# 10-701 Machine Learning, Spring 2011: Homework 4

Due: Tuesday March 1st at the begining of the class

**Instructions** There are two questions on this assignment. Please submit your writeup as two separate sets of pages according to questions, with your name and userid on each set.

## 1 EM and Bayes Nets [Yi Zhang, 40 points]

In the class we learned the famous expectation-maximization algorithm, which is widely used to estimate model parameters from partially observed data. In this question, we will use **[EM]** to denote the lecture slides on Feb 17th ("Graphical models 3: EM") from the course website. Since each page of [EM] contains two slides, we will use, e.g., the upper/lower slide on page 7 of [EM], as a pointer to the first/second slide on a specific page of lecture slides.

Given the set of observed variables $\mathbf{X}$ and the set of unobserved variables $\mathbf{Z}$, as shown by the lower slide on page 6 of [EM], the EM algorithm for estimating model parameters $\theta$ from training examples is the following procedure:

1. E-Step: for each example $k$, use $\mathbf{X}_k$ and the current $\theta$ to calculate $P(\mathbf{Z}_k|\mathbf{X}_k, \theta)$.
2. M-Step: re-estimate $\theta$ as $\theta \leftarrow \text{argmax}_{\theta'} E_{P(\mathbf{Z}|\mathbf{X},\theta)}[\log P(\mathbf{X}, \mathbf{Z}|\theta')]$ using the training examples.
3. Iterate until convergence.

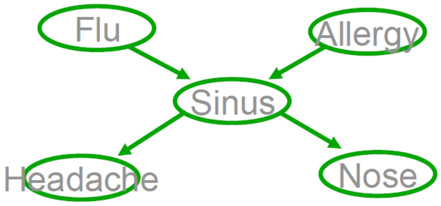


Figure 1: The Bayes net for the flu-allergy example

Now consider the familiar flu-allergy Bayes net as figure 1. Suppose $F, A, H, N$ are observed (i.e., $\mathbf{X} = \{F, A, H, N\}$) and $S$ is unobserved (i.e., $\mathbf{Z} = \{S\}$). Given a set of training examples $\{(f_k, a_k, h_k, n_k)\}_{k=1}^{K}$, the EM algorithm should proceed as follows:

1. E-Step: for each example $k$, use $(f_k, a_k, h_k, n_k)$ and the current $\theta$ to calculate $P(S_k|f_k, a_k, h_k, n_k, \theta)$.
2. M-Step: re-estimate $\theta$ as $\theta \leftarrow \text{argmax}_{\theta'} \sum_{k=1}^{K} E_{P(S_k|f_k,a_k,h_k,n_k,\theta)}[\log P(f_k, a_k, h_k, n_k, S_k|\theta')]$.
3. Iterate until convergence.

### 1.1 A simplified EM algorithm

As we saw in the lower slide on page 8 of [EM], the EM algorithm can be simplified when the unobserved variable is of boolean values. In this case, the *simplified* EM algorithm is:

1. E-Step: for each example $k$, use $\mathbf{X}_k$ and the current $\theta$ to calculate the expected value $E(\mathbf{Z}_k|\mathbf{X}_k, \theta)$.
2. M-Step: re-estimate $\theta$ similarly to MLE on observed data, but replacing each count involving the unobserved variable by its expected count.
3. Iterate until convergence.

As a result, as shown by the upper slide on page 8 of [EM], the *simplified* EM algorithm for the flu-allergy Bayes net with a set of training examples $\{(f_k, a_k, h_k, n_k)\}_{k=1}^K$ is:

1. E-Step: for each example $k$, use $(f_k, a_k, h_k, n_k)$ and the current $\theta$ to calculate the expected value $E(S_k) = E(S_k|f_k, a_k, h_k, n_k, \theta)$.
2. M-Step: re-estimate $\theta$ as MLE with expected counts, e.g., $\theta_{s|ij} \leftarrow \frac{\sum_{k=1}^K \delta(f_k=i,a_k=j)E(S_k)}{\sum_{k=1}^K \delta(f_k=i,a_k=j)}$. Recall that $\theta_{s|ij}$ stands for $P(S = 1|F = i, A = j)$.
3. Iterate until convergence.

