Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Exercise Sheet 8 - 10.12.2018	due: 17.12.2018, 14:15

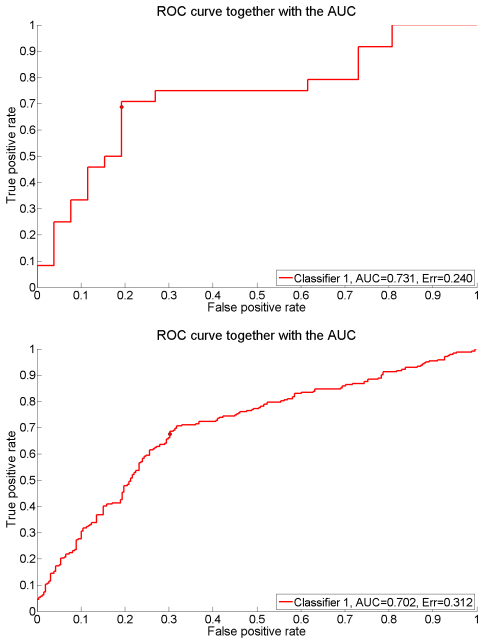
Exercise 15 - ROC and AUC

You are reviewer for an international conference on machine learning. Among the submissions are two independent papers, each proposing a solution to the important problem of automatic classification of hand-crafted contemporary art pottery. As you are an expert in this emerging new field, you are asked to write a review of the two papers. In both papers, the quality of the proposed classifier is evaluated by plotting the ROC curves. Author 1 reports an AUC of 0.73, Author 2 reports an AUC of 0.70.

Both authors made the results of their classifier available as supplementary material. However, the authors used different datasets to evaluate their methods. You can find the dataset Y_1 (the true labels) and the prediction f_1 of the first author as well as the dataset Y_2 and the prediction f_2 of the second author in the file `auc.mat`.

a. **(4 Points)** Inspect the datasets and the predictions made by the two methods. Reproduce the values of the AUC reported by the two authors by means of the function `PlotROC`. Perform 1000 runs of a random classifier on each dataset and compare the ROC curves and the AUC values obtained by the two methods with the results of the random classifiers. Report your results. Which classifier do you think works better? Justify your decision.

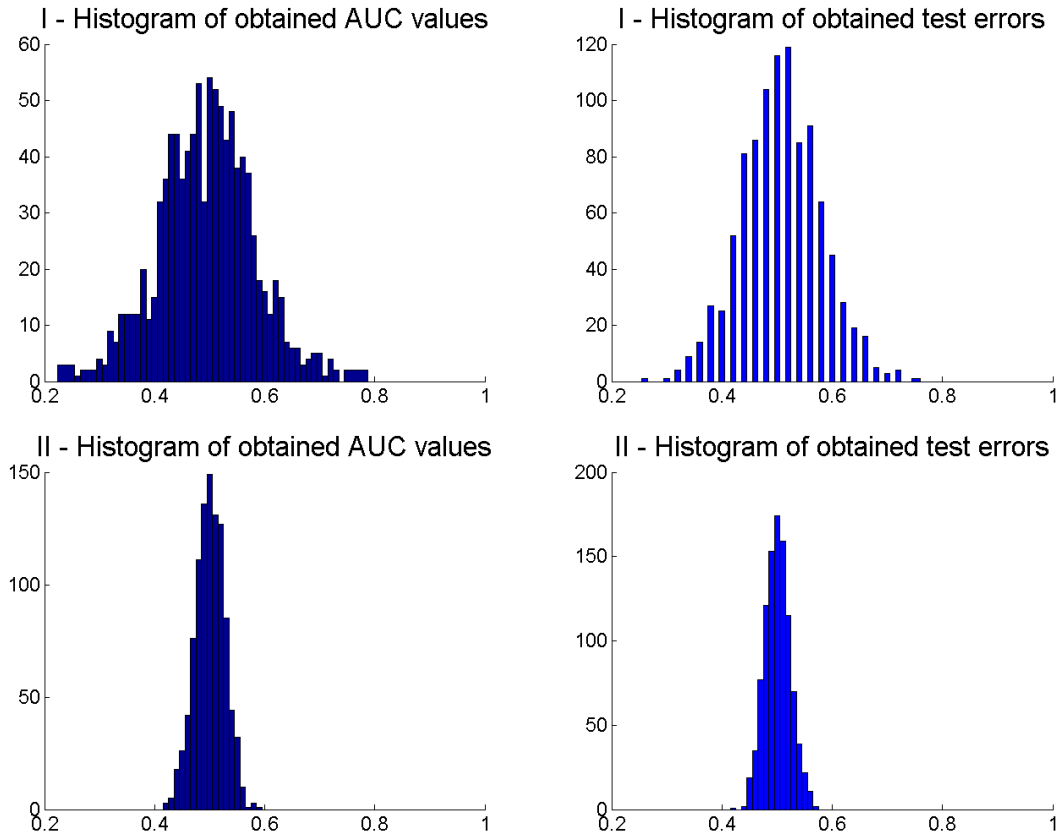
Solution: We compute the ROC curves of the classifiers proposed by Author 1 (left) and Author 2 (right):



