments in the framework of "statistical hypothesis testing" originally coined by R. Fisher?

- The **null hypothesis** H_0 is that the Lady has no such ability to recognize cups prepared by first adding milk.
    - null hypothesis means that she (hypothetically) does random selection

- The **test statistic** is a simple count of the number of successes in selecting the correct cups.

- The probability of getting the observed result ($=$ the statistic value) at random is
    - 50 % in the Experiment 1
    - 6.25 % in the Experiment 2
    - 1.43 % in the Experiment 3

Rejecting the null hypothesis based on p-value

Assuming that the null hypothesis is true, the probability of getting the observed result in Experiment 3 is only 1.43%, which *could be considered as* a good reason to **reject the null hypothesis**.

P-value
In statistical tests, p-value is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

> **The null hypothesis is rejected if the p-value is small enough**

So, we need to set a threshold for the p-value.

Test significance level α and confidence level $1 - \alpha$

Could we make wrong decision when rejecting the null hypothesis?

- Yes, in the example above there is 1.43% chance of getting the result even if the Lady selected randomly!
- Such an error is called error of the first kind or "Type I Error".

The null hypothesis should be only rejected when an error is very unlikely

- Therefore we choose a **significance level** α as a threshold for p-value
    - α is the test's probability of incorrectly rejecting the null hypothesis
- Then the null hypothesis will be only rejected when p-value $< \alpha$
- Usually, $\alpha = 5\%$ or 1% or 0.5% or something like that
- The corresponding value $1 - \alpha$ is called **confidence level**
    - which is the probability of not doing the error of the first kind

> **Remember:** The significance level α is a property of the test itself, while p-value is derived from the observed data!

A remark on Czech terms

Česká terminologie

- significance level α = *hladina významnosti testu*

- confidence level $1 - \alpha$ = *hladina spolehlivosti testu*

- interval hodnot statistiky (e.g. t-values), které lze pozorovat
  s pravděpodobností (pouze) α se nazývá *kritický obor*
  $\longrightarrow$ pokud statistika padne do kritického oboru, zamítáme H_0

- α je míra rizika, že uděláme chybu I. druhu, tj. že chybně zamítneme H_0,
  ačkoliv ona platí

- p-value = p-hodnota = *dosažená hladina testu*

Example 2 – Is your die fair?

You have got only 10 sixes when rolling a die 100 times
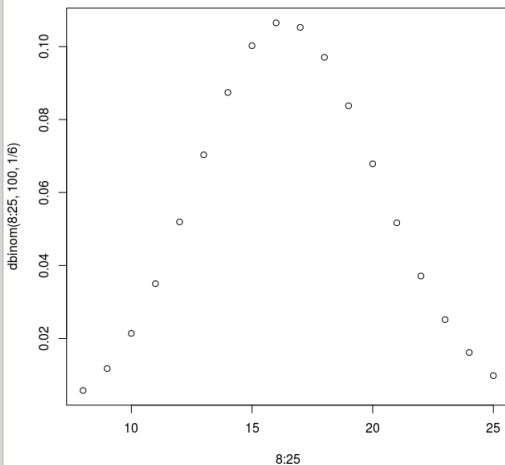
Example 3 – Classifier accuracy

Example

Test sample size $= 100$; there are 73 correctly classified instances.
– Is it possible that classifier accuracy is 76 %?

Example 4 – men's height mean

Assume that the population of men's height is normally distributed with the known variance $\sigma^2 = 100$ and an unknown mean μ. In other words, the men's height will be represented by a continuous random variable X so that

$$X \sim \mathrm{N}(\mu = ?, \ \sigma^2 = 100).$$

We have a sample of $n = 10$ men's heights:

```
> observation
 [1] 174.7 178.0 195.9 181.0 181.6 197.5 184.9 167.6 173.4 175.8
```

Are we able to reliably estimate the mean of the population?

The best estimate is given by the **sample mean**:

```
> mean(observation)
 [1] 181.04
```

Why do you believe that this estimate is the best one?
How confident are you about the estimated mean?

Men's height mean – test statistic distribution

The sample average $\bar{x} = \frac{1}{n} \sum x_i$ will be used as a **test statistic**.

- What is the distribution of the average represented by the random variable $\bar{X}$ when we randomly sample the population?

Theorem

If $X_1, \ldots, X_n$ are independent and have the same distribution $\mathrm{N}(\mu, \sigma^2)$, then

$$\bar{X} = \sum_{i=1}^{n} X_i \quad \text{has the distribution} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

When we formulate a hypothesis about the population mean, we will know the hypothetical distribution of the statistic. Hence, we will be able to compute the probability of the observed statistic.

