

Praktikum zu Moderne Methoden der Datenanalyse

Exercise 5: Hypotheses Testing and Classification

“Is this a new discovery or just a statistical fluctuation?” Statistics offers some methods to give a quantitative answer.

But these methods should not be used blindly. In particular one should know exactly what the obtained numbers mean and what they don't mean.

- **Exercise 5.1:**

The following table shows the number of winners in a horse race for different track numbers:

track	1	2	3	4	5	6	7	8
#winners	29	19	18	25	17	10	15	11

Test the hypothesis that the track number has *no* influence on the chance to win with a χ^2 test. Define a confidence level, e.g. of 95 % or 99 %, *before* you do the test.

- **Exercise 5.2:**

In a counting experiment 5 events are observed while $\mu_B = 1.8$ background events are expected. Is this a significant ($=3\sigma$) excess? Calculate the probability of observing 5 or more events when the expectation value is 1.8 using Poisson statistics.

- **Exercise 5.3:**

Determine an upper limit μ_S^{max} for the number of signal events at a 95 % confidence level for the experiment from exercise 5.2. Such a limit is defined by the expected number of signal events μ_S^{max} where the probability of measuring the observed number of events or less reaches 5 % assuming a Poisson with mean $\mu_B + \mu_S^{max}$. Use an iterative method to determine μ_S^{max} .

- **Exercise 5.4:**

Verify the limit determined in exercise 5.3 with toy experiments. In each toy experiment generate a random number according to a Poisson distribution with a mean value of $\mu_B + \mu_S^{max}$. Then count the number of experiments in which this random number is less or equal 5. The fraction of these events should be 5%.

- **Exercise 5.5:**

In an experiment two types of events, signal (S) and background (B), are observed. The measured quantity x of signal events follows a Gaussian distribution \mathcal{N} with a mean of 1 and a sigma of 1: $x_i^S \in \mathcal{N}(1, 1)$. The distribution of background events is given by a Gaussian distribution with mean of 0 and a sigma of 1: $x_i^B \in \mathcal{N}(0, 1)$.

Simulate a large number of signal events and the same number of background events and plot their x distribution. One can classify these events as signal or background candidates based on a cut at a chosen value x_c . Plot the significance α , the power β (as defined in the lecture), the signal efficiency ϵ , the signal purity p and the fraction of wrong decisions as a function of the cut value x_c . Also plot the purity versus the efficiency. Repeat the simulation and the plots with 10 times more background.

- **Exercise 5.6:**

The experiment described in exercise 5.5 is extended by the measurement of an additional variable. Each measured event is now a pair $\vec{x} = (x_1, x_2)$.

Simulate n signal events $\vec{x}_i^S \in \mathcal{N}(1, 1) \times \mathcal{N}(1, 1)$ and n background events $\vec{x}_i^B \in \mathcal{N}(0, 1) \times \mathcal{N}(0, 1)$. Apply the Fisher discriminant method to separate both classes of events. Plot the Fisher discriminant value t for signal and background and choose a cut value. Make a two dimensional scatter plot of the signal and background events in different colors together with a line indicating the chosen cut.

- **Exercise 5.7:**

In a further experiment the background distribution is changed with respect to exercise 5.6. Simulate n signal events $\vec{x}_i^S \in \mathcal{N}(1, 1) \times \mathcal{N}(1, 1)$, $n/2$ background events $\vec{x}_i^{B1} \in \mathcal{N}(0, 1) \times \mathcal{N}(0, 1)$, $n/2$ background events $\vec{x}_i^{B2} \in \mathcal{N}(2, 1) \times \mathcal{N}(2, 1)$ and write the measurement pairs together with a flag for the type of event to a ntuple. You may skip this part of the exercise and take the ntuple file `data.root` at `/home/staff/tkuhr/Praktikum/Exercise5`.

Train a neural network to distinguish between both classes. Use the root class `TMultiLayerPerceptron` which is available in root after loading the appropriate library with `gSystem->Load('libMLP.so')`.

Plot the net output value o for signal and background. Make a two dimensional scatter plot of the signal and background events and add a contour plot of the neural net output to it. The contour can be drawn by using a `TF2` for the network output. Why is the Fisher discriminant method not suitable for a classification in this case?

Fisher discriminant method:

Given is a set of events $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$ of class 1 and class 2, respectively. The covariance matrix of class j is estimated by

$$V_{km}^{(j)} = \frac{1}{N} \sum_N (x_m^{(j)} - \bar{x}_m^{(j)})(x_k^{(j)} - \bar{x}_k^{(j)})$$

with \bar{x} being the mean value and N the number of events. Then the Fisher discriminant value for a measurement \vec{x} is defined as:

$$t = \sum_{i=1}^n f_i x_i - \frac{1}{2} \sum_{i=1}^n f_i (\bar{x}_i^{(1)} + \bar{x}_i^{(2)})$$

with

$$f_i = \sum_k (V^{-1})_{ik} (\bar{x}_k^{(1)} - \bar{x}_k^{(2)}) \quad \text{and} \quad V_{mk} = \frac{1}{2} (V_{mk}^{(1)} + V_{mk}^{(2)})$$

Here n is the dimension of the measurement vector.