Indeed, as reported by the authors, for the first classifier we obtain an AUC of 0.73 and for the second classifier we obtain an AUC of 0.70. At first sight, this suggests that the first classifier has a better performance than the second classifer. However, we observe that the sizes of the testsets used by the two authors are significantly different.

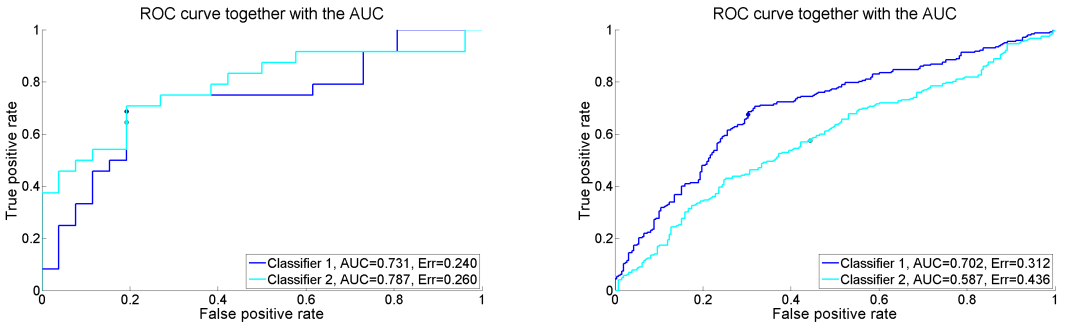
We perform the following experiment: Assume we have a random classifier which just draws random values x from a Gaussian distribution and assigns the labels based on the sign of x. Thus in Matlab we simply compute `f=randn(50,1)` resp. `f=randn(500,1)` to obtain predictions for the

two datasets. We apply our random classifier 1000 times on each dataset and compute the ROC curve as well as the AUC for each run. The following figure shows the AUC values obtained in the 1000 runs, for the first dataset (left) and the second dataset (right).



We observe that the AUC values obtained by a random classifier have much higher variance around 0.5 in the first case (on the smaller dataset) than in the second case. The maximum AUC value obtained in the 1000 runs of the random classifier was 0.787 in the first case, and 0.587 in the second case.

This means on the first dataset, even with a random classifier we can achieve a ROC curve which is better as the one reported by the first author. This is illustrated in the next figure, where we compare the ROC curves of the two classifiers (Classifier 1) with the best result from the 1000 runs of the random classifier (Classifier 2).



In other words, the result obtained by the first classifier is not significantly better than random classification, whereas in the second case, the AUC value reported by the author is much better than the one obtained by a random classifier. This suggests that the second classifier has superior performance, or at least the experimental evidence given by the first author is not sufficient to demonstrate otherwise. In our review, we would ask the first author to provide further experimental results (larger testset) which show the quality of his classifier. We will later on discuss

confidence bands for test error more rigorously which then replace these heuristic arguments. This exercise just illustrates the underlying effects of having a too small test sample.

