Fall 2019
CSC 591/791: HW 1
09/11/2019

This assignment contains 6 questions. Please read and follow the following instructions.

- **DUE DATE: Oct 4th, 11:45 PM**
- **TOTAL NUMBER OF POINTS: 135**
- **NO PARTIAL CREDIT** will be given so provide concise answers.
- Clearly list your team ID, each team member's **names and Unity IDs** at the top of your submission.
- Submit only a **single PDF file** per group, containing your answers.

---

1. (45 points) [Song Ju] [**Expectation Maximization**]
You are running a Naive Bayes classifier using 3 binary feature variables X_1, X_2, X_3 to predict the status of a nuclear power plant Y (0 for "Normal" and 1 for "Malfunction". Note that the actual status of the nuclear power plant Y cannot be directly observed but rather it can only be estimated through sensors such as the core temperatures. In this simplified scenario, let's assume there is only one sensor, Z with binary values: 0 for "Normal" vs. 1 for "Abnormal". One day you realize some of the sensor values Z are missing. Based on the nuclear power plant's manual book, the probability of missing values is much higher when $Y = 1$ than when $Y = 0$. More specifically, the exact values from the sensor specifications are:

$$P(Z\ missing|Y = 1) = .15; P(Z = 1|Y = 1) = .85$$
$$P(Z\ missing|Y = 0) = .03; P(Z = 0|Y = 0) = .97$$

   (a) (5 points) Draw a Bayes net that represents this problem with a node Y that is the unobserved label, a sensor node Z that is either a copy of Y or has the value "missing", and the three features X_1, X_2, X_3.
   Solution:

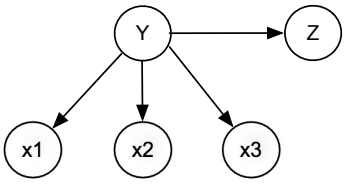


Figure 1: the Bayes Net.

(b) (5 points) What is the probability of $Y = 1$ given that our sensor data Z is missing, i.e., $P(Y = 1|Z =$ "$missing$"$)$? Derive your answer in terms of $\theta_{Y=1}$, denoting $P(Y = 1)$.

Solution:

$$
\begin{aligned}
&\mathbf{P(Y = 1|Z = missing)} \\
&= \frac{P(Y = 1, Z = missing)}{P(Z = missing)} \\
&= \frac{P(Z = missing|Y = 1) \cdot P(Y = 1)}{\sum_{Y \in \{0,1\}} P(Y, Z = missing)} \\
&= \frac{P(Z = missing|Y = 1) \cdot P(Y = 1)}{\sum_{Y \in \{0,1\}} P(Z = missing|Y) \cdot P(Y)} \\
&= \frac{P(Z = missing|Y = 1) \cdot P(Y = 1)}{P(Z = missing|Y = 0) \cdot P(Y = 0) + P(Z = missing|Y = 1) \cdot P(Y = 1)} \\
&= \frac{\mathbf{0.15} \cdot \theta_{\mathbf{Y=1}}}{\mathbf{0.03}(\mathbf{1} - \theta_{\mathbf{Y=1}}) + \mathbf{0.15} \cdot \theta_{\mathbf{Y=1}}}
\end{aligned}
$$

(c) (5 points) Based on the sensor data Z, we approximate the count and probabilities of each X_i variables under condition $Y = 1$, which are shown below.

$$
\begin{aligned}
Count(X_1 = 1|Y = 1) &= 20, \\
Count(X_2 = 1|Y = 1) &= 24, \\
Count(X_3 = 1|Y = 1) &= 4, \\
P(X_1 = 1|Y = 1) &= \frac{1}{2}, \\
P(X_2 = 1|Y = 1) &= \lambda, \\
P(X_3 = 1|Y = 1) &= 1 - 3\lambda,
\end{aligned}
$$

Assume $X_1|Y, X_2|Y, X_3|Y$ are all Bernoulli variables. Please apply log likelihood to calculate λ.

