

Final Exam

December 11th, 2012

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam. Note that there is an appendix with possibly useful formulae and computational shortcuts at the end.
- This exam booklet contains **five** problems, out of which you are expected to answer **four** problems of your choice.
- The exam ends at 10:45 AM. You have 75 minutes to earn a total of 100 points. You can earn 25 additional (bonus) points if you successfully attempt all five problems.
- If you choose to attempt all five problems, the four problems with the highest points will be considered for your final exam score and the lowest will be considered as bonus.
- Answer each question in the space provided. If you need more room, write on the reverse side of the paper and indicate that you have done so.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Name (NetID): (1 Point)

Short Questions		/24
Support Vector Machines		/25
Probability		/25
Naive Bayes		/25
Expectation Maximization		/25
Total		/100
Bonus		/25

Short Questions [24 points]

(a) [**10 points**] We define a class $C_{r,k,n}$ of r -of- k functions in the following way. Let $X = \{0, 1\}^n$. For a chosen set of k relevant variables and a given number r , an r -of- k function $f(x_1, \dots, x_n)$ is 1 if and only if at least r of the k relevant attributes are 1. We assume that $1 \leq r \leq k \leq n$.

1. [**5 points**] Phrase this problem as a problem of learning a Boolean disjunction over some feature space. Define the feature space and the learning problem. Use feature space of r -conjunctions over original variables. Now the final hypothesis is the disjunction of the $\binom{k}{r}$ features representing r -conjunctions of the k relevant variables.

2. [**5 points**] Assume you are learning this function using Winnow. What mistake bound do you obtain?

Winnow makes $O(k' \log n')$ mistakes, $k' = \binom{k}{r}$, $n' = \binom{n}{r}$.

- (b) [8 points] According to the CDC, Women who smoke are 21 times more likely to develop lung cancer compared to those who don't smoke. Furthermore, CDC tells us that about 10% of the total women smoke. If you learn that a woman has been diagnosed with lung cancer, and you know nothing else about her, what is the probability that she is a smoker?

Let A be the event that the women has cancer and let B be the event that she smokes. We are given that and $Pr[B] = 0.1$ and that

$$Pr[A|B] = 21Pr[A|\neg B]. \quad (1)$$

We want to find out $Pr[B|A]$. That is given by

$$\begin{aligned} Pr[B|A] &= \frac{Pr[A|B]Pr[B]}{Pr[A|B]Pr[B] + Pr[A|\neg B]Pr[\neg B]} \quad \text{Bayes rule} \\ &= \frac{21Pr[B]}{21Pr[B] + 1 - Pr[B]} \quad \text{from (1)} \\ &= \frac{21 \times 0.1}{1 + 20 \times 0.1} = \frac{2.1}{3} = 0.7 \end{aligned}$$

(c) [6 points] Fill in the blanks with options given below:

- | | | | | | |
|------------------------|-------------------|-------------------------------|------------------|-------------------------------|--------------------|
| (a) δ | (b) ϵ | (c) $1/\delta$ | (d) $1/\epsilon$ | (e) $1 - \delta$ | (f) $1 - \epsilon$ |
| (g) m | (h) n | (i) $\text{size}(\mathbf{C})$ | | (j) $\text{size}(\mathbf{H})$ | |
| (k) number of examples | (l) instance size | | | (m) computation time | |
| (n) linear | (o) polynomial | | | (p) exponential | |

A concept class \mathbf{C} defined over the instance space \mathbf{X} (with instances of length n) is *PAC learnable* by learner \mathbf{L} using a hypothesis space \mathbf{H} if

for all $f \in \frac{\mathbf{C}}{\{\mathbf{C} | \mathbf{H}\}}$

for all distributions \mathbf{D} over \mathbf{X} , and fixed $\delta, \epsilon \in [0, 1]$, given a sample of m examples sampled independently according to the distribution \mathbf{D} , the learner \mathbf{L} produces with a probability $\frac{\text{at least}}{\{\text{at least} | \text{at most} | \text{equal to}\}} \frac{1 - \delta}{\{\text{one of (a) to (f)}\}}$