Men's height mean – first example hypothesis

Let us consider the following null hypothesis about the populatin mean
$H_0 : \mu = 190$.

Under that assumption the distribution of the sample average is
$\bar{X} \sim \mathrm{N}(\mu = 190, \sigma^2 = 100)$.

Then, what is the probability that $\bar{X} \leq 181.04$?

The answer is 0.23 %.

```
> pnorm(181.04, 190, sqrt(10))
[1] 0.00230278
>
```

Similarly, the probability that $\bar{X} \geq 198.96$ is also 0.23 %.

Conclusion
Assuming that the null hypothesis is true, the probability of obtaining the test statistic $\bar{x}$ as extreme or more extreme as the one that was actually observed is only 0.46 % (= the p-value). Hence, with the significance level $\alpha = 5$ % we will reject the hypothesis.

Men's height mean – second example hypothesis

Let us consider the following null hypothesis about the populatin mean
$H_0 : \mu = 180$.

Under that assumption the distribution of the sample average is
$\bar{X} \sim \mathrm{N}(\mu = 180, \sigma^2 = 100)$.

Then, what is the probability that $\bar{X} \geq 181.04$?

The answer is 37.1 %.

```
> 1 - pnorm(181.04, 180, sqrt(10))
[1] 0.3711244
>
```

Similarly, the probability that $\bar{X} \leq 178.96$ is also 37.1 %.

Conclusion
We will not reject the hypothesis. Assuming that the null hypothesis is true, the
probability of obtaining the test statistic $\bar{x}$ as extreme or more extreme as the one
that was actually observed is 74.2 %. If we rejected the hypothesis, we would take
the 74.2 % risk of doing the error of the first kind.

Men's height mean – confidence interval

Obviously, hypothesized values of μ that are relatively close to the observed $\bar{x}$ *would not be rejected*. On the other hand, values that are too far from the observed $\bar{x}$ *would be rejected*.

Question: What is the interval of all possibly hypothesized values of μ that would NOT be rejected?

- to determine this interval you need to know or choose a required significance level α
- this interval is called **confidence interval** for the population mean with the confidence level $1 - \alpha$

Statistical tests – first little summary

Statistical tests are used to test a hypothesis about a population

Always we observe only a sample (often only a small one) of the population.
Then we should make a decision according to this observation.
Having the observed data we compute a test statistic.

The value of the test statistic can be

- in contradiction with the hypothesis
    $\longrightarrow$ then we **reject** the hypothesis

- NOT in contradiction with the hypothesis
    $\longrightarrow$ then we **do not reject** the hypothesis

Example 5 – Confidence interval for the mean of normal population *with known variance* σ^2

Example exercise

Given the confidence level 99 %, find the confidence interval for the men's height population mean based on the given observation. The given assumptions are:

- the men's height population variance $\sigma^2 = 100$
- the observed sample mean $\bar{x} = 181.04$
- the observed sample size $n = 10$

The confidence interval contains all possible values of the population mean that could not be rejected if hypothesized at the given confidence level.

To generally derive how to compute the confidence interval we will use standardized normal distribution and its critical values.

Expected value and variance – basic properties

For any random variables X, Y and any $a, b \in \mathbb{R}$ the following holds true:

- $\mathrm{E}(a + bX) = a + b\,\mathrm{E}X$
- $\mathrm{E}(X + Y) = \mathrm{E}X + \mathrm{E}Y$
- if X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}X\,\mathrm{E}Y$

Variance of a random variable is defined by

$$\mathrm{var}\,X = \mathrm{E}(X - \mathrm{E}X)^2.$$

If a random variable X has finite variance, then the following holds true:

- $\mathrm{var}\,X = \mathrm{E}X^2 - (\mathrm{E}X)^2$
- $\mathrm{var}\,(a + bX) = b^2\,\mathrm{var}\,X$
- $\sqrt{\mathrm{var}\,(a + bX)} = |b|\,\sqrt{\mathrm{var}\,X}$

Standardized random variable

Definition

If a random variable X has non-zero finite variance, then $Z = \dfrac{X - \mathsf{E}X}{\sqrt{\text{var}\, X}}$ is called *standardized* random variable.

Note: If Z is standardized, then $\mathsf{E}Z = 0$ and $\text{var}\, Z = 1$.

Standardized normal distribution − notation

If $X \sim \mathrm{N}(\mu, \sigma^2)$ then standardized variable $Z = \dfrac{X - \mathsf{E}X}{\sqrt{\text{var}\, X}} = \dfrac{X - \mu}{\sigma}$ has the distribution $\mathrm{N}(0, 1)$. Usual notation for standardized normal distribution is

- φ for the density,
- Φ for the distribution function,
- Φ^{-1} for the quantile function.