Solution:

$$
\begin{aligned}
&P(X_1, X_2, X_3|Y=1, \lambda) \\
&= (\frac{1}{2})^{Count(X_1=1|Y=1)} * (\lambda)^{Count(X_2=1|Y=1)} * (1-3\lambda)^{Count(X_3=1|Y=1)} \\
&= (\frac{1}{2})^{20} * (\lambda)^{24} * (1-3\lambda)^4
\end{aligned}
$$

$$
\begin{aligned}
&log(P(X_1, X_2, X_3|Y=1, \lambda)) \\
&= 20log(\frac{1}{2}) + 24log(\lambda) + 4log(1-3\lambda)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial logP}{\partial \lambda} = \frac{24}{\lambda} + \frac{-4*3}{1-3\lambda} = 0 \\
&\lambda = \frac{2}{7}
\end{aligned}
$$

(d) (15 points) Despite the approximation in part (c), we would like to theoretically learn the best choice of parameters for $P(Y), P(X_1|Y)$, $P(X_2|Y)$, and $P(X_3|Y)$. Assume $Y, X_1|Y, X_2|Y, X_3|Y$ are all Bernoulli variables and let us denote θ to include the following parameters:

$$
\begin{aligned}
\theta_{Y=y} &= P(Y=y), \\
\theta_{X_1=x_1|Y=y} &= P(X_1=x_1|Y=y), \\
\theta_{X_2=x_2|Y=y} &= P(X_2=x_2|Y=y) \\
\theta_{X_3=x_3|Y=y} &= P(X_3=x_3|Y=y).
\end{aligned}
$$

Write the log-likelihood of X, Y and Z given θ, in terms of θ and $P(Z|Y)$, first for a single example $(X_1=x_1, X_2=x_2, X_3=x_3, Z=z, Y=y)$, then for n i.i.d. examples $(X_1^i=x_1^i, X_2^i=x_2^i, X_3^i=x_3^i, Z^i=z^i, Y^i=y^i)$ for $i=1, ..., n$.

Solution:

From Figure 1, we know X and Z are independent with each other. And X_1, X_2, X_3 are mutually independent. The log probability of X, Y, Z give θ is:

$$
\begin{aligned}
\log P(X, Y, Z|\theta) &= \log[P(Z|Y) \cdot P(X|Y, \theta) \cdot P(Y|\theta)] \\
&= \log[P(Z|Y) \cdot P(X_1|Y, \theta) \cdot P(X_2|Y, \theta) \cdot P(X_3|Y, \theta) \cdot P(Y|\theta)] \\
&= \log P(Z|Y) + \log P(X_1|Y, \theta) + \log P(X_2|Y, \theta) + \log P(X_3|Y, \theta) + \log P(Y|\theta)
\end{aligned}
$$

For a specific single example $(X_1=x_1, X_2=X_2, X_3=x_3, Z=z, Y=y)$:

$$
\begin{aligned}
&\log P(X_1=x_1, X_2=X_2, X_3=x_3, Z=z, Y=y|\theta) \\
&= \log P(Z=z|Y=y) + \log \theta_{X_1=x_1|Y=y} + \log \theta_{X_2=x_2|Y=y} + \log \theta_{X_3=x_3|Y=y} + \log \theta_{Y=y}
\end{aligned}
$$

For n IID examples:

$$\log P(X, Y, Z|\theta) = \sum_{i=1}^{n} \{\log P(Z^i = z^i|Y^i = y^i) + \log \theta_{X_1=x_1^i|Y=y^i} + \log \theta_{X_2=x_2^i|Y=y^i} + \log \theta_{X_3=x_3^i|Y=y^i} + \log \theta_{Y=y^i}\}$$

(e) (15 points) Provide the E-step and M-step for performing expectation maximization of θ for this problem. For the $(t+1)$th iteration, in the E-step compute the distribution $Q_{t+1}(Y|Z, X)$ using

$$Q_{t+1}(Y = 1|Z, X) = E[Y|Z, X_1, X_2, X_3, \theta_t]$$

using your Bayes net from part (a) and conditional probability from part (b) for the unobserved class label Y of a single example.

