CMPUT 466/551 — Assignment 4

Instructors: R Greiner, B Póczos
Due Date: 5:00pm, Monday, 7/Dec/09
The following exercises are intended to further your understanding of PAC learning, Belief Networks, Expectation Maximization, Principle Component Analysis, and Independent Component Analysis.
Relevant reading: Lecture notes;
HTF: Chapter 14.5, 18 (skim);
(Bishop: Chapter 7.1.5, 8, 12)
Total points: UGrad: 55 Grad: 55

Question 1 *[10 points] Universal Set; tools from PAC learning*
A set $S = \{\mathbf{x}_1, \ \ldots, \ \mathbf{x}_m\}$ of binary d-tuples (*i.e.*, each $\mathbf{x}_k = \langle x_1^{(k)}, \ldots, x_d^{(k)} \rangle \ \in \{0,1\}^d$) is a (d,k)-*universal set* if, for every assignment to any subset of k variables, S includes an element that agrees with that assignment. That is, pick any of the $\binom{d}{k}$ size-k subsets of the d variables — call them $\{X_{i_1}, \ldots, X_{i_k}\}$ where each $i_j \in \{1, \ldots, d\}$ — and then pick any one of the 2^k assignments to these variables, say $t_{i_j} \in \{0,1\}$ for each j. Then there is (at least) one element $\mathbf{x} \in S$ such that $x_{i_j} = t_{i_j}$ for all $j = 1..d$.

As an example, consider the set of $d = 4$ tuples:

$$S \quad = \quad \left[\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{array} \right]$$

To be $(4,2)$-universal set, it would have include all $2^2 = 4$ assignments to each of the $\binom{4}{2} = 6$ pairs, $\langle x_i, x_j \rangle$. Fortunately, S does include all $2^2 = 4$ assignments to $\langle x_1, x_2 \rangle$ — *i.e.*, it includes $\langle x_1, \ x_2 \rangle = \langle 0,0 \rangle$, $\langle 0,1 \rangle$, $\langle 1,0 \rangle$ and $\langle 1,1 \rangle$. It also includes all 4 assignments to $\langle x_1, \ x_3 \rangle$, $\langle x_1, \ x_4 \rangle$, $\langle x_2, \ x_3 \rangle$, and $\langle x_3, \ x_4 \rangle$. However, this S is NOT a $(4,2)$-universal set as it does not include every possible assignment to $\langle x_2, x_4 \rangle$: it includes $\langle x_2, \ x_4 \rangle = \langle 0,0 \rangle$ and $\langle 1,1 \rangle$, but it does *not* include either $\langle 0,1 \rangle$ or $\langle 1,0 \rangle$.

Now consider

$$S' \quad = \quad \left[\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{array} \right]$$

and notice this S' is a $(4,2)$-universal set.

There are elaborate algorithms that are guaranteed to produce such (d,k)-universal sets. But how hard is it, really?

Suppose you just generate a set of $m(d,k)$ binary d-tuples, RANDOMLY — *i.e.*, each $x_i^{(k)}$ is drawn uniformly from $\{0,1\}$. How large does $m(d,k)$ have to be, to be $1-\delta$ confident that this set is a (d,k)-universal set?

(Of course, you should expect this to be at least 2^k.)

[Hint: 1. What is the chance that a random d-tuple (think "row in the matrix") does NOT include a particular assignment to a particular k-tuple of columns?
2. How many such "conditions" need to be satisfied?
3. Use this to bound the chance that a sample containing $m(d,k)$ instances does NOT qualify — i.e.*, that there is a particular k-tuple of columns that does NOT contain a particular assignment.*
You may want to prove, then use, that $\log(1-\epsilon) < -\epsilon$ holds for all $\epsilon \in (0,1)$.]

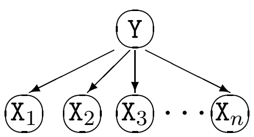
Question 2 *[4 points] Belief Networks (Independencies)*
Given variables A, B, C, we say that A is independent of B, given C — written "$A \perp B \,|\, C$" — iff $\forall a,b,c\ P(\, A=a \,|\, B=b, C=c\,) = P(\, A=a \,|\, C=c\,)$.