**Question [15 pts]:** Given that $S$ in the flu-allergy Bayes net is a boolean variable, prove that the simplified EM algorithm for estimating $\theta_{s|ij}$ is indeed equivalent to the standard EM algorithm. In other words, show that $\theta_{s|ij} \leftarrow \frac{\sum_{k=1}^K \delta(f_k=i,a_k=j)E(S_k)}{\sum_{k=1}^K \delta(f_k=i,a_k=j)}$ in the M-Step of the simplified EM indeed gives the corresponding parameter in $\theta \leftarrow \text{argmax}_{\theta'} \sum_{k=1}^K E_{P(S_k|f_k,a_k,h_k,n_k,\theta)}[\log P(f_k, a_k, h_k, n_k, S_k|\theta')]$ from the M-Step of the standard EM algorithm.

Note: you are required to prove this only for $\theta_{s|ij}$, and no need to worry about other parameters in $\theta$.

Hint: the lower slide on page 4 of [EM] should be helpful. The slide shows how to derive the parameter $\theta_{s|ij}$ in $\theta \leftarrow \text{argmax}_{\theta'} \sum_{k=1}^K [\log P(f_k, a_k, h_k, n_k, s_k|\theta')]$ when all variables are observed.

★ **SOLUTION:** To find $\text{argmax}_{\theta'_{s|ij}} \sum_{k=1}^K E_{P(S_k|f_k,a_k,h_k,n_k,\theta)}[\log P(f_k, a_k, h_k, n_k, S_k|\theta')]$ in the M-Step, we need to take the derivative w.r.t. $\theta'_{s|ij}$ and set it to zero. Let's write $P(S_k|\theta) = P(S_k|f_k, a_k, h_k, n_k, \theta)$ for short. Note that $P(S_k|\theta)$ does not depend on $\theta'$ (or $\theta'_{s|ij}$).

$$\begin{aligned}
& \frac{\partial \sum_{k=1}^K E_{P(S_k|\theta)}[\log P(f_k, a_k, h_k, n_k, S_k|\theta')]}{\partial \theta'_{s|ij}} \\
= & \frac{\partial \sum_{k=1}^K E_{P(S_k|\theta)}[\log P(S_k|f_k, a_k, \theta')]}{\partial \theta'_{s|ij}} \qquad \text{(only } P(S_k|f_k, a_k) \text{ matters)} \\
= & \frac{\partial \sum_{k=1}^K [P(S_k = 1|\theta) \log P(S_k = 1|f_k, a_k, \theta') + P(S_k = 0|\theta) \log P(S_k = 0|f_k, a_k, \theta')]}{\partial \theta'_{s|ij}} \qquad \text{(expectation over } P(S_k|\theta)) \\
= & \frac{\partial \sum_{k=1}^K [E(S_k) \log P(S_k = 1|f_k, a_k, \theta') + (1 - E(S_k)) \log P(S_k = 0|f_k, a_k, \theta')]}{\partial \theta'_{s|ij}} \qquad \text{(definition of } E(S_k)) \\
= & \frac{\partial \sum_{k=1}^K \delta(f_k = i, a_k = j)[E(S_k) \log P(S_k = 1|f_k, a_k, \theta') + (1 - E(S_k)) \log P(S_k = 0|f_k, a_k, \theta')]}{\partial \theta'_{s|ij}} \qquad \text{(only } f_k = i, a_k = j) \\
= & \sum_{k=1}^K \delta(f_k = i, a_k = j)[E(S_k) \frac{\partial \log P(S_k = 1|f_k, a_k, \theta')}{\partial \theta'_{s|ij}} + (1 - E(S_k)) \frac{\partial \log P(S_k = 0|f_k, a_k, \theta')}{\partial \theta'_{s|ij}}] \\
= & \sum_{k=1}^K \delta(f_k = i, a_k = j)[E(S_k) \frac{\partial \log \theta'_{s|ij}}{\partial \theta'_{s|ij}} + (1 - E(S_k)) \frac{\partial \log(1 - \theta'_{s|ij})}{\partial \theta'_{s|ij}}] \\
= & \sum_{k=1}^K \delta(f_k = i, a_k = j)[E(S_k) \frac{1}{\theta'_{s|ij}} - (1 - E(S_k)) \frac{1}{(1 - \theta'_{s|ij})}]
\end{aligned}$$

Set the above formula to zero, we have $\theta'_{s|ij} = \frac{\sum_{k=1}^K \delta(f_k=i,a_k=j)E(S_k)}{\sum_{k=1}^K \delta(f_k=i,a_k=j)}$.