In the M-step, compute:

$$\theta_{t+1} = argmax_\theta \sum_{i=1}^{n} \sum_{y} Q(Y^i = y|Z^i, X^i) logP(X_1^i, X_2^i, X_3^i, Y^i, Z^i|\theta_t)$$

using all of the examples $(X_1^1, X_2^1, X_3^1, Y^1, Z^1), ..., (X_1^n, X_2^n, X_3^n, Y^n, Z^n)$. Note: it is OK to leave your answers in terms of $Q(Y|Z, X)$.

Solution:

E Step (for a single example):

$$\begin{aligned} Q_{t+1}(Y^i = 1|Z^i, X^i) =& P(Y^i = 1|X_1^i = x_1^i, X_2^i = X_2^i, X_3^i = x_3^i, Z^i = z^i, \theta) \\ =& \frac{P(Z^i = z^i|Y^i = 1) \cdot \theta_{X_1=x_1^i|Y=1} \cdot \theta_{X_2=x_2^i|Y=1} \cdot \theta_{X_3=x_3^i|Y=1} \cdot \theta_{Y=1}}{\sum_{y\in\{0,1\}} P(Z^i = z^i|Y^i = y) \cdot \theta_{X_1=x_1^i|Y=y} \cdot \theta_{X_2=x_2^i|Y=y} \cdot \theta_{X_3=x_3^i|Y=y} \cdot \theta_{Y=y}} \\ Q_{t+1}(Y^i = 0|Z^i, X^i) =& 1 - Q_{t+1}(Y^i = 1|Z^i, X^i) \end{aligned}$$

M Step:

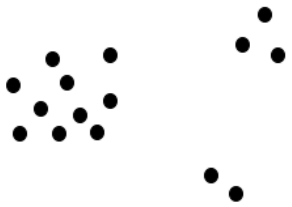
$$\theta_{Y=y} = \frac{\sum_{i=1}^{n} Q_{t+1}(Y^i = y|Z^i, X^i)}{n}$$

$$\theta_{X_1=x_1|Y=y} = \frac{\sum_{i=1}^{n} \delta(X_1^i = x_1) Q_{t+1}(Y^i = y|Z^i, X^i)}{\sum_{i=1}^{n} Q_{t+1}(Y^i = y|Z^i, X^i)}$$

$$\theta_{X_2=x_2|Y=y} = \frac{\sum_{i=1}^{n} \delta(X_2^i = x_2) Q_{t+1}(Y^i = y|Z^i, X^i)}{\sum_{i=1}^{n} Q_{t+1}(Y^i = y|Z^i, X^i)}$$

$$\theta_{X_3=x_3|Y=y} = \frac{\sum_{i=1}^{n} \delta(X_3^i = x_3) Q_{t+1}(Y^i = y|Z^i, X^i)}{\sum_{i=1}^{n} Q_{t+1}(Y^i = y|Z^i, X^i)}$$

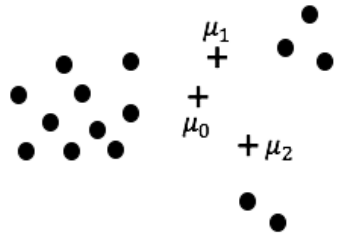
2. (20 points) [Song Ju] **Gaussian Mixture Model (GMM)**
Consider the set of training data in the graph below, let's assume it contains three clusters. For GMM, the means and variances of three Gaussians are μ_0 and σ_0, μ_1 and σ_1, and μ_2 and σ_2, respectively. Additionally, we have π_0, π_1, π_2 to denote the mixture proportions of the three Gaussians (i.e., $p(x) = \pi_0 N(\mu_0, \sigma_0 I) + \pi_1 N(\mu_1, \sigma_1 I) + \pi_2 N(\mu_2, \sigma_2 I)$), where I is the identity matrix and $\pi_0 + \pi_1 + \pi_2 = 1$. We will also use θ to refer to the entire collection of parameters $(\mu_0, \mu_1, \mu_2, \sigma_0, \sigma_1, \sigma_2, \pi_0, \pi_1, \pi_2)$ defining the mixture model $p(x)$.



