CS/CNS/EE 155: Probabilistic Graphical Models
Problem Set 2

Handed out:	21 Oct 2009
Due:	4 Nov 2009

1 Tree Augmented Naive Bayes [40 points]

In this problem, you should hand in a printout of your MATLAB implementation. Also email a zip archive with the source code to the TAs. The training data set is given in a file called `trainingData.txt`, available on the course webpage. There are 200 training examples. Each row of the data in the file is a training example. Given a sample, the $1st$ column is the class variable C , and the $2nd$ to the $6th$ columns are the attributes A_1, A_2, A_3, A_4, A_5. The testing data set is given in a file called `testingData.txt`. There are 100 testing samples, with the same format for each sample.

1. **Learning a Naive Bayes model**. You are asked to learn a naive Bayesian network based on a given training data set. The structure of the naive Bayes Network is given as follows:

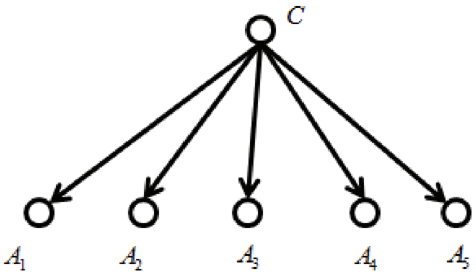

Figure 1: Naive Bayes network.

Estimate the parameters for the conditional probability distributions in the network using MLE on the training data. Based on the constructed naive Bayesian network you can classify samples by applying Bayes rule to compute conditional class probabilities $P(C|A_1, A_2, A_3, A_4, A_5)$, and predicting the label with the highest probability.

Please write down the parameters θ_C and $\theta_{A_1|C}$, and the percentage of classification error on the testing data set.

2. **Learning a Tree Augmented Naive Bayes (TAN) model**. Tree augmented naive Bayes models are formed by adding directional edges between attributes. After removing the class variable, the attributes should form a tree structure (no V-structures). See Fig. 2 as an example.

Use the following procedure to learn the tree augmented naive Bayes model for the training data, then draw the structure of the obtained model.

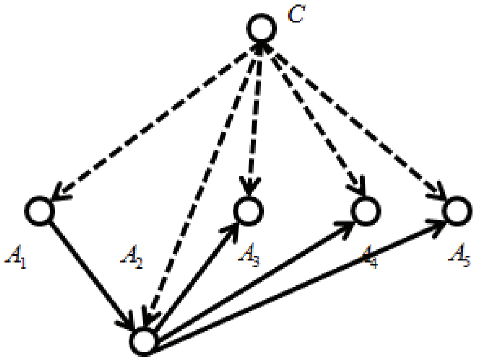

Figure 2: An example of a tree augmented naive Bayes network.

(a) Compute $I_{\widehat{P}_D}(A_i; A_j|C)$ between each pair of attributes, $i \neq j$, where $I_{\widehat{P}_D}(A_i; A_j|C)$ is the conditional mutual information (with respect to the empirical distribution $\widehat{P}_D$ on the training data) between A_i, A_j given the class variable.

$$I_{\widehat{P}_D}(X; Y|C) = \sum_{x,y,c} \widehat{P}_D(x, y, c) \log \frac{\widehat{P}_D(x, y|c)}{\widehat{P}_D(x|c)\widehat{P}_D(y|c)}$$

(b) Build a complete undirected graph in which the vertices are the attributes A_1, A_2, A_3, A_4, A_5. Annotate the weight of an edge connecting A_i and A_j by $I_{\widehat{P}_D}(A_i; A_j|C)$.

(c) Build a maximum weighted spanning tree.

(d) Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.

(e) Construct a tree augmented naive Bayes model by adding a vertex labeled by C and adding an directional edge from C to each A_i.

3. **TAN for classification**. Based on the structure above, you are asked to estimate the parameters for conditional probability distributions using the training data set (using MLE). Then you can classify the testing data set by computing the highest probability $P(C|A_1, A_2, A_3, A_4, A_5)$.

   What is the percentage of classification error for the testing data set? Please compare this result with that using naive Bayesian structure, and explain why it performs better or worse.

4. **Inference**. Based on the TAN model constructed above (the network structure, θ_C and $\theta_{A_i|Pa_i}$ with $1 \leq i \leq 5$), answer the following questions:

   (a) If $A_1 = 1, A_2 = 0, A_3 = 0$ are observed, what is the most likely assignment for (C, A_4, A_5)?

   (b) If $A_1 = 0, A_2 = 0$ are observed, what is the most likely assignment for A_5? In this case, what's the probability for this most likely assignment?

2 Score equivalence [20 points]

1. **K2 prior.** Show that the Bayesian score with a K2 prior in which we have a Dirichlet prior $Dirichlet(1, 1, \ldots, 1)$ for each set of multinomial parameters is not score-equivalent.

