**CS446: Machine Learning** **Fall 2016**

# Final Exam

December $6^{th}$, 2016

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **four** problems. You need to solve all problems to get 100%.
- Please make sure that your exam booklet contains **20 pages.**
- The exam ends at 1:45 PM. You have 75 minutes to earn a total of 100 points.
- 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.**
- A list of potentially useful functions has been included in the **appendix at the back.**

**Good Luck!**

**Name (NetID):** (1 Point)

| | | |
|---|---|---|
| Naïve Bayes | | /25 |
| Expectation Maximization | | /25 |
| Multiclass Classification and Graphical Models | | /25 |
| Short Questions | | /24 |
| **Total** | | /100 |

**Naïve Bayes** [25 points]
In this question, we consider the problem of classifying Black Friday deals ($Y$) into two categories: valid deals ($A$), and scams ($B$).
For every deal, we have two attributes: **number of views** ($X_1$), and **time taken to receive 100 views** ($X_2$).
We assume that the number of views ($X_1$) is related to each category ($A$, $B$) via a *geometric distribution* with a category-specific parameter ($\theta^A$, $\theta^B$ resp.) and that the time taken to receive 100 views ($X_2$) is related to each category ($A$, $B$) via an *exponential distribution* with a category-specific parameter ($\lambda^A$, $\lambda^B$ resp.).
Also, $(\gamma^A, \gamma^B)$ are our prior beliefs for each of the categories (A, B), resp.
The summary of the model assumptions is given below:

$$Pr[Y = y] = \gamma^y \quad \forall y \in \{A, B\}$$

$$Pr[X_1 = x_1|Y = y] = (\theta^y)(1 - \theta^y)^{(x_1-1)} \quad \forall y \in \{A, B\}$$
$$Pr[X_2 = x_2|Y = y] = (\lambda^y)e^{-x_2\lambda^y} \quad \forall y \in \{A, B\}$$

(a) **[15 points]** Assume $\mathbf{D_A}$ to be the set of training instances with label A, and $\mathbf{D_B}$ to be the set of training instances with label B

i. **(5 points)** Under the given naïve Bayes assumption, and using the notation of $x_1^i$ and $x_2^i$ to represent the values of $X_1$ and $X_2$ respectively for the $i^{th}$ training instance, write down the expression for the log likelihood (LL) of the dataset.

$$LL = LL(\gamma^A) + LL(\gamma^B) + LL(\theta^A) + LL(\theta^B) + LL(\lambda^A) + LL(\lambda^B)$$

where,

$$LL(\gamma^A) = \sum_{i \in D_A} \Big( ln(\gamma^A) \Big)$$

$$LL(\gamma^B) = \sum_{i \in D_B} \Big( ln(\gamma^B) \Big)$$

$$LL(\theta^A) = \sum_{i \in D_A} \Big( (x_1^i - 1)ln(1 - \theta^A) + ln(\theta^A) \Big)$$

$$LL(\theta^B) = \sum_{i \in D_B} \Big( (x_1^i - 1)ln(1 - \theta^B) + ln(\theta^B) \Big)$$

$$LL(\lambda^A) = \sum_{i \in D_A} \Big( ln(\lambda^A) - x_2^i\lambda^A \Big)$$

$$LL(\lambda^B) = \sum_{i \in D_B} \Big( ln(\lambda^B) - x_2^i\lambda^B \Big)$$

ii. **(5 points)** Now, assume the following notation:

$$
\begin{aligned}
|D_A| &= n_A \\
|D_B| &= n_B \\
\sum_{i \in D_A} x_1^i &= f_A \\
\sum_{i \in D_B} x_1^i &= f_B \\
\sum_{i \in D_A} x_2^i &= g_A \\
\sum_{i \in D_B} x_2^i &= g_B
\end{aligned}
$$

Using this notation, **derive the expressions** for the MLE estimates of the parameters of your model.