a hypothesis $g \in \frac{\mathbf{H}}{\{\mathbf{C} | \mathbf{H}\}}$

with error ($\text{Error}_{\mathbf{D}} = \Pr_{\mathbf{D}}[f(x) \neq g(x)]$) $\frac{\text{at most}}{\{\text{at least} | \text{at most} | \text{equal to}\}} \frac{\epsilon}{\{\text{one of (a) to (f)}\}}$

where the $\frac{\text{number of examples}}{\{\text{one of (k) to (m)}\}}$ is $\frac{\text{polynomial}}{\{\text{one of (n) to (p)}\}}$ in

$\frac{n}{\{\text{four of (a) to (j)}\}}$, $\frac{1/\delta}{\{\text{four of (a) to (j)}\}}$, $\frac{1/\epsilon}{\{\text{four of (a) to (j)}\}}$, and $\frac{\text{size}(\mathbf{C})}{\{\text{four of (a) to (j)}\}}$.

Support Vector Machines [25 points]

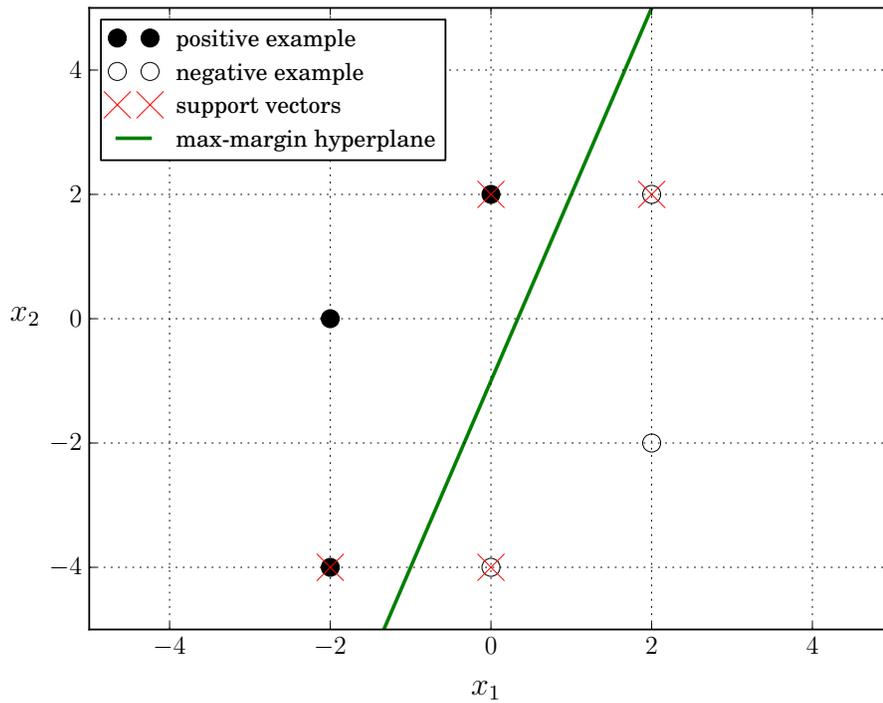
We are given the following set of training examples $D = \{(x_1^{(i)}, x_2^{(i)}, y^{(i)})\}, i = 1, \dots, m$, where $x_j^{(i)}$ are integer-valued features, and $y^{(i)}$ are binary labels.

x_1	x_2	y
-2	-4	+
-2	0	+
0	2	+
2	2	-
2	-2	-
0	-4	-

Our objective is to learn a hyperplane $w_1x_1 + w_2x_2 + b = 0$ using the hard-SVM objective:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} (w_1^2 + w_2^2) \\ & \text{subject to} && y^{(i)} (w_1x_1^{(i)} + w_2x_2^{(i)} + b) \geq 1, \quad i = 1, \dots, m. \end{aligned}$$

Use the grid below to answer the following questions (you will place a few points and lines on this grid).