Prove or disprove the following statements. (You may assume that these variables are discrete, and that every probability is non-zero — *i.e.*, $P(\, X=x\,) > 0$.)

a [2]: $A \perp B \,|\, C \quad \Longrightarrow \quad B \perp A \,|\, C.$

b [2]: $A \perp B \,|\, C \quad \Longrightarrow \quad A \perp C \,|\, B.$

Question 3 *[10 points] NaiveBayes + Conditional Likelihood*
As you recall, the parameters $\Theta = \{\theta_y\} \cup \{\theta_{x_i|y}\}$ for the standard NaiveBayes model



are trained *generatively*, to optimize (log) likelihood of the training data $S = \{\, \langle \mathbf{x}_i, y_i \rangle \,\}$; *i.e.*,

$$\Theta_{ML}^{(*)} = \operatorname*{argmax}_{\Theta} P(\, S \,|\, \Theta\,) = \operatorname*{argmax}_{\Theta} \sum_{\langle \mathbf{x}, y\rangle \in S} \log P_\Theta(\, y, \mathbf{x}\,)$$

Of course, we will later use this NaiveBayes model for the *discriminative* task of predicting y given $\mathbf{x}$. This suggests it might make sense to, instead, seek the parameters that optimize *conditional* likelihood

$$\Theta_{MCL}^{(*)} = \operatorname*{argmax}_{\Theta} \sum_{\langle \mathbf{x}, y\rangle \in S} \log P_\Theta(\, y \,|\, \mathbf{x}\,) \tag{1}$$

Consider the simple case where everything is binary — $y \in \{0,1\}$ and $x_{i,j} \in \{0,1\}$. Also, let $\beta_y = \log \theta_y$ and $\beta_{x_i|y} = \log \theta_{x_i|y}$ be the logs of the corresponding θ parameters (which you may assume are all non-zero).

a [3]: Express the value of $P_\Theta(\, y=1 \,|\, \mathbf{x}\,)$ in terms of these β_χ parameters.

b [6]: Write $f_+(\mathbf{x}) = P_\Theta(\, y=1 \,|\, \mathbf{x}\,)$ as an explicit function of the values $\mathbf{x}$. You may assume that $\mathbf{x} = \langle 1, x_1, \ldots, x_n \rangle$.

[Hint: Observe $\beta_{x_i=a|y} \;=\; \beta_{x_i=0|y} \;+\; a\,(\beta_{x_i=1|y} - \beta_{x_i=0|y})$ *for* $a \in \{0,1\}$*.]*

c [1]: Quickly describe an algorithm for finding the optimal values for these parameters — *i.e.*, that optimize Equation 1.

Question 4 *[15 points] Mixture of Gaussians; EM*
You are to compute maximum likelihood estimates of the parameters $\theta, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2$ of the following distribution of the discrete variable G that represents a person's gender, and the continuous variable X that represents a person's height:

$$
\begin{aligned}
P(\,G = 1\,) \;&=\; \theta \\
P(\,G = 0\,) \;&=\; (1-\theta) \\
P(\,X = x \,|\, G = 1\,) \;&=\; P_{\mathcal{N}}(x;\, \mu_1, \sigma_1^2) \\
P(\,X = x \,|\, G = 0\,) \;&=\; P_{\mathcal{N}}(x;\, \mu_0, \sigma_0^2)
\end{aligned}
$$

where $P_{\mathcal{N}}(x;\, \mu, \sigma^2)$ is the Gaussian probability distribution function with mean μ and variance σ^2. This model is a *mixture* of two Gaussian distributions, one for females and one for males.

Several sub-questions below ask for "high-level pseudo-code" for some algorithm. It is critical that your code here be simple and concise — while Matlab is not required, the person grading your assignment will probably be thinking this way. Note also that each function should be only a few lines. Finally, you are ALLOWed to actually implement your code, if you wish. (This is not required.)

a [2]: What is the marginal distribution of X — *i.e.*, what is the pdf $p(X = x)$?

b [2]: What is the distribution $P(\,G = g \,|\, X = x\,)$?

c [5]: Suppose that, in order to make your assignment extremely easy, your TA has gone out and measured people's height (at a local bar, say) and given you a list of i.i.d. instance of height+genders pairs $\langle x_i,\, g_i \rangle$, $i \in \{1..N\}$ where $x_i \in \Re^+$ is the height of the person i and $g_i = 1$ holds if i is female, and $g_i = 0$ if i is male. Assume these are drawn from the above distribution. Express the maximum likelihood estimates of the above five parameters in terms of x_i and g_i. (You don't need to derive them, just write them down.) Write high-level pseudo-code for the function

```
function [theta, mu_1, sig2_1, mu_0, sig2_0, loglike] = maxlike(x, g)
```

that returns the maximum likelihood parameter estimates, as well as the log likelihood of the data given those estimates. You should treat the vector `g` as a vector of probabilities, where the ith entry gives the probability that person i is female — *i.e.*, don't use an '`if`' statement to determine which Gaussian distribution to use, but rather treat $\mathtt{g}_i$ as an indicator variable.