(a) (10 points) Would K-Means ($K = 3$) and our 3-cluster GMM trained using EM produce the same cluster centers (means) for this data set above? Justify your answer. (Answer without any justification will get zero point.)

Solution: Either algorithm will find the clusters just fine. But the difference lies in that k-means uses hard assignment of each point to a single cluster, whereas GMM uses soft assignment, where every point has non-zero (though possibly small) probability of being in each cluster. So in k-means, the means of the clusters are determined by an average of the points assigned to that cluster, but in GMM the means of each cluster are (differently) weighted averages of all points. This has the effect of skewing the center of the left cluster to the right, the center of the right cluster to the left, the center of the bottom cluster to the top.

(b) (10 points) In the following, we apply EM to train our 3-cluster GMM on the data below. The '+' points indicate the current means μ_0, μ_1, and μ_2 of the three Gaussians after the kth iteration of EM.



(b.1) (3 points) On the figure, draw the directions in which μ_0, μ_1 and μ_2 will move in the next EM iteration.

Solution: μ_0 moves to the left, μ_1 moves to the right, and μ_2 moves to the bottom.

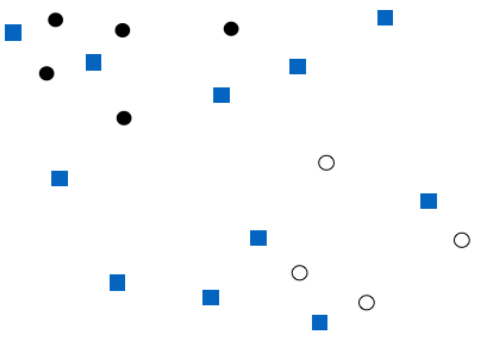
(b.2) (3 points) Will the marginal likelihood of the data, $\prod_j P(x^j|\theta)$ increase or decrease on the next EM iteration? Explain your reasoning.

Solution: Increase. Each iteration of the EM algorithm increases to likelihood of the data, unless you happen to be exactly at a local optimum.

(b.3) (4 points) Will the estimate of π_0 increase or decrease on the next EM iteration? Explain your reasoning.

Solution: Yes, π_0 will increase. π_0 is determined by adding the probabilities of all points that they are in cluster 1. When μ_0 moves to left, the probabilities of points in cluster 1 will increase but cluster 2 and 3 will decrease. Because more points are becoming closer to μ_0 and the increase is faster than decrease.

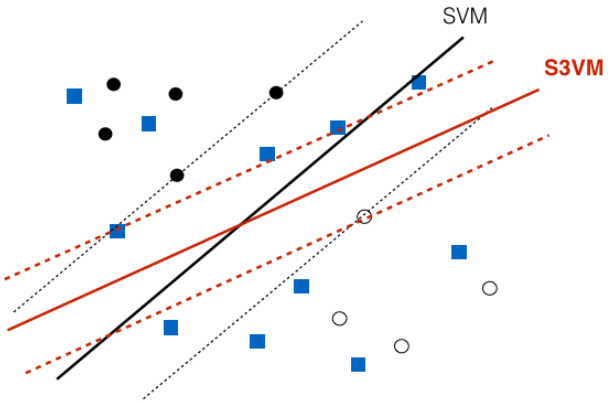
3. (26 points) [Farzaneh Khoshnevisan] **Semi-supervised Learning**
Consider the following figure which contains labeled (L) (class 1: black circles, class 2: hollow circles) and unlabeled (U) (blue squares) data. In this question, you will use two semi-supervised methods: S^3VM and co-training, to utilize the unlabeled data for further improvement of a SVM classifier.