## 1.2 Simulating the simplified EM algorithm

We are given the following $K = 8$ training examples as in Table 1, where only two examples contain unobserved values, namely, $H_7$ and $N_8$. We like to simulate a few steps of the simplified EM algorithm by hand. Note that this is not a programming question.

Table 1: Training examples for the flu-allergy Bayes net

| k | F | A | S | H | N |
|---|---|---|---|---|---|
| $k = 1$ | 1 | 0 | 1 | 1 | 1 |
| $k = 2$ | 0 | 1 | 1 | 1 | 0 |
| $k = 3$ | 1 | 1 | 1 | 1 | 1 |
| $k = 4$ | 0 | 0 | 0 | 0 | 0 |
| $k = 5$ | 0 | 0 | 0 | 1 | 0 |
| $k = 6$ | 0 | 0 | 0 | 0 | 1 |
| $k = 7$ | 1 | 1 | 1 | **?** | 1 |
| $k = 8$ | 1 | 1 | 1 | 1 | **?** |

**Question A [7 pts]:** given that all variables are boolean, how many parameters we need to estimate in the flu-allergy Bayes net? Also, list all the parameters we need to estimate, e.g., one parameter will be $\theta_{s|11}$, which stands for $P(S = 1|F = 1, A = 1)$.

★ **SOLUTION:** We have 10 parameters:

$$\begin{aligned}
\theta_f &= P(F = 1) \\
\theta_a &= P(A = 1) \\
\theta_{s|00} &= P(S = 1|F = 0, A = 0) \\
\theta_{s|01} &= P(S = 1|F = 0, A = 1) \\
\theta_{s|10} &= P(S = 1|F = 1, A = 0) \\
\theta_{s|11} &= P(S = 1|F = 1, A = 1) \\
\theta_{h|0} &= P(H = 1|S = 0) \\
\theta_{n|1} &= P(H = 1|S = 1) \\
\theta_{n|0} &= P(N = 1|S = 0) \\
\theta_{n|1} &= P(N = 1|S = 1)
\end{aligned}$$

**Question B [6 pts]:** Now we like to simulate the first E-step of the simplified EM algorithm. Before we start, we initialize all the parameters as 0.5, and then proceed to execute the E-step. What are the expectations we need to calculate in this E-step? Also, list the actual values for these expectations. (Note that only two examples contains unobserved variables).

★ **SOLUTION:** $E(H_7|f_7, a_7, s_7, n_7, \theta) = E(H_7|s_7 = 1, \theta_{h|1}) = 0.5$, and $E(N_8|f_8, a_8, s_8, h_8, \theta) = E(N_8|s_8 = 1, \theta_{n|1}) = 0.5$.

**Question C [6 pts]:** Now we like to simulate the first M-step. List the estimated values of all model parameters we obtain in this M-step. (Note that we use the expected count only when the variable is unobserved in an example).

★ **SOLUTION:**

$$\begin{aligned}
\theta_f &= P(F=1) = 0.5 \\
\theta_a &= P(A=1) = 0.5 \\
\theta_{s|00} &= P(S=1|F=0,A=0) = 0 \\
\theta_{s|01} &= P(S=1|F=0,A=1) = 1 \\
\theta_{s|10} &= P(S=1|F=1,A=0) = 1 \\
\theta_{s|11} &= P(S=1|F=1,A=1) = 1 \\
\theta_{h|0} &= P(H=1|S=0) = 1/3 \\
\theta_{n|1} &= P(H=1|S=1) = 4.5/5 = 0.9 \\
\theta_{n|0} &= P(N=1|S=0) = 1/3 \\
\theta_{n|1} &= P(N=1|S=1) = 3.5/5 = 0.7
\end{aligned}$$

**Question D [6 pts]:** Last, let's simulate the second E-step. List the actual values for all the expectations we calculate in this E-step.

