Probabilistic Graphical Models

Solution to HWK#4
Assigned Saturday, April 11, 2020
Due: Thursday, April 30, 2020 (7 AM EST)

You may install and use "+brml" package

Problem 1 (Conditional Likelihood training)

Consider a situation in which we partition observable variables into disjoint sets $\boldsymbol{x}$ and $\boldsymbol{y}$ and that we want to find the parameters that maximize the conditional likelihood,

$$CL(\theta) = \frac{1}{N}\sum_{n=1}^{N} \log P(y^n \mid x^n, \theta),$$

for a set of training data $\{(\boldsymbol{x^n}, \boldsymbol{y^n})$, n = 1, . . . , N$\}$. All data is assumed generated from the same distribution $P(x, y \mid \theta^0) = P(y \mid x, \theta^0)P(x \mid \theta^0)$ for some unknown underlying parameter θ^0 . In the limit of a large amount of i.i.d. training data, does CL(θ) have an optimum at θ^0 ?

Solution:

Following the standard maximum likelihood example, one can write

$$CL(\theta) \overset{N\to\infty}{=} \left\langle \log P(y \mid x, \theta) \right\rangle_{P(y\mid x, \theta^0)}$$

We assume for concreteness that the variables are continuous (the same holds for discrete variables on replacing integration with summation) and rewrite the above as

$$CL(\theta) = \int p(x|\theta^0) \int p(y|x, \theta^0) \log p(y|x, \theta)$$

We can write this as

$$CL(\theta) = -\int p(x|\theta^0)\mathrm{KL}\big(p(y|x,\theta^0)|p(y|x,\theta)\big) + \mathsf{const.}$$

When $\theta = \theta^0$ the KL term is zero, and $CL(\theta)$ is maximal. Hence $\theta = \theta^0$ corresponds to an optimum of the conditional likelihood. Howeve this does not mean that one can necessarily learn all parameters. For example if a distribution is parameterised as

$$p(x, y|\theta) = p(y|x, \theta_1)p(x|\theta_2)$$

then the conditional likelihood of $p(y|x, \theta)$ is independent of θ_2.

Problem 2

A local supermarket specializing in breakfast cereals decides to analyze the buying patterns of its customers. They make a small survey asking 6 randomly chosen people their age (older or younger than 60 years) and which of the breakfast cereals (Cornflakes, Frosties, Sugar Pu↵s, Branflakes) they like. Each respondent provides a vector with entries 1 or 0 corresponding to whether they like or dislike the cereal. Thus a respondent with (1101) would like Cornflakes, Frosties and Branflakes, but not Sugar Pu↵s. The older than 60 years respondents provide the following data (1000), (1001), (1111), (0001). The younger than 60 years old respondents responded (0110), (1110). A novel customer comes into the supermarket and says she only likes Frosties and Sugar Pu↵s. Using naive Bayes trained with maximum likelihood, what is the probability that she is younger than 60?

Solution:

Looking at the data, the estimates using maximum likelihood are

$$P(C=1 \mid Young)=0.5,\ P(F=1 \mid Young)=1,\ P(SP=1 \mid Young)=1,\ P(B=1 \mid Young)=0$$

and

$$P(C=1 \mid OlD)=0.75,\quad P(F=1 \mid OlD)=0.25,\quad P(Sp=1 \mid OlD)=0.25,$$

$$P(B=1 \mid OlD)=0.75$$

and $P(Young) = 2/6$ and $P(OlD) = 4/6$. Plugging this into Bayes formula, we get