(a) [10 points] Finding the hard-SVM hyperplane:

1. [2 points] Place the training examples on the grid, and indicate the support vectors.
See the figure.
2. [3 points] Draw the hyperplane produced by the hard-SVM on the grid.
See the figure.
3. [5 points] Find the values of $w_1, w_2, b \in \mathbb{R}$ that optimize the hard-SVM objective.

The support vectors are at $(-2, -4), (0, 2), (0, -4), (2, 2)$. At the support vectors, the constraints in the optimization of the hard-SVM will be an equality. This gives us the following system of equations:

$$\begin{aligned}+1(-2w_1 - 4w_2 + w_3) &= 1 \\+1(0w_1 + 2w_2 + w_3) &= 1 \\-1(2w_1 + 2w_2 + w_3) &= 1 \\-1(0w_1 - 4w_2 + w_3) &= 1.\end{aligned}$$

By solving this system of equations, we get $w_1 = -1, w_2 = \frac{1}{3}, w_3 = \frac{1}{3}$.

(b) [10 points] Experimental evaluation:

1. [2 points] Provide the classification rule used to classify an example with features x_1, x_2 , using the hyperplane produced by hard-SVM.

$$y = \begin{cases} +1 & \text{if } w_1x_1 + w_2x_2 + b \geq 0 \\ -1 & \text{otherwise.} \end{cases}$$

2. [2 points] What will the error of your classifier be on the training examples D (expressed as the fraction of training examples misclassified)?

0

3. [2 points] Draw on the grid, the hyperplane that will be produced by hard-SVM when you use all training examples except $a = (0, 2, +)$. Using this hyperplane, will you classify a correctly?

The hyperplane will be $x_1 = -1$. No, a will be misclassified.

4. [2 points] Draw on the grid, the hyperplane that will be produced by hard-SVM when you use all training examples except $b = (2, 2, -)$. Using this hyperplane, will you classify b correctly?

The hyperplane will be the same as the one you computed in (a). Yes, b will be classified correctly.

5. [2 points] What will be the average error if you use 6-fold cross validation on the training set D ?

The classifier will make an error when either the support vector at $(0, 2)$, or $(0, -4)$ is left out for validation. In these two cases, the error will be $\frac{1}{6}$, and in all other cases, the error will be 0. The average error will be $\frac{2}{6} \times \frac{1}{6} = \frac{1}{18}$

(c) [5 points] Soft-SVM formulation:

1. [3 points] Write the soft-SVM objective below. Circle either min or max.

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \max \left(0, 1 - y^{(i)}(w_1x_1^{(i)} + w_2x_2^{(i)} + b) \right)$$

2. [2 points] For what range of positive C values, will the hyperplane produced by this soft-SVM objective be most similar to the hyperplane produced by hard-SVM. Circle one of the following.

very small

moderate

very large

Probability [25 points]

You are given the following sample S of data points in order to learn a model. This question will use this data.

Example	A	B	C
1	1	1	0
2	0	1	1
3	1	0	0
4	0	0	0
5	1	1	0
6	0	0	0
7	1	0	1
8	0	1	1
9	1	1	0
10	0	0	0
11	1	1	1
12	0	0	0

(a) [**3 points**] What would be your estimate for the probability of the following data points, given the sample S , if you were not given any information on a model? (That is, you would estimate the probability directly from the data.)

1. $P(A = 1, B = 1, C = 0)$

We just count the number of times $A = 1, B = 1$, and $C = 0$ occurs in the dataset. By counting we get $P(A = 1, B = 1, C = 0) = \frac{3}{12} = 0.25$.

2. $P(A = 0, B = 1, C = 1)$

By counting we get $P(A = 0, B = 1, C = 1) = \frac{2}{12} = \frac{1}{6}$.

3. $P(A = 0, B = 0, C = 1)$

By counting we get $P(A = 0, B = 0, C = 1) = 0$.

- (b) [10 points] Consider the following graphical model M over three variables A, B , and C .

$$A \rightarrow B \rightarrow C$$

1. [5 points] What are the parameters you need to estimate in order to completely define the model M ? Circle these parameters from Table 1.