★ **SOLUTION:** $E(H_7|f_7, a_7, s_7, n_7, \theta) = E(H_7|s_7 = 1, \theta_{h|1}) = 0.9$, and $E(N_8|f_8, a_8, s_8, h_8, \theta) = E(N_8|s_8 = 1, \theta_{n|1}) = 0.7$.

## 2 Midterm review questions [Tom Mitchell, 60 points]

This question contains some short questions adapted from previous midterm exams – a good way to review for our own on March 3.

1. Give a *one sentence* reason why [12 pts]:

   - we might prefer Decision Tree learning over Logistic Regression for a particular learning task.
   - we might prefer Logistic Regression over Naive Bayes for a particular learning task.
   - we choose parameters that minimize the sum of squared training errors in Linear Regression.

   ★ **SOLUTION:**

   (a) We prefer Decision Tree over Logistic Regression when we expect the decision boundary to be nonlinear.

   (b) We prefer Logistic Regression over Naive Bayes when the conditional independence assumption does not hold for the data. (Note that Logistic Regression does not make the conditional independence assumption).

   (c) We choose parameters that minimize the sum of squared training errors because it corresponds to find the MLE (maximum likelihood estimate) of $P(Y|X)$ assuming that $P(Y|X)$ follows a Gaussian distribution.

2. Suppose we train several classifiers to learn $f : X \rightarrow Y$, where $X$ is the feature vector $X =< X_1, X_2, X_3 >$. Which of the following classifiers contains sufficient information to allow calculating $P(X_1, X_2, X_3, Y)$? If you answer yes, give a brief sketch of how. If you answer no, state what is missing. [12 pts]

   - Gaussian Naive Bayes
   - Logistic Regression
   - Linear Regression

★ **SOLUTION:**

(a) YES. Naive Bayes estimates $P(X_1|Y)$, $P(X_2|Y)$, $P(X_3|Y)$ and $P(Y)$ and calculates $P(X_1, X_2, X_3, Y)$ by:

$$P(X_1, X_2, X_3, Y) = P(X_1|Y)P(X_2|Y)P(X_3|Y)P(Y)$$

(b) NO. Since Logistic Regression only estimates $P(Y|X_1, X_2, X_3)$, we cannot calculate $P(X_1, X_2, X_3, Y)$ without any information of $P(X_1, X_2, X_3)$.

(c) NO. The reason is similar to Logistic Regression. In Linear Regression, we assume that $P(Y|X_1, X_2, X_3) \sim N(w_0 + w_1X_1 + w_2X_2 + w_3X_3, \sigma^2)$ and Linear Regression estimates $\{w_0, w_1, w_2, w_3\}$ via MLE. But still, we do not have any information of $P(X_1, X_2, X_3)$ and hence cannot calculate $P(X_1, X_2, X_3, Y)$.

3. True or False? If true, give a 1-2 sentence explanation. If false, a counterexample. *Your answer must fit into the space below the question* [12 pts].

- As the number of data points grows to infinity, the MLE estimate of a parameter approaches the MAP estimate, for all possible priors.

- The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

- There is *no* training data set for which a decision tree learner and logistic regression will output the same decision boundary.

★ **SOLUTION:**

(a) False. Suppose the prior probability is assigned to be 1 at a single point and zero at all other points (i.e. $P(\theta = \hat{\theta}) = 1)$). In this case, MLE and MAP are different as the number of data points approaches infinity.

(b) False. Assume that we have $N$ data points. Since each tree node contains at least 1 data point (otherwise, we will not split its parent node), there will be at most $N - 1$ splits, each of which increases the depth at most by 1. Therefore, the depth of the tree is at most $N$. [Note that if the number of the features is more than the number of data points, the depth of the tree is still at most $N$. In this case, we will not split on all the features.]

(c) False. A simple counter example is as follows: the data set contains two data points: $\{X = -1, Y = -1\}$ and $\{X = 1, Y = 1\}$. The decision surfaces for decision tree and logistic regression are the same.

