Statistical Methods Exercises 5 Autumn 2020

Return your solutions by 12.00 Finnish time on Thursday 15.10.2020 to Moodle course page: `https://moodle.helsinki.fi/course/view.php?id=30207`

1. **Fisher discriminant for event classification.** In a physics analysis, the measured events have to be classify into two different classes, class a and b. Each event is described by two variables x_1 and x_2. Events from these two classes can be classified with a Fisher discriminant. This method provides weighting factors c_1 and c_2, so that a test statistic x_{new} built as a linear sum of the weighting factors and the variable values, i.e. $x_{new} = c_1 * x_1 + c_2 * x_2$, allows best separation between the two classes. The weighting factors can be defined as a vector $\mathbf{c} = (c_1, c_2)$:

$$\mathbf{c} \;\propto\; (\mathbf{V_a} + \mathbf{V_b})^{-1}(\mu_\mathbf{a} - \mu_\mathbf{b})$$

where $\mathbf{V_a}$ and $\mathbf{V_b}$ are the 2×2 covariance matrices for class a and b and $\mu_\mathbf{a}$ and $\mu_\mathbf{b}$ are the 2×1 expectation value vectors for class a and b. To build the new variable, you can use two training data samples, `train_sample20_class_a.txt` and `train_sample20_class_b.txt` that contain the x_1 (1. column) and x_2 (2. column) values of events only from class a and only from class b, respectively. The data sample files can be found in the moodle course area with the exercise paper.

(i) Draw the data points from the two training samples as different coloured dots in the x_1x_2-plane (x_1 first variable, x_2 second variable).

(ii) Calculate the vector $\mathbf{c}$. Use the mean values and (co)variances of the training data as estimates for the expectation values and covariances. Plot the x_{new} value for the events of class a and b in a histogram.

(iii) What is the requirement on x_{new} to have a 96 % rejection of events from class b? What is the corresponding acceptance for class a events? Plot the data points of the two test samples in the x_1x_2-plane using different colours. Add the x_{new} requirement giving 96 % rejection of class b into the same plot. *Exercise gives max 9 points instead of usual 6.*

2. **Kolmogorov-Smirnov test**. During the occurrance of the 1987 Supernova (SN1987A), interactions from cosmic neutrinos were seen in two large underground experiments: IMB in USA and Kamiokande (KAM) in Japan. Most of the cosmic neutrino events are expected to produce recoil particles with an angular distribution:

$$dN/d\cos\theta \propto 1 + \gamma\cos\theta\,,$$

where $\gamma \approx 0.1$, w.r.t. direction towards the source, here assumed to be SN1987A. The recoil angles θ of the 8 IMB events were 80, 44, 56, 65, 33, 52, 42 and 104 degree (= "data1") and of the 12 KAM events 18, 40, 108, 70, 135, 68, 32, 30, 38, 122, 49 and 91 degree (="data2"). The uncertainty in the θ determination can be neglected.

(i) Test using the Kolmogorov-Smirnov test whether the two results are compatible (i.e. they could orginate from the same distribution). Calculate first the Kolmogorov-Smirnov distance (KS-distance) for data1 vs. data2. The KS-distance is given by maximum difference

$$\mathrm{diff}_{max} \quad = \quad max|F_{\mathrm{data1}}(x) - F_{\mathrm{data2}}(x)|,$$

where $F_{\mathrm{data1}}(x)$ and $F_{\mathrm{data2}}(x)$ are the values of the normalized cumulative distribution function for data1 and data2 at a specific value of x.

(ii) What is the corresponding P-value, P_{KS} Hint: use e.g. `http://www.mathworks.com/matlabcentral/fileexchange/4369-kolmogorov-distributionfunctions` to calculate P_{KS}. The P-value is then estimated as:

$$P_{KS} \quad = \quad 1 - F_{\mathrm{K}}(\mathrm{diff}_{max}\sqrt{\frac{N_{\mathrm{event,data1}} \cdot N_{\mathrm{event,data2}}}{N_{\mathrm{event,data1}} + N_{\mathrm{event,data2}}}}),$$

where $F_{\mathrm{K}}(x)$ is the cumulative distribution function for the Kolmogorov distribution. Are the two angle distribution compatible to originate from same distribution i.e. to originate from the same source?

(iii) Test using the KS test whether the experimental data is compatible with the expected angular distribution (when all data is treated as one sample). Calculate KS-distance for data1+data2 vs. expectation.

(iv) What is the corresponding P-value for the comparison data1+data2 vs expectation? NB! $\sqrt{N_{\mathrm{event,data1}} \cdot N_{\mathrm{event,data2}}/(N_{\mathrm{event,data1}} + N_{\mathrm{event,data2}})}$ should here be replaced by $\sqrt{N_{\mathrm{event,data1}} + N_{\mathrm{event,data2}}}$. Is the combuned experimental distribution compatible with the expected distribution?