$$P(Young \mid C=0, F=1, SP=1, B=0) \ \alpha\ 0.5*1*1*(1/6)$$

$$P(Old \mid C=0, F=1, SP=1, B=0) \ \alpha\ 0.25*0.25*0.25*0.25*(4/6)$$

Using the fact that these probabilities sum to 1, this gives

$$p(Young \mid C=0,\ F=1,\ SP=1,\ B=0) = 64/65$$

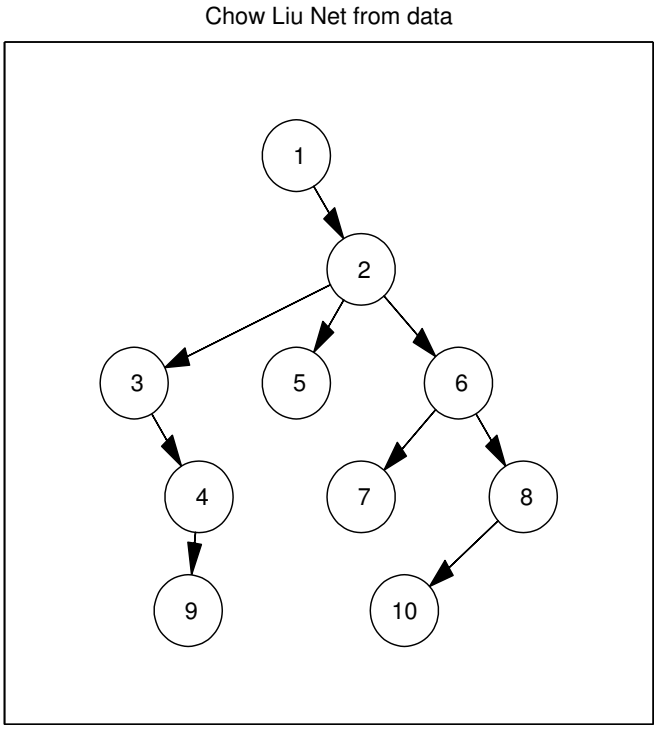
Problem 3

Write a MATLAB routine A = ChowLiu(X) where X is a D by N data matrix containing a multivariate data point on each column that returns a Chow-Liu maximum likelihood tree for X. The tree structure is to be returned in the sparse matrix A. You may find the routine spantree.m in "+brml" package useful. The file ChowLiuData.mat contains a data matrix for 10 variables. Use your routine to find the maximum likelihood Chow Liu tree, and draw a picture of the resulting DAG with edges oriented away from variable 1.

Solution:

See the code "ChowLiu.m" from the solution folder and run drawNet(ChowLiu(X)) from the +BRML package.

See also demoChowLiu.m for a demo from +BRML package.



Problem 4

A Naive Bayes Classifier for binary attributes $x_i \in \{0, 1\}$ is parameterized by $\theta_i^1 = p(x_i = 1 | \text{class} = 1)$, and $\theta_i^0 = p(x_i = 1 | \text{class} = 0)$. Furthermore, we have $p_1 = p(\text{class} = 1)$ and $p_0 = p(\text{class} = 0)$. Show that the decision to classify a data point x as class 1 holds if $w^T x + b > 0$ for some $\boldsymbol{w}$ and b, and state explicitly $\boldsymbol{w}$ and b as a function of $\theta^1, \theta^0, p_1, p_0$.

Solution:

The decision boundary s given by:

$$p_1 \prod_i p(x_i|1) = p_0 \prod_i p(x_i|0)$$

$$p_1 \prod_i \theta_i^{1\,x_i} (1 - \theta_i^1)^{1-x_i} = p_0 \prod_i \theta_i^{0\,x_i} (1 - \theta_i^0)^{1-x_i}$$

Taking log, and defining $\alpha_i = \log \theta_i$, $\beta_i = \log(1 - \theta_i)$, we have

$$\log p_1 + \sum_i x_i \alpha_i^1 + (1 - x_i)\beta_i^1 = \log p_0 + \sum_i x_i \alpha_i^0 + (1 - x_i)\beta_i^0$$

$$\sum_i x_i \underbrace{\left(\alpha_i^1 - \beta_i^1 - \alpha_i^0 + \beta_i^0\right)}_{w_i} + \underbrace{\log p_1 - \log p_0 + \sum_i \left(\beta_i^1 - \beta_i^0\right)}_{b} = 0$$

Problem 5

This question concerns spam filtering. Each email is represented by a vector $x = (x_1,...,x_D)$ where $x_i \in \{0, 1\}$. Each entry of the vector indicates if a particular symbol or word appears in the email. The symbols/words are money, cash, !!!, viagra, . . . , etc. so that for example $x_2 = 1$ if the word 'cash' appears in the email. The training dataset consists of a set of vectors along with the class label c, where c = 1 indicates the email is spam, and c = 0 not spam. Hence, the training set consists of a set of pairs $(\boldsymbol{x^n, c^n})$, n = 1,...,N. The naive Bayes model is given by

$$p(c, x) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

1. Derive expressions for the parameters of this model in terms of the training data using maximum likelihood. Assume that the data is independent and identically distributed

$$p(c^1, \ldots, c^N, x^1, \ldots, x^N) = \prod_{n=1}^{N} p(c^n, x^n)$$

   Explicitly, the parameters are

$$p(c = 1), p(x_i = 1|c = 1), p(x_i = 1|c = 0), i = 1, \ldots, D$$

2. Given a trained model $\boldsymbol{p(x, c)}$, explain how to form a classifier $\boldsymbol{p(c|x)}$.

3. If 'viagra' never appears in the spam training data, discuss what effect this will have on the classification for a new email that contains the word 'viagra'. Explain how you might counter this effect. Explain how a spammer might try to fool a naive Bayes spam filter.