   Hint: Construct a data set for which the score of the network $X \to Y$ differs from the score of the network $X \leftarrow Y$.

2. **BDe score equivalence.** Assume that we have a BDe prior specified by an equivalent sample size α and prior distribution P'. Prove the following:

   (a) Consider networks over the variables X and Y. Show that the BDe score of $X \to Y$ is equal to that of $X \leftarrow Y$.

   (b) Show that if $\mathcal{G}$ and $\mathcal{G}'$ are identical except for a covered edge reversal of $X \to Y$, then the BDe score of both networks is equal.

   (c) Show that the proof of score equivalence follows from the result in Part 2b and Theorem 3.9 of [KF09].

   (d) Given the above results, what are the implications for learning optimal trees?

3 The Gaussian Distribution [30 points]

Preamble. The multivariate Gaussian (or Normal) distribution over the D dimensional continuous random vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ has a joint probability distribution given by

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu}$ is the mean vector of length D, and $\boldsymbol{\Sigma}$ is the (symmetric and positive definite) covariance matrix, of size $D \times D$. We sometimes write $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as a shorthand for the above.

Additionally, it is sometimes preferable to work with the *precision matrix*, which is just the inverse of the covariance, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. In this case, we use the notation $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.

We also point out the matrix inversion lemma

$$(Z + UWV^\top)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\top Z^{-1}U)^{-1}V^\top Z^{-1}$$

which will be useful below in rewriting the joint distribution as a factored distribution.

Let $\mathbf{x}$ and $\mathbf{y}$ be two jointly Gaussian random vectors

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim p(\mathbf{x}, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}^\top & \tilde{B} \end{bmatrix}^{-1}\right).$$

1. Show that the marginal distribution of $\mathbf{x}$ is $\mathcal{N}(\boldsymbol{\mu}_x, A)$, i.e.,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, A),$$

i.e.,

$$\int p(\mathbf{x},\mathbf{y})d\mathbf{y} = \int \mathcal{N}\left(\left[\begin{array}{c} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{array}\right], \left[\begin{array}{cc} A & C \\ C^\top & B \end{array}\right]\right) d\mathbf{y} = \mathcal{N}(\boldsymbol{\mu}_x, A)$$

2. (a) Show that the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_x + CB^{-1}(\mathbf{y}-\boldsymbol{\mu}_y), A - CB^{-1}C^\top).$$

(b) Equivalently, show that $\mathbf{x}|\mathbf{y}$ in terms of the precision matrix

$$\boldsymbol{\Lambda} = \left[\begin{array}{cc} \tilde{A} & \tilde{C} \\ \tilde{C}^\top & \tilde{B} \end{array}\right]$$

is

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_x - \tilde{A}^{-1}\tilde{C}(\mathbf{y}-\boldsymbol{\mu}_y), \tilde{A}^{-1}).$$

3. **Conjugate prior of multivariate Gaussian with known covariance**

Preamble. For a multivariate Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})$ in which both the mean $\boldsymbol{\mu}$ and the precision $\boldsymbol{\Lambda}$ are unknown, the conjugate prior is the *Gaussian-Wishart* distribution:

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{\mu}_0,\beta,\mathbf{W},\nu) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0,(\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu)$$

where the Wishart distribution is given by

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W},\nu) = B(\mathbf{W},\nu)|\boldsymbol{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)$$

and

$$B(\mathbf{W},\nu) = |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1}$$

and $\mathbf{W}$ is a $D\times D$ symmetric, positive definite matrix (see the appendix of [Bis06]). The parameter ν is called the *number of degrees of freedom* of the distribution and is restricted to $\nu > D-1$.

Conjugate prior for the mean of a 1D Gaussian. In the case of 1-dimensional Gaussians, where the variance σ^2 is known, the data $\mathcal{D} = (x_1,\ldots,x_N)$, and the likelihood is given by

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right),$$

show that the conjugate prior for the mean is given by

$$p(\mu) = \mathcal{N}(\mu|\mu_0,\sigma_0^2),$$

and provide the parameters for the posterior, $p(\mu|\mathcal{D})$.

4. **The exponential family** The exponential family of distributions over $\mathbf{x}$, given parameters $\boldsymbol{\eta}$, is defined to be the set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^\top\mathbf{u}(\mathbf{x}))$$

Show that the multivariate Gaussian can be expressed as a member of the exponential family.

4 MAP versus MPE [10 points]

Show that the MAP (*maximum a-posteriori*) assignment does not necessarily equal the MPE (Most Probable Explanation). I.e., construct a Bayes net such that the most likely configuration of all variables does not agree with the most likely assignment to a single variable (marginalizing out the remaining variables).

References

[Bis06] C. Bishop. *Pattern Recognition and Machine Learning.* Springer Science+Business Media, LLC, New York, NY, 2006.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press, Cambridge, MA, 2009.