- $\theta^A$, $\theta^B$:

  Setting the derivatives of $LL(\theta^A)$ and $LL(\theta^B)$ to zero, yields the following maximum likelihood estimates :-

$$
\begin{aligned}
\theta^A &= \frac{n_A}{f_A} \\
\theta^B &= \frac{n_B}{f_B}
\end{aligned}
$$

- $\lambda^A$, $\lambda^B$:

  Setting the derivatives of $LL(\lambda^A)$ and $LL(\lambda^B)$ to zero, yields the following maximum likelihood estimates :-

$$
\begin{aligned}
\lambda^A &= \frac{n_A}{g_A} \\
\lambda^B &= \frac{n_B}{g_B}
\end{aligned}
$$

- $\gamma^A$, $\gamma^B$:
  The estimates of $\gamma^A$ and $\gamma^B$ can be directly calculated by just counting the number of instances with each of the labels :-

$$\gamma^A = \frac{n_A}{n_A + n_B}$$
$$\gamma^B = \frac{n_B}{n_A + n_B}$$

iii. **(5 points)** Assume that the given data in Table 1 is generated by a naïve Bayes model. Use this data and your MLE expressions obtained above to compute the prior probabilities $(\gamma^A, \gamma^B)$ and parameter values $(\theta^A, \theta^B, \lambda^A, \lambda^B)$. That is, fill out Table 2. (Keep the solutions as fractions.)

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 2 | 12 | $A$ |
| 4 | 5 | $A$ |
| 3 | 7 | $A$ |
| 12 | 11 | $B$ |
| 1 | 1 | $B$ |
| 7 | 8 | $B$ |
| 12 | 4 | $B$ |

Table 1: Dataset for Poisson naïve Bayes

|  |  |
|---|---|
| $\gamma^A = 3/7$ | $\gamma^B = 4/7$ |
| $\theta^A = 1/3$ | $\theta^B = 1/8$ |
| $\lambda^A = 1/8$ | $\lambda^B = 1/6$ |

Table 2: Parameters for naïve Bayes

(b) **[5 points]** Derive an algebraic expression for the naïve Bayes predictor for $Y$ in terms of the parameters of the model.
That is, predict Y = A iff ________________________________

$$\alpha > 1, where$$

$$\text{Let } \alpha = \frac{\Pr(Y = A)\Pr(X_1, X_2|Y = A)}{\Pr(Y = B)\Pr(X_1, X_2|Y = B)}$$

$$\alpha = \frac{\left(\gamma^A\right)\left((\theta^A)(1-\theta^A)^{(x_1-1)}\right)\left((\lambda^A)e^{-x_2\lambda^A}\right)}{\left(\gamma^B\right)\left((\theta^B)(1-\theta^B)^{(x_1-1)}\right)\left((\lambda^B)e^{-x_2\lambda^B}\right)}$$

(c) **[3 points]** Based on the parameter values from Table 2, compute

$$\frac{\Pr(Y = A|X_1 = 2, X_2 = 16)}{\Pr(Y = B|X_1 = 2, X_2 = 16)}$$

use $16 \approx 24(\ln(2))$ to simplify your calculations

$$= \frac{\Pr(Y = A)\Pr(X_1 = 2, X_2 = 16|Y = A)}{\Pr(Y = B)\Pr(X_1 = 2, X_2 = 16|Y = B)}$$

$$= \left(\frac{\gamma^A}{\gamma^B}\right)\left(\frac{1-\theta^A}{1-\theta^B}\right)^{(x_1-1)}\left(\frac{\theta^A}{\theta^B}\right)\left(\frac{\lambda^A}{\lambda^B}\right)e^{-x_2(\lambda^A-\lambda^B)}$$

Solving above, we get :-

$$= \frac{16}{7}$$

(d) **[2 points]** What will the classifier predict as the value of $Y$, given the above data point i.e. $X_1 = 2, X_2 = 16$?