Exercise 16 - Statistical Tests

We are in the setting of binary classification $\mathcal{Y} = \{-1, 1\}$. We want to develop statistical tests to evaluate if there is any dependency between the features X and the labels Y.

a. **(1 Point)** Prove the following statement:

$$X \text{ and } Y \text{ are independent if and only if } p(x|y=1) = p(x|y=-1)$$

Thus we can reduce the question if X and Y are independent to the question if the class conditional distributions are equal.

b. **(4 Points)** We only have one feature, that is $x \in \mathbb{R}$. Suppose that $p(x|y=1) \sim \mathcal{N}(\mu_+, \sigma^2)$ and $p(x|y=-1) \sim \mathcal{N}(\mu_-, \sigma^2)$ with known variance σ^2. Develop a modified version of the Gauss-test to test if $p(x|y=1) = p(x|y=-1)$ (null-hypothesis). That requires to develop a test-statistic, derivation of the distribution of the test-statistic under the null-hypothesis, definition of a rejection region for a given significance level α, computation of the p-value. The number of samples from positive and negative class can be different.

c. **(2 Points)** Write a Matlab Function

$$\texttt{pval = DoPermutationTest(X,Y)},$$

which given the data $(x_i, y_i)_{i=1}^n$ does a permutation test for testing whether $p(x|y=1)$ is equal to $p(x|y=-1)$ using the Fisher-score as the test-statistic (1000 permutations drawn uniform at random) and returns the p-value.

d. **(3 Points)** Compare the two developed tests, where

- first with data sampled from $p(x|y=1) \sim \mathcal{N}(\mu, 1)$ and $p(x|y=-1) \sim \mathcal{N}(0, 1)$ (100 points from $Y=1$ and 200 points from $Y=-1$) and vary $\mu = \texttt{0:0.1:0.5}$ and for each μ repeat the experiment 100 times. How often do you reject H_0 for the significance level 0.05 for both tests for each μ.
- second with data sampled from $p(x|y=1) \sim \texttt{Uni}[\mu, 1+\mu]$ and $p(x|y=-1) \sim \texttt{Uni}[0, 1]$, where `Uni` denotes the uniform distribution (100 points from $Y=1$ and 200 points from $Y=-1$) and vary $\mu = \texttt{0:0.03:0.15}$ and for each μ repeat the experiment 100 times. How often do you reject H_0 for the significance level 0.05 for both tests for each μ.

Discuss the result of the two tests. Plot the number of rejections of H_0 as a function of μ.

Send the code for c) and the plots of d).

Hints:

- In order to get random permutations use the function `randperm(n)` which returns a permutation of the number $1, \ldots, n$.
- For b) you may use that if $X \sim \mathcal{N}(\mu, \Sigma)$ where $X \in \mathbb{R}^n$, then for $A \in \mathbb{R}^{m \times n}$ the variable AX is distributed as $\mathcal{N}(A\mu, A\Sigma A^T)$.

Solution:

a. X and Y are independent, if $p(x, y) = p(x)\mathrm{P}(y)$, $\forall x \in \mathcal{X}, y \in \mathcal{Y}$. In general one has, $p(x, y) = p(x \,|\, y)\mathrm{P}(y)$. Thus X and Y are independent, if and only if $p(x \,|\, y) = p(x)$ for all $y \in \mathcal{Y}$, which implies $p(x \,|\, y=1) = p(x \,|\, y=-1)$. On the other hand if $p(x \,|\, y=1) = p(x \,|\, y=-1)$, then $p(x) = p(x \,|\, y=1)\mathrm{P}(y=1) + p(x \,|\, y=-1)\mathrm{P}(y=-1) = p(x \,|\, y=1)$.

b. Under the null hypothesis, the means of both samples have to agree. Thus in analogy to the original Gauss-test we consider the test statistics,

$$T'(X) = \frac{1}{n_+} \sum_{\{i \,|\, Y_i=1\}}^{n} X_i - \frac{1}{n_-} \sum_{\{j \,|\, Y_j=-1\}} X_j,$$

where $n_+ = |\{i \,|\, Y_i = 1\}|$ and $n_- = |\{j \,|\, Y_j = -1\}|$ (number of samples from positive and negative class). Using the hint, we can write $T'(X) = \langle a, X \rangle$ where $a_i = \begin{cases} \frac{1}{n_+} & \text{if } Y_i = 1, \\ -\frac{1}{n_-} & \text{if } Y_i = -1. \end{cases}$. According to the hint we get $T'(X) \sim \mathcal{N}(\mu_+ - \mu_-, \sigma^2 \|a\|_2^2)$, where $\|a\|_2^2 = \frac{1}{n_+} + \frac{1}{n_-}$. Then we do the usual standardization procedure, in order that the test statistic has unit variance,

$$T(X) = \frac{1}{\sqrt{\sigma^2 \|a\|^2}} T'(X) = \sqrt{\frac{n_+ n_-}{\sigma^2 (n_+ + n_-)}} T'(X).$$

