Mathematics of Cellular Networks

Prof. Dr. Volkhard Helms
Thorsten Will
Sommer Semester 2014

Saarland University
Chair for Computational Biology

Exercise Sheet 5

Due: Juli 1, 2014 10:15

Send a single PDF attachment with your name(s), your solution(s). Hand in all source code via mail to thorsten.will@bioinformatik.uni-saarland.de.

1 Maximum Likelihood Estimation (60 points)

(a) Many statistical methods assume an independent and identical distribution (i.i.d.) behind the process that generates the observable data.
Explain the meaning of i.i.d. in your own words and elaborate why it is a crucial prior assumption in the context of MLE.
Additionally, state an example for which the assumption of an i.i.d. process is approximately true and an example for which it is clearly not the case. (5+5 points)

(b) Suppose you observed a set of i.i.d. datapoints $x_1, x_2, ..., x_n$ that are normally distributed. Prove that the empirical mean and variance of the samples are the most likely parameters μ and σ for that distribution. (5+5 points)

$$P_{gaussian}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(c) Given a set of i.i.d. samples $x_1, x_2, ..., x_n$, derive the most likely parameter λ for three further popular distributions: Poisson-, exponential- and a log-normal distribution with fixed variance. (5+5+5 points)

$$P_{Poisson}(x|\lambda) = \frac{\lambda^x}{x!} e^{\lambda}$$
$$P_{exp}(x|\lambda) = \lambda e^{-\lambda x}$$
$$P_{log-norm}(x|\lambda) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2}}$$

(d) The file 'samples.txt' from the website contains datapoints generated by one of the distributions in part (c). Determine and report the optimal parameter λ for each of the distributions, then plot a histogram of the data together with the best fit of each one. Decide by visual inspection how the data is most likely distributed. You are allowed to use any software package of your choice to plot the distributions.
Briefly describe how one could decide on one of the distributions without manual inspection. (5+15+5 points)

Hint: Depending on your programming language of choice it may be necessary to sample the two continuous distributions. Use a sufficient number of samples if this is the case.

2 MLE in Bayesian Networks (40 points)

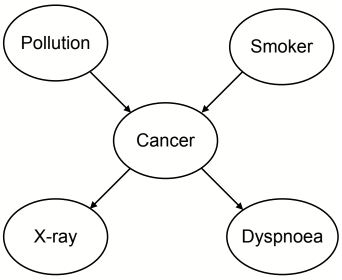

Figure 1: Bayesian network G for lung cancer

Figure 1 shows a simple Bayesian network G that depicts some causal dependencies in the context of lung cancer. It contains the random variables pollution:{low, high}, smoker:{true, false}, cancer:{true, false}, x-ray:{positive, negative} and dyspnoe:{true, false} (breathlessness).

(a) Derive and report the independent local likelihood terms of G by global likelihood decomposition.
What happens if the training data does not contain all assignments of a random variable? (10 points)

(b) Apply MLE on the samples in the file 'cancer_bn.txt' from the website to compute all priors and conditional probabilities for the model as given by G.
Report the parameters in the form of tables. The layout of the file is specified in the first row and should be self-explanatory. (15 points)

(c) Given the state of pollution, smoker, x-ray and dyspnoea for a patient, how can you query if he/she may have cancer?
If a patient lives in a highly polluted area and is a smoker with dyspnoea but there is no sign of cancer in the x-ray what do you tell him/her? (5 points)

(d) A sufficient amount of samples and a well-balanced distribution of the observed assignments of all relevant features are always important in statistical learning.
In Bayesian networks the number of samples that are at least needed to determine each conditional probability increases exponentially with the number of parents of the specific node (complexity of the model corresponds to the number of edges in the graph). Explain the origin of that dependency.
Also, a heavily unbalanced distribution of individual random variables can easily lead to skewed likelihoods in Bayesian networks even if one has many samples. Why? (5+5 points)