Y = A

**Expectation Maximization** [25 points]

Consider the following generative probabilistic model:

$$W \to X \leftarrow Z.$$

over the Boolean variables $W$, $X$, $Z$, where the data is generated according to:

- The variable $W$ is set to 1 with probability $\alpha$, and 0 with probability $1 - \alpha$.
- The variable $Z$ is set to 1 with probability $\beta$, and 0 with probability $1 - \beta$.
- If $(W, Z) = (1, 1)$ then $X = 1$ with probability $\lambda_{11}$
  If $(W, Z) = (0, 1)$ then $X = 1$ with probability $\lambda_{01}$
  If $(W, Z) = (1, 0)$ then $X = 1$ with probability $\lambda_{10}$
  If $(W, Z) = (0, 0)$ then $X = 1$ with probability $\lambda_{00}$

This information is summarized below.

$$\begin{aligned} P(W = 1) &= \alpha \\ P(Z = 1) &= \beta \\ P(X = 1|W = 1, Z = 1) &= \lambda_{11} \\ P(X = 1|W = 0, Z = 1) &= \lambda_{01} \\ P(X = 1|W = 1, Z = 0) &= \lambda_{10} \\ P(X = 1|W = 0, Z = 0) &= \lambda_{00} \end{aligned}$$

You need to estimate the parameters of this model. However, **one of the variables, $Z$, is not observed**. You are given a sample of $m$ data points:

$$\{(w^{(j)}, x^{(j)})|w, x \in \{0, 1\}\}_{j=1}^{m}$$

In order to estimate the parameters of the model, $\alpha$, $\beta$, $\lambda_{11}$, $\lambda_{01}$, $\lambda_{10}$, $\lambda_{00}$, you will derive update rules for them via the EM algorithm.

(a) **(3 points)** Choose the correct expression for $P(w^{(j)}, x^{(j)})$ in terms of the unknown parameters $\alpha$, $\beta$, $\lambda_{11}$, $\lambda_{01}$, $\lambda_{10}$, $\lambda_{00}$. (Circle one of the four options given below.)

i. $$\begin{aligned} P(w^{(j)}, x^{(j)}) = {} & (1-\beta)\,[\alpha\lambda_{11}^{x_j}(1-\lambda_{11})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}(1-\lambda_{01})^{1-x_j}]^{1-w_j} \\ & + \beta[\alpha\lambda_{10}^{x_j}(1-\lambda_{10})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{00}^{x_j}(1-\lambda_{00})^{1-x_j}]^{1-w_j} \end{aligned}$$

ii. $$\begin{aligned} P(w^{(j)}, x^{(j)}) = {} & \beta[\alpha\lambda_{11}^{x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}]^{1-w_j} \\ & + (1-\beta)[\alpha\lambda_{10}^{x_j}]^{w_j}[(1-\alpha)\lambda_{00}^{x_j}]^{1-w_j} \end{aligned}$$

iii. $$\begin{aligned} P(w^{(j)}, x^{(j)}) = {} & \beta[\alpha\lambda_{11}^{x_j}(1-\lambda_{11})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}(1-\lambda_{01})^{1-x_j}]^{1-w_j} \\ & + (1-\beta)[\alpha\lambda_{10}^{x_j}(1-\lambda_{10})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{00}^{x_j}(1-\lambda_{00})^{1-x_j}]^{1-w_j} \end{aligned}$$

iv. $$P(w^{(j)}, x^{(j)}) = \beta[\alpha\lambda_{11}^{x_j}(1-\lambda_{11})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}(1-\lambda_{01})^{1-x_j}]^{1-w_j}$$

Ans: iii

(b) **(4 points)** Let $f_z^j = P(Z = z|w^{(j)}, x^{(j)})$, the probability that the hidden variable $Z$ has value $z$. Choose the correct expression for $f_1^{(j)}$ in terms of the unknown parameters $\alpha$, $\beta$, $\lambda_{11}$, $\lambda_{01}$, $\lambda_{10}$, $\lambda_{00}$. (Circle one of the four options given below.)