Then $T(X) \sim \mathcal{N}(\sqrt{\frac{n_+ n_-}{\sigma^2 (n_+ + n_-)}}(\mu_+ - \mu_-), 1)$. Moreover, under the null hypothesis $T(X) \sim \mathcal{N}(0, 1)$. Thus we can use the same definition of rejection region as for the usual Gauss test,

$$B(\alpha) = (-\infty, q_{\frac{\alpha}{2}}] \cup [q_{1-\frac{\alpha}{2}}, \infty)$$

The p-value can be computed as

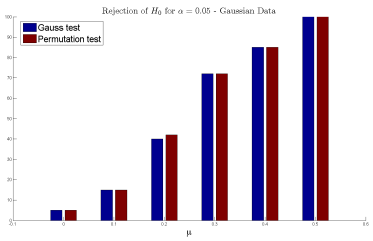
$$\text{p-value} = \mathrm{P}(|T(X)| > T_{\text{obs}}),$$

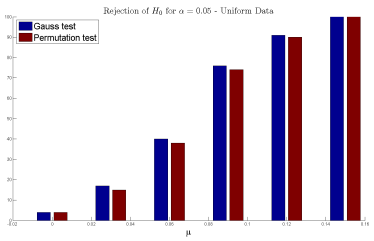
where T_{obs} is the computed value of the test statistic on the sample and q_γ are the quantiles of $\mathcal{N}(0, 1)$.

c.
```
function [PVal,FScorePerm,OrgScore]=DoPermutationTest(Xtrain,Ytrain)
% does a permutation test in order to determine if the sample of both
% classes comes from the same distribution
% input:  Xtrain,Ytrain, Xtrain has to be real-valued, Ytrain must only have
% two different values
% output:  p-value, the scores from the permutation test, the original score

classes = unique(Ytrain);
end

% compute the Fisher Score for the original data
OrgScore=FisherScore(Xtrain,Ytrain,classes);


NUM=1000; NumTrain=length(Ytrain);
for i=1:NUM
end
%curtime=clock;
%rand('state',ceil(100*curtime(6)));
idx = randperm(NumTrain);
YtrainPerm=Ytrain(idx);
% compute the Fisher score for the permuted data
FScorePerm(i)=FisherScore(Xtrain,YtrainPerm,classes);
end
PVal = 1/NUM*sum(FScorePerm > OrgScore);


function FScore=FisherScore(Xtrain,Ytrain,classes)
ixpos = find(Ytrain==classes(1));
```

```
ixneg = find(Ytrain==classes(2));
vecpos=Xtrain(ixpos);
vecneg=Xtrain(ixneg);

mupos = mean(vecpos);
muneg = mean(vecneg);

sigma2pos = mean(vecpos.^2)-mupos^2; %var(Xtrain(ixpos));
sigma2neg = mean(vecneg.^2)-muneg^2; %var(Xtrain(ixneg));

FScore=(mupos-muneg)^2/(sigma2pos+sigma2neg);
```

d. We plot the number of rejections of the null hypothesis in dependence of the parameter μ. On the left we see the result of the Gaussian data. The traditional Gauss-test works slightly better than the permutation test, which basically uses a test statistic

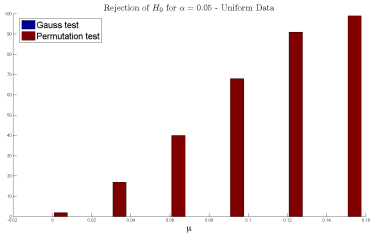




similar to the one of the Gauss test. This is to be expected as the Gauss-test is constructed to have high power and the data meets the assumptions of the test.

For the uniform data again Gauss test and permutation test have the same behavior. Even though the Gauss test is basically not applicable we have already that many samples that the central limit theorem kicks in and thus the distribution of the sums is well approximated by Gaussians. Note that one had to use the correct variance estimate for the uniform distribution, if $X \sim \text{Uni}(0,1)$, then $\text{Var}(X) = \frac{1}{12} = \sigma^2$. The downside of the permutation test is the long time it takes to compute the quantities. Moreover, we would have to sample even more permutations for the smaller p-values.

If you used the Gauss-test with the wrong variance estimate of $\sigma^2 = 1$ you have got this result



which shows the (obvious) fact that the variance has a huge impact on the test result.