Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt
Sabyasachi Patjoshi
sabyasachi2k13@gmail.com, martin.vogt@bit.uni-bonn.de

Due: May 13, 10:30 (by the end of the lecture)

Exercise Series 3

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should help solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.
A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

1. A company discovers a novel biomarker diagnosis assay, which has a sensitivity of 95% to detect a certain disease when it is present. The test also yields 5% false-positive results. If 5% of the population has the disease, what is the probability that a person with a positive test result actually has the disease? What is the probability that a person with a negative test result has the disease? Tip: Make use of Bayes theorem.
   (5 points + 1 bonus point)

2. Suppose that we have 2 urns A and B, each containing black and white balls. Urn A contains three times as many black as white balls. Urn B contains two times as many white as black balls. We select one urn randomly and pick 10 balls blindly, replacing each ball after it has been selected (sampling with replacement). The result is: 5 black, 5 white balls. What is the probability that we selected urn A? Tip: use Bayesian reasoning.
   (5 points + 2 bonus points)

3. A biased die is rolled n times. The probabilities for 1, 3, 6 are 0.2 each. The probabilities for the other numbers are also identical to each other.
   a) What are the probabilities for 2, 4, 5? (1 point)
   b) What is the probability to observe 3 times 6, 1 times 5, 2 times 4 and 1 times 3? (2 points)
   c) What is the probability to observe 2 times 1 and 2 times 5 with the outcome of a fifth die role being unknown? Tip: Use marginalization! (3 points)
   d) What is the expected value and variance of the die? (2 points)
   e) A second biased die with the same properties is thrown. What is the covariance of these dies? (2 points)

   (5 bonus points)

4. Implement a R function `multinom`, which takes as arguments
   i) a sample matrix (each row being one sample vector)
   ii) a probability parameter vector **p** for a multinomial distribution
   iii) the number of trials n.

   The function computes and returns:
   - the expectation value
   - the population variance/covariance matrix
   - the sample mean (R functions `apply` and `mean`)
   - the sample covariance matrix (R function `cov`)
   - the value of the probability mass function for each sample (Tip: you can use R function `dmultinom`)

In addition `multinom` should show an appropriate bar diagram visualzing **p**.
Test your function with 50 random drawings from the multinomial distribution in task 3 (R function `rmultinom`) and show your obtained results. How close are the empirical mean, variance and covariances to the theoretically expected results? Which behavior would you expect, if the number of samples is increased and decreased, respectively? Please discuss.

(10 points + 3 bonus points)