4. In class we defined *conditional independence* by saying that random variable $X$ is conditionally independent of $Y$ given $Z$ if and only if:

$$P(X|Y, Z) = P(X|Z) \tag{1}$$

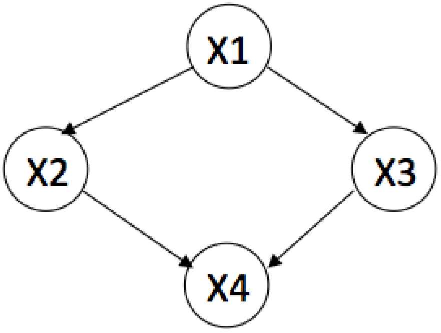
Prove that if $P(XY|Z) = P(X|Z)P(Y|Z)$, then $X$ is conditionally independent of $Y$ given $Z$ (*hint: this is a two-line proof*) [4 pts].

★ **SOLUTION:** If $P(XY|Z) = P(X|Z)P(Y|Z)$, we have

$$P(X|Y,Z) = \frac{P(XY|Z)}{P(Y|Z)} = \frac{P(X|Z)P(Y|Z)}{P(Y|Z)} = P(X|Z).$$

Therefore, $X$ is conditionally independent of $Y$ given $Z$.

5. Consider the Bayes network below, defined over four Boolean variables [20 pts].



- How many parameters are needed to define $P(X1, X2, X3, X4)$ for this Bayes Net?
- Give the formula that calculates $P(X1 = 1, X2 = 0, X3 = 1, X4 = 0)$ using only the Bayes net parameters. Use notation like $P(X1 = 0|X2 = 1, X4 = 0)$ to refer to each Bayes net parameter you use in your formula.
- Give the formula that calculates $P(X1 = 1, X4 = 0)$ using only the Bayes net parameters.
- Give the formula that calculates $P(X2 = 1|X3 = 0)$ using only the Bayes net parameters.

★ **SOLUTION:**

(a) 9 parameters: $P(X1 = 1)$, $P(X2 = 1|X1 = 1)$, $P(X2 = 1|X1 = 0)$, $P(X3 = 1|X1 = 1)$, $P(X3 = 1|X1 = 0)$, $P(X4 = 1|X2 = 1, X3 = 1)$, $P(X4 = 1|X2 = 1, X3 = 0)$, $P(X4 = 1|X2 = 0, X3 = 1)$, $P(X4 = 1|X2 = 0, X3 = 0)$. [Note that the parameters $P(Xk = 0|\cdots)$ can be then computed $1 - P(Xk = 1|\cdots)$.]

(b)

$$\begin{aligned}
& P(X1 = 1, X2 = 0, X3 = 1, X4 = 0) \\
= \; & P(X1 = 1)P(X2 = 0|X1 = 1)P(X3 = 1|X1 = 1)P(X4 = 0|X2 = 0, X = 1) \\
= \; & P(X1 = 1)\,(1 - P(X2 = 1|X1 = 1))\,P(X3 = 1|X1 = 1)\,(1 - P(X4 = 1|X2 = 0, X = 1))
\end{aligned}$$

(c)

$$\begin{aligned}
& P(X1 = 1, X4 = 0) \\
= \; & \sum_{x_2=0}^{1}\sum_{x_3=0}^{1} P(X1 = 1, X2 = x_2, X3 = x_3, X4 = 0) \\
= \; & P(X1 = 1)\sum_{x_2=0}^{1} P(X2 = x_2|X1 = 1)\sum_{x_3=0}^{1} P(X3 = x_3|X1 = 1)P(X4 = 0|X2 = x_2, X3 = x_3)
\end{aligned}$$

(d)

$$\begin{aligned} & P(X2=1|X3=0) \\ = \quad & \frac{P(X2=1, X3=0)}{P(X3=0)} \\ = \quad & \frac{\sum_{x_1=0}^{1} P(X1=x_1)P(X2=1|X1=x_1)P(X3=0|X1=x_1)}{\sum_{x_1=0}^{1} P(X1=x_1)P(X3=0|X_1=x_1)} \end{aligned}$$