

Praktikum zu Moderne Methoden der Datenanalyse

Exercise 1: Priors and Monte Carlo

Bayes Theorem

$$P(A|B) = P(B|A) \cdot \frac{P(A)}{P(B)}$$

shows that the conditional probability $P(A|B)$ of observing an event A in case an event B has happened depends on the conditional probability $P(B|A)$ and the a priori probabilities (called priors) $P(A)$ and $P(B)$ for the events A and B respectively.

The priors are of crucial importance for the interpretation of measurements. A theory usually predicts a probability $P(B|A)$ to observe a measurement B for a given assumption (theory / parameter set) A . However in general one is interested in the opposite: Given a measurement B what can be deduced about the theory? How “probable” is it that theory A is correct?

- **Exercise 1.1:**

“Should I carry an umbrella or should I risk to get wet?” A possible answer to this question is to look at the weather forecast. But as we all know it is not always reliable.

Let’s assume that if it will rain, the forecast predicts this correctly in 80 % of the cases. If it will not rain, the forecast is assumed to be accurate in 90 % of the cases. In Sun-City the a priori rain probability is only 5 %, in Equal-City it’s 50 % and in Rain-City it’s 95 %. Calculate (on a sheet of paper or in a root macro) the four probabilities that it will (not) rain if (no) rain is predicted for the three cities.

There are two different risks of a wrong decision:

- Carry an umbrella, but it does not rain.
- Don’t carry an umbrella in case it rains.

Which are the three possible strategies and which of them is the optimal one to minimize the risk of a wrong decision in each of the three cities? Calculate and compare the risk for each of the three possible strategies. Determine as well the optimal strategy when the second risk is considered 10 or 100 times more serious.

- **Exercise 1.2:**

“Does it make sense to do a preventive medical check-up for cancer of the intestine? How sure can I be that I have cancer when I get a positive test result?”

Let’s assume the test gives a positive result for 50 % of the patients with cancer. If the patient is healthy the test confirms this in 97 % of the cases. At the age of 75-79 0.2 % of the people have cancer of the intestine. At the age of 30-34 the rate is only 0.002 %. Calculate the probability to have cancer if the test result is positive for the two age ranges.

In case of a positive test result a follow-up examination can be performed which is here assumed to be 100 % accurate. Unfortunately an early discovery of cancer due to this examination does not help most of the patients. Only 20 % of the patients can be healed in addition. On the other hand a follow-up examination causes complications in 0.2 % of the cases.

Given a number of $N = 100\,000$ patients calculate the expected number of patients with positive initial test results, the number of patients with detected cancer after the follow-up examination, the number of additionally healed patients, and the number of patients with complications due to the follow-up examination for each of the two age ranges.

- **Exercise 1.3:**

“Fewer than 1 in 1000 women who are abused by their mates go on to be killed by them” said Alan M. Dershowitz, the advocate of O. J. Simpson. So it is very unlikely that O. J. Simpson killed his wife, isn’t it?

Calculate the probability that he killed his wife using the following assumptions: O. J. Simpson has abused his wife. The probability to be killed by a beating husband is 0.1 %. The average “lifetime” of a marriage is 3 years if the husband abuses and kills his wife. 25 000 murders are committed in the US each year. 12 % of them are committed by the husband of the victim. Half of the victims not killed by a husband are women killed by somebody else. In total 130 000 000 women live in the US. The probability to be killed by somebody else does not depend on the behaviour of the mate.

Hint: Calculate from the number of woman per year who are killed by somebody else than her husband the probability that this happens. Then calculate the probability to be killed by a beating husband and combine both probabilities in order to determine the statistical probability of O. J. Simpsons guilt. Make sure you use in your calculation all information provided in the exercise.

- **Exercise 1.4:**

Verify your calculations for at least one of the previous three exercises experimentally by writing a Monte Carlo. Simulate N weather events, examined patients, abused wives respectively. For each decision (rain, forecast, cancer test result, etc.) generate

a uniformly distributed random numbers between 0 and 1 with `gRandom->Rndm()` (class `TRandom`) and compare it to the corresponding probability given in the above exercises. Make sure that you only use the values given in the text or calculated from the hint in exercise 1.3 in your code. Then count the number of events of different category (rain and no rain predicted, patient healed, etc.). Finally use these numbers to determine the fraction of wrong weather forecasts, murders by a beating husband, etc. and compare these numbers to the calculated probabilities. Repeat the simulation for different values of N .

- **Exercise 1.5:**

How well do the experimentally determined values correspond to the calculated ones? Repeat the simulation many times for a fixed value of N . For each simulated measurement fill the value of at least one selected quantity into a histogram (root class `TH1F`) and investigate the obtained distribution.

Some useful relations about probabilities:

$$\begin{aligned}
 P(A \vee B) &= P(A) + P(B) - P(A \wedge B) \\
 P(A \vee \neg A) &= P(A) + P(\neg A) = 1 \quad (\neg A = \text{not } A) \\
 P(A \wedge B) &= P(A) \cdot P(B|A) = P(B) \cdot P(A|B) \quad \Rightarrow \text{Bayes Theorem} \\
 P(B|A) &= P(B) \quad \text{if } A \text{ and } B \text{ are independent}
 \end{aligned}$$