i. $f_1^{(j)} = \dfrac{\beta[\alpha\lambda_{11}^{x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}]^{1-w_j}}{P(w^{(j)}, x^{(j)})}$

ii. $f_1^{(j)} = \dfrac{\beta[\alpha\lambda_{11}^{x_j}(1-\lambda_{11})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{01}^{x_j}(1-\lambda_{01})^{1-x_j}]^{1-w_j}}{P(w^{(j)}, x^{(j)})}$

iii. $f_1^{(j)} = \dfrac{(1-\beta)[\alpha\lambda_{10}^{x_j}]^{w_j}[(1-\alpha)\lambda_{00}^{x_j}]^{1-w_j}}{P(w^{(j)}, x^{(j)})}$

iv. $f_1^{(j)} = \dfrac{(1-\beta)[\alpha\lambda_{10}^{x_j}(1-\lambda_{10})^{1-x_j}]^{w_j}[(1-\alpha)\lambda_{00}^{x_j}(1-\lambda_{00})^{1-x_j}]^{1-w_j}}{P(w^{(j)}, x^{(j)})}$

Ans: ii

(c) **(10 points)** Choose the correct expression for the expected log likelihood (LL) of the entire dataset, $\{(w^{(1)}, x^{(1)}), (w^{(2)}, x^{(2)}), ..., (w^{(m)}, x^{(m)})\}$ given the new parameter estimates $\widetilde{\alpha}$, $\widetilde{\beta}$, $\widetilde{\lambda}_{11}$, $\widetilde{\lambda}_{01}$, $\widetilde{\lambda}_{10}$, $\widetilde{\lambda}_{00}$. (Circle one of the four options given below.)

i.
$$\begin{aligned} E[LL] = &\sum_{j=1}^{m} f_1^j \log\left(\beta \left[\alpha \lambda_{11}^{x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{01}^{x_j}\right]^{1-w_j}\right) \\ &+ \sum_{j=1}^{m} \left(1 - f_1^j\right) \log\left((1-\beta) \left[\alpha \lambda_{10}^{x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{00}^{x_j}\right]^{1-w_j}\right) \end{aligned}$$

ii.
$$\begin{aligned} E[LL] = &\sum_{j=1}^{m} f_1^j \log\left(\beta \left[\alpha \lambda_{11}^{x_j} \left(1 - \lambda_{11}\right)^{1-x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{01}^{x_j} \left(1 - \lambda_{01}\right)^{1-x_j}\right]^{1-w_j}\right) \\ &+ \sum_{j=1}^{m} \left(1 - f_1^j\right) \log\left((1-\beta) \left[\alpha \lambda_{10}^{x_j} \left(1 - \lambda_{10}\right)^{1-x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{00}^{x_j} \left(1 - \lambda_{00}\right)^{1-x_j}\right]^{1-w_j}\right) \end{aligned}$$

iii.
$$\begin{aligned} E[LL] = &\sum_{j=1}^{m} f_1^j \log\left((1-\beta) \left[\alpha \lambda_{11}^{x_j} \left(1 - \lambda_{11}\right)^{1-x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{01}^{x_j} \left(1 - \lambda_{01}\right)^{1-x_j}\right]^{1-w_j}\right) \\ &+ \sum_{j=1}^{m} \left(1 - f_1^j\right) \log\left(\beta \left[\alpha \lambda_{10}^{x_j} \left(1 - \lambda_{10}\right)^{1-x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{00}^{x_j} \left(1 - \lambda_{00}\right)^{1-x_j}\right]^{1-w_j}\right) \end{aligned}$$

iv.
$$E[LL] = \sum_{j=1}^{m} f_1^j \log\left(\beta \left[\alpha \lambda_{11}^{x_j} \left(1 - \lambda_{11}\right)^{1-x_j}\right]^{w_j} \left[(1-\alpha)\,\lambda_{01}^{x_j} \left(1 - \lambda_{01}\right)^{1-x_j}\right]^{1-w_j}\right)$$

Ans: ii

(d) **(8 points)** Maximize the LL and select the correct update rule for $\beta$ according to the EM algorithm. (Circle one of the four options given below.)