Note: You may assume this sample includes at least one male, and at least one female.

d [3]: Suppose that, while out at the bar, a clumsy patron spilled a drink on the half of the sheet of paper on which your TA was recording the *genders*, rendering this gender data unavailable. However, the TA notices that *if only we knew the parameters of the distribution*, we could determine the probability that each data point was female, say. (He assumes that you students have already completed part (b).) Write high-level pseudo-code for

```
function [g] = expectation(x, theta, mu_1, sig2_1, mu_0, sig2_0)
```

that computes the expected value of each g_i given x_i and the five parameters, which in our case also happens to be the probability $P(\,G = 1\,|\,x\,)$.

e [3]: You now have the components necessary to run "Expectation Maximation" (EM) to estimate the parameters. Write the high-level pseudo-code

```
function [theta, mu_1, sig2_1, mu_0, sig2_0, g, loglike]  =
    emiteration(x, theta, mu_1, sig2_1, mu_0, sig2_0)
```

that takes the current parameter guesses and the observed data vector `x` and returns a new set of parameter estimates, along with the vector of expectations `g` and the log likelihood of the data given the new parameters.

Question 5 *[10 points] PCA/ICA: Independence, Correlation*
Definitions:

- Y and Z are independent $\;\Leftrightarrow\; p(y,z) \;=\; p(y)\,p(z)$
- (correlation) $corr(Y,Z) \;=\; \frac{\mathbb{E}\left[(Y-\mathbb{E}[Y])\;(Z-\mathbb{E}[Z])\right]}{var(Y)^{1/2}\;var(Z)^{1/2}}$

  $corr(Y,Z) \;=\; 0$ means Y and Z are uncorrelated.
  Note that the numerator is the "covariance" $cov(Y,Z) \;=\; \mathbb{E}\big[(Y-\mathbb{E}[Y])\;(Z-\mathbb{E}[Z])\big]$.

a [2]: Prove: Y and Z are independent $\;\Rightarrow\; \mathbb{E}[\,g(Y)\;h(Z)\,] \;=\; \mathbb{E}[\,g(Y)\,]\,\mathbb{E}[\,h(Z)\,]$, where $g(\cdot)$ and $h(\cdot)$ are arbitrary functions (provided only that their expected values are well defined).

b [2]: Prove: $corr(Y,Z) = 0 \;\;\Leftrightarrow\;\; \mathbb{E}[\,Y\,Z\,] \;=\; \mathbb{E}[Y]\,\mathbb{E}[Z]$

c [1]: Prove: Y and Z are independent $\;\Rightarrow\;$ Y and Z are uncorrelated.

d [3]: Show an example where Y and Z are uncorrelated but Y and Z are not independent.

e [2]: Prove: if (Y_1, Y_2) are jointly Gaussian, then
Y_1 and Y_2 are independent $\;\Leftrightarrow\;$ Y_1 and Y_2 are uncorrelated.

Question 6 *[6 points] PCA can be used for whitening*
Let $\mathbf{A} \in \Re^{N\times M}$ be a full rank matrix, $N \geq M$.
Let $\mathbf{s} \in \Re^{M}$ be a random variable such that $\mathbb{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_M$, and let $\mathbf{x} \;=\; \mathbf{A}\,\mathbf{s} \;\in \Re^{N}$. Prove:
$\exists\; \mathbf{Q} \in \Re^{M\times N}$ such that, using $\mathbf{A}^* = \mathbf{Q}\mathbf{A}$, if $\mathbf{x}^* \doteq \mathbf{Q}\mathbf{x}$ then:

$$\begin{aligned}
\mathbf{x}^* &= \mathbf{A}^*\mathbf{s} \\
\mathbf{A}^*\mathbf{A}^{*T} &= \mathbf{I}_M \\
\mathbb{E}[\mathbf{x}^*\mathbf{x}^{*T}] &= \mathbf{I}_M
\end{aligned}$$