Solution:

1. This is Standard Naive Bayes training using maximum likelihood method as in the lecture.

2.

$$p(c|x) = \frac{p(c) \prod_i p(x_i|c)}{\sum_c p(c) \prod_i p(x_i|c)}$$

3.

The classification is certain that it is not spam. This is because the probability is zero, which annihilates all terms in the product in Naive Bayes. We can use a Bayesian method on the tables to help with this. Placing for example a uniform Dirichlet prior on the tables has the effect of simply adding 1 to all the data counts. This means that there would no longer be any zero counts in the data and the problem is thereby diminished. A spammer could attempt to fool a simple Naive Bayes classifier by for example including the spam text at the beginning of the email, and then appending a large amount of normal text at the end of the email. This would have the effect of biasing the statistics of the email towards a normal email, meaning that the Naive Bayes classifier would class this as a normal ‘ham’ email.

Problem 6 (Parameter learning in partially observed BN)
(Printer Nightmare). Cheapco is, quite honestly, a pain in the neck. Not only did they buy a dodgy old laser printer from StopPress and use it mercilessly, but try to get away with using substandard components and materials. Unfortunately for StopPress, they have a contract to maintain Cheapco's old warhorse, and end up frequently sending the mechanic out to repair the printer. They decide to make a statistical model of Cheapco's printer, so that they will have a reasonable idea of the fault based only on the information that Cheapco's secretary tells them on the phone. In that way, StopPress hopes to be able to send out to Cheapco only a junior repair mechanic, having most likely diagnosed the fault over the phone.

Based on the manufacturer's information, StopPress has a good idea of the dependencies in the printer, and what is likely to directly affect other printer components. The belief network in the figure represents these assumptions. However, the specific way that Cheapco abuse their printer is a mystery, so that the exact probabilistic relationships between the faults and problems is idiosyncratic to Cheapco. StopPress has the following table of faults in which each column represents a visit.

In this problem, consider the BN given in Figure in below. Further, the following table represents data gathered on the printer, where ? indicates that the entry is missing. Each column represents a data point.

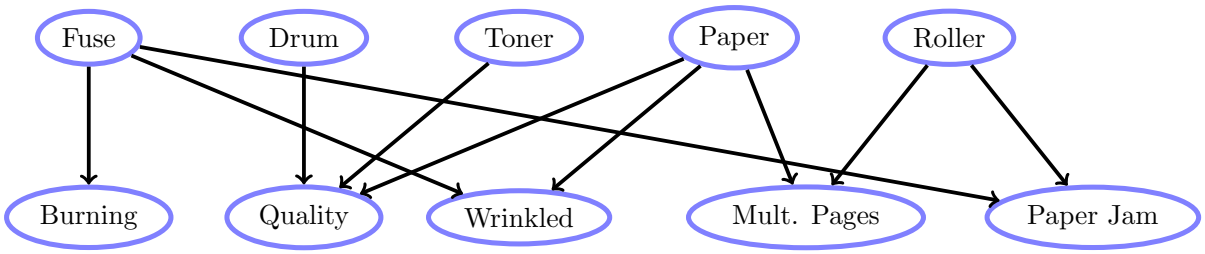


Figure: Printer Nightmare belief network. All variables are binary. The upper variables without parents are possible problems (diagnoses), and the lower variables consequences of problems (faults).

fuse assembly malfunction	?	?	?	1	0	0	?	0	?	0	0	?	1	?	1
drum unit	?	0	?	0	1	0	0	1	?	?	1	1	?	0	0
toner out	1	1	0	?	?	1	0	1	0	?	0	1	?	0	?
poor paper quality	1	0	1	0	1	?	1	0	1	1	?	1	1	?	0
worn roller	0	0	?	?	?	0	1	?	0	0	?	0	?	1	1
burning smell	0	?	?	1	0	0	0	0	0	?	0	?	1	0	?
poor print quality	1	1	1	0	1	1	0	1	0	0	1	1	?	?	0
wrinkled pages	0	0	1	0	0	0	?	0	1	?	0	0	1	1	1
multiple pages fed	0	?	1	0	?	0	1	0	1	?	0	0	?	0	1
paper jam	?	0	1	1	?	0	1	1	1	1	0	?	0	1	?

The table is contained in EMprinter.mat, using states 1, 2, nan; in place of 0, 1, ?.

Use the EM algorithm to learn all CPTs of the network. Given no wrinkled pages, no burning smell and poor print quality, what is the probability there is a drum unit problem?

Solution:

See the code "demoEMprinter.m"

Based on 50 EM iterations, the probability of a drum unit problem, given the evidence and the learned table, is 0.618 to the decimal places.