(1) $P[A = 1]$	(5) $P[B = 1]$	(9) $P[C = 1]$
(2) $P[A = 1 B = b] \quad b \in \{0, 1\}$	(6) $P[B = 1 C = c] \quad c \in \{0, 1\}$	(10) $P[C = 1 A = a] \quad a \in \{0, 1\}$
(3) $P[A = 1 C = c] \quad c \in \{0, 1\}$	(7) $P[B = 1 A = a] \quad a \in \{0, 1\}$	(11) $P[C = 1 B = b] \quad b \in \{0, 1\}$
(4) $P[A = 1 B, C = b, c] \quad b, c \in \{0, 1\}$	(8) $P[B = 1 A, C = a, c] \quad a, c \in \{0, 1\}$	(12) $P[C = 1 A, B = a, b] \quad a, b \in \{0, 1\}$

Table 1: Options to choose from to explain model M .

2. [5 points] Use the data to estimate the parameters you have circled in (b).1.

The parameters we need to estimate are $P_M[A = 1]$, $P_M[B = 1|A = 1]$, $P_M[B = 1|A = 0]$, $P_M[C = 1|B = 1]$ and $P_M[C = 1|B = 0]$. We estimate these parameters as:

$$P_M[A = 1] = \frac{1}{2} = 0.5$$

$$P_M[B = 1|A = 1] = \frac{4}{6} = \frac{2}{3}$$

$$P_M[B = 1|A = 0] = \frac{2}{6} = \frac{1}{3}$$

$$P_M[C = 1|B = 1] = \frac{3}{6} = 0.5$$

$$P_M[C = 1|B = 0] = \frac{2}{6} = \frac{1}{3}$$

(c) [6 points] Use the parameters chosen in (b).1 to write down expressions for the probabilities of the same data points according to model M and compute these probabilities using the estimated parameters.

1. $P_M(A = 1, B = 1, C = 0)$

Using the conditional independence structure of the graphical model M , we get $P_M(A = 1, B = 1, C = 0) = P_M[A = 1]P_M[B = 1|A = 1]P_M[C = 0|B = 1] = 0.5 * \frac{2}{3} * 0.5 = \frac{1}{6}$.

2. $P_M(A = 0, B = 1, C = 1)$

$P_M(A = 0, B = 1, C = 1) = P_M[A = 0]P_M[B = 1|A = 0]P_M[C = 1|B = 1] = 0.5 * \frac{1}{3} * 0.5 = \frac{1}{12}$.

3. $P_M(A = 0, B = 0, C = 1)$

$P_M(A = 0, B = 0, C = 1) = P_M[A = 0]P_M[B = 0|A = 0]P_M[C = 1|B = 0] = 0.5 * \frac{2}{3} * \frac{1}{3} = \frac{1}{9}$.

(d) [6 points] Use the parameters chosen in (b).1 to write down the expressions for the following probabilities for model M and compute these probabilities.

1. $P_M(B = 1)$

$P_M(B = 1) = P_M(B = 1|A = 1)P_M(A = 1) + P_M(B = 1|A = 0)P_M(A = 0) = 0.5\frac{2}{3} + 0.5\frac{1}{3} = 0.5$

2. $P_M(A = 1|B = 0)$

$P_M(A = 1|B = 0) = \frac{P_M(B=0|A=1)P_M(A=1)}{P_M(B=0)} = \frac{0.5 * \frac{1}{3}}{0.5} = \frac{1}{3}$

3. $P_M(A = 0|B = 0, C = 0)$

Since A is independent of C given B , so we have $P_M(A = 1|B = 0, C = 0) = P_M(A = 1|B = 0) = \frac{1}{3}$.
 $P_M(A = 0|B = 0) = 1 - P_M(A = 1|B = 0) = \frac{2}{3}$.