Quantiles and critical values

Definition

Quantile function of a random variable X is defined as

$$F_X^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\},$$

where F_X is the distribution function of X and $\alpha \in (0, 1)$.

Value $F^{-1}(\alpha)$ is called α-*quantile*.

Note: 0.5-quantile $F^{-1}(0.5)$ is called *median*.

Definition

Critical value of the standard normal distribution is defined as $z(\alpha) = \Phi^{-1}(1 - \alpha)$.

Note: If $Z \sim \mathrm{N}(0, 1)$, then

- $\Pr\{Z > z(\alpha)\} = \alpha$
- $\Pr\{|Z| > z(\alpha/2)\} = \alpha$

Computing confidence interval

Example exercise

Given the confidence level 99 %, find the confidence interval for the men's height population mean based on the given observation. The given assumptions are:

- the men's height population variance $\sigma^2 = 100$
- the observed sample mean $\bar{x} = 181.04$
- the observed sample size $n = 10$

The confidence interval contains all possible values of the population mean that could not be rejected if hypothesized at the given confidence level.

Idea of the solution
The confidence interval is a symmetric interval around the observed sample mean $\bar{x}$. Hence, we are looking for the confidence interval radius r so that the null hypothesis $H_0 : \mu = \mu_0$ is rejected if and only if $|\bar{x} - \mu_0| > r$.

Then the confidence interval will be given by $(\bar{x} - r, \ \bar{x} + r)$.

Assuming that $H_0 : \mu = \mu_0$ is valid, we have $\bar{X} - \mu_0 \sim \mathrm{N}(0, \sigma^2/n)$.

Since the significance level α, which is the probability of incorrectly rejecting the null hypothesis, is given, we should choose the confidence interval radius r so that

$$\Pr\{|\bar{X} - \mu_0| > r\} = \alpha,$$

and after standardization

$$\Pr\left\{\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \frac{r}{\sigma/\sqrt{n}}\right\} = \alpha,$$

which means that $\dfrac{r}{\sigma/\sqrt{n}} = z(\alpha/2)$, and thus $r = \dfrac{\sigma}{\sqrt{n}}\, z(\alpha/2)$.

Therefore the confidence interval for the population mean μ is

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}}\, z(\alpha/2),\ \bar{x} + \frac{\sigma}{\sqrt{n}}\, z(\alpha/2)\right),$$

and $\Pr\left\{\bar{X} - \dfrac{\sigma}{\sqrt{n}}\, z(\alpha/2) < \mu < \bar{X} + \dfrac{\sigma}{\sqrt{n}}\, z(\alpha/2)\right\} = 1 - \alpha$.

Example 6 – testing the classifier accuracy mean

You have two models, **A** and **B**, and for each of them 10 results – accuracies obtained from 10-fold cross-validation experiment.

```
> A.acc
 [1] 0.853 0.859 0.863 0.871 0.832 0.848 0.863 0.860 0.850 0.849
> mean(A.acc)
 [1] 0.8548

> B.acc
 [1] 0.851 0.848 0.862 0.871 0.835 0.836 0.860 0.859 0.841 0.843
> mean(B.acc)
 [1] 0.8506
```

The average accuracy of **A** is 85.48 %, while the average accuracy of **B** is only 85.06 %.

Question: Is model A *really* better than model B?

Using t-distribution as the principle of t-test

What if you do NOT know the variance?

When we get k different results from the cross-validation experiment, we can assume that the values are (approximately) normally distributed. Then we use t-test.
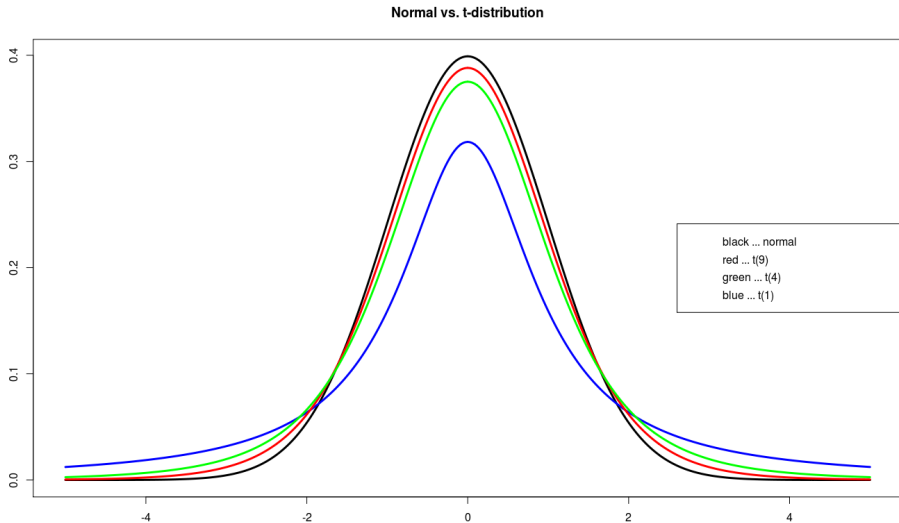
Theorem