(a) (10 points) Explain how would the semi-supervised SVM (S^3VM) perform on this data as compared to the supervised SVM, by plotting the separating hyper-planes

produced by both algorithms. For SVM, please draw marginal boundaries and separating hyper-plane by solid lines. Using a different color, draw the marginal boundaries and the separating hyper-plane of S^3VM using dash lines.

Solution: For supervised SVM, we only consider the labeled data and try to maximize the margins. The separating hyper-plane would split the margin into half (black line). For S^3VM, we start with SVM separator and greedily move it to low density area while maximizing margins for both labeled and unlabeled data points. The separating hyper-plane splits the width of the margin into half (red line).



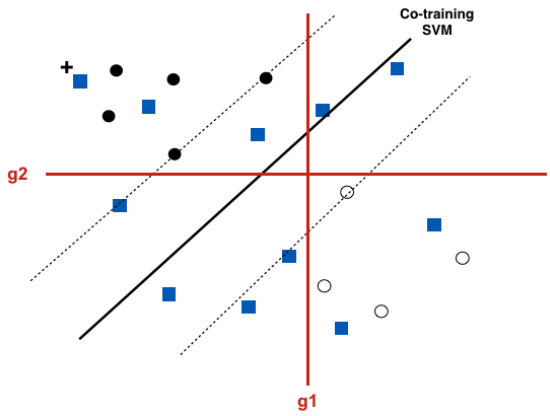
(b) (10 points) In applying co-training, g_1 and g_2 are SVM classifiers, and p_i/n_i represents the number of positive or negative points to label at iteration i, respectively. In this example, assume that $p_1, n_1 = 2$ and $p_2, n_2 = 1$.

(b.1) (5 points) Explain What is the main underlying assumption in applying co-training? How can we apply it to the above data (what are the two classifiers)?

Solution: Co-training partitions the feature space into two separate sets and its underlying assumption is for the two feature sets to be independent given the class. It also assumes that each set of feature are sufficient for classification. Here, the most natural way is to use one classifier (an SVM) for the x-axis and the second (another SVM) using the y-axis.

(b.2) (5 points) Identify the label of each point $x_i \in U$ and the final SVM separating hyper-plane after applying 2 iterations of semi-supervised co-training. Assume that at iteration i, p_i and n_i points are labeled based on the farthest distance from the separation hyper-plane.

Solution: In literature, there are multiple versions of co-training algorithm for semi-supervised learning. Here we stick to the pseudo-code illustrated in lecture notes. In this algorithm, only the intersection of p and n points are added to the set of labeled points L. In this example only one sample is labeled after 2 iterations. (other versions of co-training are also acceptable solutions.)



(c) (6 points) Compare the separating hyper-planes produced by S^{3}VM from part (a) and co-training from part (b.2). Explain why you think these two algorithms perform similarly/differently?
Solution: They perform differently because of multiple reasons: 1) in S^{3}VM all of the unlabeled data are considered for training while in co-traning algorithm, only a limited number of unlabeled data are contributing at a time. 2) the objective of S^{3}VM is to maximize the margin, while co-training in this example is based on the distance from separating hyper-plane. 3) The underlying assumption of independency of x and y given the class label might not hold in this example, thus resulting in less accurate classification.

4. (15 points) [Farzaneh Khoshnevisan] **Generalized Sequential Pattern**
Consider a data sequence:

$$S = (\{A, B\}\{B, C\}\{D, E\}\{A, D\}\{B, E, F\})$$

and the following time constraints:

- min_gap=0 (interval between last event in e_i and first event in e_{i+1} is > 0)
- max_gap=2 (interval between first event in e_i and last event in e_{i+1} is $<= 2$)
- max_span=5 (interval between first event in e_1 and last event in e_{last} is $<= 6$)
- $w_s = 1$ (time between first and last events in e_i is $<= 1$)

For each of the sequences $w = (e_1, ..., e_{last})$ below, determine whether they are subsequences of S, and if not, which constraint is excluding them.