Naive Bayes [25 points]

You would like to study the effects of *irrigation*, *fertilization* and *pesticide* use with respect to the **yield** of a farm. Suppose you are provided with a collection $D = \{D_1, \dots, D_m\}$ of m data points corresponding to m different farms. Each farm has three binary attributes *IsIrrigated* (X_1), *IsFertilized* (X_2) and *UsesPesticide* (X_3), and each has either a high yield ($V = 1$) or a low yield ($V = 0$). The label is **Yield**. A natural model for this is the **multi-variate Bernoulli model**.

Below is a table representing a *specific collection* S of data points for 8 farms to illustrate how a collection might look like.

#	<i>IsIrrigated</i> (X_1)	<i>IsFertilized</i> (X_2)	<i>UsesPesticide</i> (X_3)	Yield (V)
1	No (0)	Yes (1)	No (0)	High (1)
2	Yes (1)	Yes (1)	No (0)	High (1)
3	No (0)	Yes (1)	No (0)	Low (0)
4	No (0)	Yes (1)	No (0)	High (1)
5	No (0)	No (0)	Yes (1)	Low (0)
6	Yes (1)	No (0)	Yes (1)	Low (0)
7	No (0)	No (0)	No (0)	Low (0)
8	No (0)	Yes (1)	No (0)	High (1)

- (a) [6 points] Circle *all* the parameters from the table below that you will need to estimate in order to completely define the model. You may assume that $i \in \{1, 2, 3\}$ for all entries in the table.

(1) $\alpha_i = \Pr(X_i = 1)$	(7) $\beta = \Pr(V = 1)$
(2) $\gamma_i = \Pr(X_i = 0)$	(8) $\varphi = \Pr(V = 0)$
(3) $p_i = \Pr(X_i = 1 \mid V = 1)$	(9) $q_i = \Pr(V = 1 \mid X_i = 1)$
(4) $r_i = \Pr(X_i = 0 \mid V = 1)$	(10) $s_i = \Pr(V = 0 \mid X_i = 1)$
(5) $t_i = \Pr(X_i = 1 \mid V = 0)$	(11) $u_i = \Pr(V = 1 \mid X_i = 0)$
(6) $w_i = \Pr(X_i = 0 \mid V = 0)$	(12) $y_i = \Pr(V = 0 \mid X_i = 0)$

We need (3), (4), (5), (6), (7), (8).

- (b) [3 points] How many **independent** parameters do you have to estimate to learn this model?

Each of X_1 , X_2 and X_3 are binary, so we need to learn $2 - 1 = 1$ independent parameters for each value of $V = v$ for the associated conditional distribution $\Pr(X_i \mid V = v)$. We also need to learn one parameter for $\Pr(V)$ (say $\Pr(V = 1)$) since V is binary also.

This means we need a total of $2 + 2 + 2 + 1 = 7$ independent parameters.

- (c) [5 points] Write explicitly the naïve Bayes classifier for this model as a function of the model parameters selected in (a):

$$\begin{aligned}
 & \Pr(V = v \mid X_1 = x_1, X_2 = x_2, X_3 = x_3) \\
 &= \frac{\Pr(V = v) \prod_{i=1}^3 \Pr(X_i = x_i \mid V = v)}{\Pr(X_1, X_2, X_3)} \\
 &\propto \Pr(V = v) \prod_{i=1}^3 \Pr(X_i = x_i \mid V = v) \\
 &= v \left(\beta \prod_{i=1}^3 p_i^{x_i} r_i^{(1-x_i)} \right) + (1 - v) \left(\varphi \prod_{i=1}^3 t_i^{x_i} w_i^{(1-x_i)} \right)
 \end{aligned}$$

- (d) [5 points] Write the expression for L , the log likelihood of the entire data set D , using the parameters that you have identified in (a).

