Mathematics of Cellular Networks

Prof. Dr. Volkhard Helms
Thorsten Will
Sommer Semester 2014

Saarland University
Chair for Computational Biology

Exercise Sheet 4

Due: June 17, 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 Basic Probability Theory (25 points)

(a) Consider tossing an arbitrary coin (P("head") $= p_H$) until head occurs for the first time. Define the sample space Ω and a random variable $X : \Omega \rightarrow \mathbb{N}^+$ that describes the number of tosses until head occurs. Determine $P(X = 4)$ and $P(X < 4)$ in terms of p_H. (5 points)

(b) Consider two events α and β such that $P(\alpha) = p_\alpha$ and $P(\beta) = p_\beta$. What are the possible maximum and minimum values of $P(\alpha \cup \beta)$ and $P(\alpha \cap \beta)$ if you have no further information? Characterize the situations in which the extreme values occur. (5 points)

(c) Consider the following joint distribution of two random variables X and Y:

X	Y	$P(X,Y)$
0	0	0.1
0	1	0.2
1	0	0.3
1	1	0.4

What are the marginal distributions of X and Y? Are X and Y independent?
Determine expectation and variance of X. (5 points)

(d) The density of random variable X on $\mathbb{R}$ is given by

$$p(x) = \begin{cases} 1/3 & x \in [1,3] \\ 2/3 & x \in [-3,-1] \\ 0 & \text{otherwise} \end{cases}$$

and Y is a boolean random variable whose probability is conditionally dependent given X:

$$P(Y|X=x) = \begin{cases} 1/2 & x \in [1,3] \\ 1/3 & x \in [-3,-1] \\ 0 & \text{otherwise} \end{cases}$$

$$P(\overline{Y}|X=x) = \begin{cases} 1/2 & x \in [1,3] \\ 2/3 & x \in [-3,-1] \\ 0 & \text{otherwise} , \end{cases}$$

where $\overline{Y}$ is an abbreviation for the negation of Y.
Determine the conditional densities $p(x|Y)$ and $p(x|\overline{Y})$. (10 points)

2 Naive Bayes Classifier (50 points)

A naive Bayes classifier is a simple classifier based on the application of Bayes' theorem and the (naive) assumption of independence among all features.
You will implement and evaluate such a classifier on the basis of two artificial datasets that relate 100 SNPs (single nucleotide polymorphisms) with a genetic predisposition to a disease.

The following abbreviations will be used within the remaining exercise:
D : "disease predisposition", $\overline{D} = H$: "healthy", S : state of all SNPs, S_i : variant of individual SNP i.

(a) We want to determine a possible predisposition to a disease given the states of several SNPs. In practice, log-likelihood ratios are used in binary classification:

$$\log \frac{P(D|S)}{P(H|S)}.$$

Derive a term that uses observable probabilities such as $P(S_i|D)$ to calculate the log-likelihood ratio from training data. How does the actual classification work? (15 points)

(b) Shortly discuss: What are the practical advantages of the logarithm and the likelihood ratio within this framework? State two reasons why this particular type of classifier may perform poorly on a real disease dataset. (10 points)

(c) Use the file 'training.tsv' to determine all necessary priors and likelihoods from part a). The file layout is explained in 'README.txt'. Report $P(D)$ and $P(H)$ as well as the five SNP variants (number, variant and ratio) with the highest likelihood ratios. Examine and comment on the results of the training-phase. Which SNPs seem to be worthwhile targets for further research? (15 points)

(d) Predict the predisposition of the patients in 'test.tsv' with your previously trained model and report the accuracy of the classifier. (10 points)

3 Bayesian networks (25 points)

After a recent increase of your local crime rate (the daily chance of becoming a victim of burglary is still only $P(B) = 0.001$) you invested in a new alarm system. The system is fairly reliable but can be influenced by vibrations. Due to years of coal mining, minor earthquakes are happening with $P(E) = 0.002$ in the Saarland. The manufacturer of the alarm is aware of that and supplied the following incomplete table of conditional alarm probabilities:

		$P(A\|B,E)$
$\overline{B}$	E	0.29
B	$\overline{E}$	0.94
B	E	0.95

You asked your neighbors if they could call you if they hear the alarm. Peter, a student that likes to hear loud music all the day, will miss to notice 30% of the actual alarm signals but will rarely warn you mistakenly when there was no alarm ($P(PC|\overline{A}) = 0.01$, PC:"Peter calls"). On the contrary, your other neighbor Hans is overcautious and will report 5 times more false alarms than Peter, but will miss only every tenth real alarm.

(a) Draw the Bayesian network with all boolean random variables as described, complete all probability distributions as far as possible. By using the probability that both called you for an actual alarm but neither a burglary nor an earthquake happened ($P(HC, PC, A, \overline{B}, \overline{E}) = 0.00062$) you should be able to fill an important gap. (10 points)

(b) How likely is it that Hans calls ($P(HC)$)? How likely is it that there is a burglary given that Hans calls ($P(B|HC)$)? How likely is it that there is a burglary given that Hans and Peter call ($P(B|HC, PC)$)? Concisely explain how you obtained the results. (10 points)

(c) Bayesian networks are directed acyclic graphs and thus can be sorted topologically. This means there is a linear ordering of the nodes (imagine giving each node a number) where an ancestor always comes before its descendants in the ordering (ancestor has lower number). Give such an ordering for your network. How can this ordering be worthwhile to reduce the runtime of certain probability queries? Briefly communicate the main idea. Can you imagine other problems where topological sorting may be useful in computational biology? (5 points)

Hints to c):
How could one break down the effort as in part b), what visualization may be useful for regulatory networks?