(a) (3 points) $w = (\{A\}\{B\}\{C\}\{D\}\{E\})$
Solution: No, gap between $\{B\}\{C\}$ is 0, which violates min_gap criteria.

(b) (3 points) $w = (\{B\}\{D\}\{E\})$
Solution: Yes, if we choose middle D and last E.

(c) (3 points) $w = (\{D, E\}\{D, E\})$
Solution: Yes

(d) (3 points) $w = (\{A\}\{C, D, E\}\{A, F\})$
Solution: No, gap between C and F is 3, which violates max_gap constraint.

(e) (3 points) $w = (\{A, B, C, D\}\{E, F\})$
Solution: No, gap between A and D is 2, which violates w_s constraint.

5. (15 points) [Farzaneh Khoshnevisan] **Generalized Sequential Pattern**

Consider the following frequent 3-sequences:

$< \{1,2,3\} >, < \{1,3\}\{4\} >, < \{1\}\{2,4\} >, < \{1\}\{4\}\{5\} >, < \{2,3\}\{4\} >, < \{2,3\}\{5\} >$
$, < \{2,4\}\{4\} >, < \{2\}\{4,5\} >, < \{3\}\{4,5\} >, < \{3\}\{4\}\{5\} >, < \{4\}\{4,5\} >$.

(a) (5 points) List all the candidate 4-sequences produced by the candidate generation step of the GSP algorithm.

Solution:

$< \{1,2,3\} >, < \{1,3\}\{4\} >=< \{1,2,3\}\{4\} >,$
$< \{1,2,3\} >, < \{2,3\}\{5\} >=< \{1,2,3\}\{5\} >,$
$< \{1,3\}\{4\} >, < \{3\}\{4,5\} >=< \{1,3\}\{4,5\} >,$
$< \{1,3\}\{4\} >, < \{3\}\{4\}\{5\} >=< \{1,3\}\{4\}\{5\} >,$
$< \{1\}\{2,4\} >, < \{2,4\}\{4\} >=< \{1\}\{2,4\}\{4\} >,$
$< \{2,3\}\{4\} >, < \{3\}\{4,5\} >=< \{2,3\}\{4,5\} >,$
$< \{2,3\}\{4\} >, < \{3\}\{4\}\{5\} >=< \{2,3\}\{4\}\{5\} >,$
$< \{2,4\}\{4\} >, < \{4\}\{4,5\} >=< \{2,4\}\{4,5\} >.$

(b) (5 points) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints).

Solution:

$< \{1,2,3\}\{4\} >, < \{1,2,3\}\{5\} >, < \{1,3\}\{4,5\} >, < \{1,3\}\{4\}\{5\} >, < \{1\}\{2,4\}\{4\} >$
$, < \{2,3\}\{4\}\{5\} >, < \{2,4\}\{4,5\} >$

(c) (5 points) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming max_gap = 1).

Solution: In Candidate pruning with time constraints, we only need to verify contiguous k-1-sequences.
$< \{1,2,3\}\{4\} >, < \{1,2,3\}\{5\} >, < \{1,3\}\{4,5\} >, < \{1\}\{2,4\}\{4\} >, < \{2,3\}\{4\}\{5\} >$
$, < \{2,4\}\{4,5\} >$

6. (14 points) [Farzaneh Khoshnevisan] **Decision Theory**

Imagine you want to purchase a start-up company. With 5% chance the company's value will be up and you will make \$10,000K, and 95% chance that the company will fail and you will loose \$600K. Additionally, it would cost you \$200K to hire a group of experts for consultation who will help you determine how promising the company will be. This group of the experts are known to be accurate 75% of time.

Your goal is to maximize the expected value of your decision. What, if any, the best action(s) should you take, and what is your expected value? Assume that you are risk-neutral.

Draw a decision tree that supports your conclusion and show all the probabilities and utility values for every node.

Solution: You should hire a group of experts. If they said YES, then purchase it, otherwise do NOT purchase it. The expected value is U($31.44).