Since we wrote the naïve Bayes classifier in (b), we can easily write

$$\begin{aligned}
 L = \sum_{j=1}^m & \left(v_j \left[\ln \beta + \sum_{i=1}^3 \left(x_i^{(j)} \ln p_i + (1 - x_i^{(j)}) \ln r_i \right) \right] + \right. \\
 & \left. (1 - v_j) \left[\ln \varphi + \sum_{i=1}^3 \left(x_i^{(j)} \ln t_i + (1 - x_i^{(j)}) \ln w_i \right) \right] \right)
 \end{aligned}$$

where $x_i^{(j)}$ refers to value of X_i for D_j .

- (e) [6 points] We would like to train a Naïve Bayes classifier on S to help us predict the yield on a new farm S_9 .

- [3 points] What is the decision rule for the Naïve Bayes classifier trained on S ?

$$v_{\text{NB}} = \arg \max_{v \in \{0,1\}} \Pr(V = v) \prod_{i=1}^3 \Pr(X_i = x_i \mid V = v)$$

- [3 points] Predict the yield for the following farm using the decision rule written earlier.

#	<i>IsIrrigated</i> (X_1)	<i>IsFertilized</i> (X_2)	<i>UsesPesticide</i> (X_3)	Yield (V)
9	Yes (1)	Yes (1)	Yes (1)	?

We observe that $\Pr(X_3 = 1 \mid V = 1) = 0$, so the value of the objective function for $V = 1$ is 0.

Computing for $V = 0$, we get $\frac{4}{8} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} > 0$. So by the decision rule we predict the yield of this farm to be Low.

Expectation Maximization [25 points]

Assume that a set of 3-dimensional points (x, y, z) is generated according to the following probabilistic generative model over Boolean variables $X, Y, Z \in \{0, 1\}$:

$$Y \leftarrow X \rightarrow Z$$

- (a) [4 points] What parameters from Table 2 will you need to estimate in order to completely define the model?

(1) $P(X=1)$	(2) $P(Y=1)$	(3) $P(Z=1)$	
(4) $P(X Y=b)$	(5) $P(X Z=b)$	(6) $P(Y X=b)$	(7) $P(Y Z=b)$
(8) $P(Z X=b)$	(9) $P(Z Y=b)$	(10) $P(X Y=b, Z=c)$	(11) 3

Table 2: Options to choose from. $b, c \in \{0, 1\}$.

Parameters needed are $P(X = 1) = \alpha$, $P(Y = 1|X = 1) = \gamma_{11}$, $P(Y = 1|X = 0) = \gamma_{10}$, $P(Z = 1|X = 1) = \delta_{11}$, $P(Z = 1|X = 0) = \delta_{10}$

- (b) [4 points] You are given a sample of m data points sampled independently at random. However, when the observations are given to you, the value of X is always omitted. Hence, you get to see $\{(y^1, z^1), \dots, (y^m, z^m)\}$. In order to estimate the parameters you identified in part (a), in the course of this question you will derive update rules for them via the EM algorithm for the given model.

Express $\Pr(y^j, z^j)$ for an observed sample (y^j, z^j) in terms of the unknown parameters.

$$\begin{aligned}
 P(y^j, z^j) &= P(y^j, z^j, x^j = 0) + P(y^j, z^j, x^j = 1) \\
 &= P(y^j, z^j | x^j = 0)P(x^j = 0) + P(y^j, z^j | x^j = 1)P(x^j = 1) \\
 &= P(y^j | x^j = 0)P(z^j | x^j = 0)P(x^j = 0) + \\
 &\quad P(y^j | x^j = 1)P(z^j | x^j = 1)P(x^j = 1) \\
 &= \gamma_{10}^{y^j} (1 - \gamma_{10})^{(1-y^j)} \delta_{10}^{z^j} (1 - \delta_{10})^{(1-z^j)} (1 - \alpha) + \\
 &\quad \gamma_{11}^{y^j} (1 - \gamma_{11})^{(1-y^j)} \delta_{11}^{z^j} (1 - \delta_{11})^{(1-z^j)} \alpha
 \end{aligned}$$

- (c) [4 points] Let $p_i^j = Pr(X = i | y^j, z^j)$ be the probability that hidden variable X has the value $i \in \{0, 1\}$ for an observation $(y^j, z^j), j \in \{1, \dots, m\}$. Express p_i^j in terms of the unknown parameters.