If $x_1, \ldots, x_n$ is a random sample of size n selected from a normally distributed population, then

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n} \ \sim \ t_{n-1},$$

where n is the sample size, $\bar{X}$ is the sample mean, S is the sample standard deviation, μ is the population mean, T is called t-statistic, and t_{n-1} stands for t-distribution with $n - 1$ degrees of freedom.

Normal vs. t-distribution

black ... normal
red ... t(9)
green ... t(4)
blue ... t(1)

Using t-test – practical procedure

First, compute **t-value** $T = \dfrac{\bar{X} - \mu}{S}\sqrt{n}$.

Then compare the t-value with the critical value $t_k(\alpha)$.

Definition

Critical value $t_k(\alpha)$ of the t-distribution t_k is defined by the equation
$\Pr\{|T| \geq t_k(\alpha)\} = \alpha$, where α is the test significance level.

Therefore $\Pr\{-t_{n-1}(\alpha) < \dfrac{\bar{X} - \mu}{S}\sqrt{n} < t_{n-1}(\alpha)\} = 1 - \alpha$

Note: Critical value corresponds to a given significance level and determines the boundary between those samples resulting in a test statistic that leads to rejecting the null hypothesis and those that lead to a decision not to reject the null hypothesis. If the calculated value from the statistical test is greater than the critical value, then the null hypothesis is rejected in favour of the alternative hypothesis, and vice versa.

Confidence interval for the mean μ using t-test

- If $\bar{x}$ is the sample mean of a sample of the size n randomly chosen from a normally distributed population and α is a significance level, then **confidence interval** for the population mean μ is

$$\left(\bar{x} - \frac{S}{\sqrt{n}} \, t_{n-1}(\alpha), \ \bar{x} + \frac{S}{\sqrt{n}} \, t_{n-1}(\alpha) \right)$$

- The probability that the (true) population mean μ lies inside the confidence interval is equal to $1 - \alpha$, which is called **confidence level**.

$$\Pr\left\{ \bar{X} - \frac{S}{\sqrt{n}} \, t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S}{\sqrt{n}} \, t_{n-1}(\alpha) \right\} = 1 - \alpha$$

Example 6 – checking the confidence interval

To test if the difference between the models **A** and **B** is **statistically significant** we will check **confidence intervals** for the mean accuracy.

```
### Could the true mean of A accuracy be 0.8506?
> t.test(A.acc, mu=0.8506)
    One Sample t-test

data:  A.acc
t = 1.2195, df = 9, p-value = 0.2537
alternative hypothesis: true mean is not equal to 0.8506
95 percent confidence interval:
 0.8470088 0.8625912
sample estimates:
mean of x
   0.8548
```

We cannot reject the null hypothesis that the mean of A accuracy is equal to 0.8506. The t-test says that the true mean of A accuracy could be between 0.847 and 0.863, which is the confidence interval at the significance level $\alpha = 5\%$.

Using t-tests

When you get k different results from the cross-validation experiment, what can you conclude then?

❶ **One Sample t-test**
– to test if the mean of a (normally distributed) population is equal to a given value

❷ **Paired Two-Sample t-test**
– to test if the difference of the means of two populations is equal to zero (or to another given value)
– assuming that the given samples contain paired individuals

Example 7 – two-sample t-test
Using the same input data as in Example 6

```
### Could the true mean of the difference be equal to zero?
> t.test(A.acc, B.acc)
    Welch Two Sample t-test

data:  A.acc and B.acc
t = 0.8157, df = 17.803, p-value = 0.4254
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.006625999  0.015025999
sample estimates:
mean of x mean of y
   0.8548    0.8506
```

We cannot reject the null hypothesis that the mean of the difference between A accuracy and B accuracy is equal to 0.

Because the t-test says that the true mean of the difference could be between -0.0066 and 0.015, which is the confidence interval at the significance level $\alpha = 5\%$.