$$\begin{aligned}
p_0^j &= P(x^j = 0 | y^j, z^j) \\
&= \frac{P(y^j, z^j | x^j = 0) P(x^j = 0)}{P(y^j, z^j)} \\
&= \frac{P(y^j | x^j = 0) P(z^j | x^j = 0) P(x^j = 0)}{P(y^j | x^j = 0) P(z^j | x^j = 0) P(x^j = 0) + P(y^j | x^j = 1) P(z^j | x^j = 1) P(x^j = 1)} \\
&= \frac{(1-\alpha)\gamma_{10}y^j(1-\gamma_{10})^{(1-y^j)}\delta_{10}z^j(1-\delta_{10})^{(1-z^j)}}{(1-\alpha)\gamma_{10}y^j(1-\gamma_{10})^{(1-y^j)}\delta_{10}z^j(1-\delta_{10})^{(1-z^j)} + \alpha\gamma_{11}y^j(1-\gamma_{11})^{(1-y^j)}\delta_{11}z^j(1-\delta_{11})^{(1-z^j)}}
\end{aligned}$$

- (d) [4 points] Let (x^j, y^j, z^j) represent the completed j^{th} example, $j \in \{1, \dots, m\}$. Derive an expression for the expected log likelihood (LL) of the completed data set $\{(x^j, y^j, z^j)\}_{j=1}^m$, given the parameters in (a).

$$\begin{aligned}
E(LL) &= E\left(\sum_{j=1}^m \log P(x^j, y^j, z^j)\right) \\
&= \sum_{j=1}^m p_0^j \log P(x^j = 0, y^j, z^j) + p_1^j \log P(x^j = 1, y^j, z^j) \\
&= \sum_{j=1}^m p_0^j (\log P(y^j | x^j = 0) + \log P(z^j | x^j = 0) + \log P(x^j = 0)) + \\
&\quad \sum_{j=1}^m p_1^j (\log P(y^j | x^j = 1) + \log P(z^j | x^j = 1) + \log P(x^j = 1)) \\
&= \sum_{j=1}^m x_0^j y^j \log \gamma_{10} + x_0^j (1 - y^j) \log(1 - \gamma_{10}) + x_0^j z^j \log \delta_{10} + x_0^j (1 - z^j) \log(1 - \delta_{10}) - \\
&\quad x_1^j y^j \log \gamma_{11} + x_1^j (1 - y^j) \log(1 - \gamma_{11}) + x_1^j z^j \log \delta_{11} + x_1^j (1 - z^j) \log(1 - \delta_{11}) + \\
&\quad p_1^j \log \alpha + p_0^j \log(1 - \alpha)
\end{aligned}$$

- (e) [9 points] Maximize LL , and determine update rules for any two unknown parameters of your choice (from those you identified in part (a)). We will determine α and γ_{10} by differentiating $E(LL)$ by α and γ_{10} .

For α , we get

$$\begin{aligned}\sum_{j=1}^m \frac{p_1^j}{\alpha} - \frac{p_0^j}{1-\alpha} &= 0 \\ \alpha &= \sum_{j=1}^m \frac{p_1^j}{m}\end{aligned}$$

For γ_{10} , we get

$$\begin{aligned}\sum_{j=1}^m p_0^j y^j \frac{1}{\gamma_{10}} - p_0^j (1-y^j) \frac{1}{1-\gamma_{10}} &= 0 \\ \gamma_{10} &= \frac{\sum_{j=1}^m p_0^j y^j}{\sum_{j=1}^m p_0^j}\end{aligned}$$

Some formulae you may need

- $P(A, B) = P(A|B)P(B)$

- $Entropy(S) = -p^+ \log(p^+) - p^- \log(p^-) = -\sum_{i=1}^k p_i \log(p_i)$, where k is number of values

- $Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

- $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

- $\log_2(3) \approx \frac{3}{2}$

This page is intentionally left blank. You may use it as additional space for some of the solutions.