i. $\beta = \frac{\sum_{j=1}^{m} 1 - f_1^j}{m}$

ii. $\beta = \frac{\sum_{j=1}^{m} f_1^j}{m}$

iii. $\beta = m \sum_{j=1}^{m} f_1^j$

iv. $\beta = \sum_{j=1}^{m} 1 - f_1^j$

Ans: ii

**Multiclass Classification and Graphical Models** [25 points] The goal of this problem is to develop a model for a multiclass classification problem. Each data point consists of five binary features, $x = (x_1, \ldots, x_5) \subseteq \{0,1\}^5$, and is assigned one of four possible labels $y \in \{A, B, C, D\}$.

(a) (13 points) In this part we consider a discriminative learning approach.

i. **(1 point)** Learning using a discriminative approach can be viewed as estimating

$P(y|x)$
$\{P(x{,}y) \mid P(y|x) \mid P(x|y)\}$

ii. **(10 points)** You are now tasked with choosing a discriminative model for this problem. Given your machine learning expertise, you have already narrowed down your modeling choices to:

- one versus all (OvA),
- all versus all (AvA),
- a minimal size error correcting output code (ECOC), and
- and multiclass SVM (MSVM)

Furthermore, you plan on using linear classifiers of the form $h(\mathbf{x}) = \mathbf{1}[\mathbf{w} \cdot \mathbf{x} + \theta \geq 0])$ for every binary classification problem that arises from these models. The goal of this question is to determine the number of parameters required to represent its hypothesis.

Note that the number of parameters is the number of real-valued variables whose values you are choosing during the learning process; for example, a single linear classifier of the form mentioned before has 6 parameters, consisting of each of the five dimensions of $\mathbf{w}$ as well as $\theta$.

**Question:** What is the *total number of parameters* required to represent each of these four hypotheses for this problem? In each case, explain how you derive your results.

- one versus all (OvA):
  There are 4 binary classification problems (one for each class), each of which requires 6 parameters. Thus, the total number of parameters here is 24.

- all versus all (AvA):
  There are $\binom{4}{2} = 6$ binary classification problems (one for each pair of classes), each of which requires 6 parameters. Thus, the total number of parameters here is 36.

- a *minimal* size error correcting output code (ECOC) (that is, use the smallest number of hypotheses needed for ECOC in this case):
  Four classes can be represented using a two-bit binary code; this means there will be two binary classification problems, each of which requires 6 parameters. Thus, the total number of parameters here is 12.

- multiclass SVM (MSVM)
  Representing MSVM requires 4 vectors (one per class) of length 6 (for the features/bias). Thus, the total number of parameters here is 24.

iii. **(2 points)** Suppose you decide to use the minimal ECOC model. Briefly discuss any potential issues with using this model to solve the classification problem.

The model requiring the minimal number of parameters is the ECOC model. This model divides the data into two different binary classification problems; however, depending on how the codes are assigned to the classes (and the form the data takes), these problems may not be separable. In other words, choosing this model means we sacrifice *expressivity* in favor of having a smaller representation.

(b) (12 points) In this part we will consider a generative approach.

i. **(1 point)** Learning using a generative approach can be viewed as estimating

$P(x, y)$
$\{P(x,y) \mid P(y|x) \mid P(x|y)\}$

ii. **(4 points)** We model the problem using a Bayesian network. After some thought, you narrow down the candidate graphs to the following two choices:

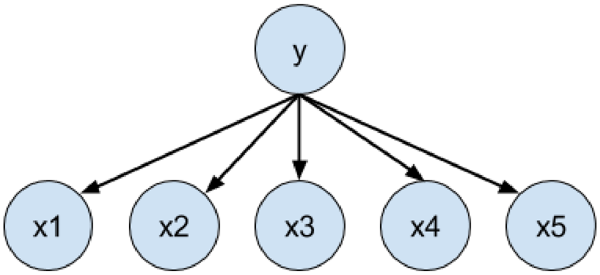


Figure 1: Option A

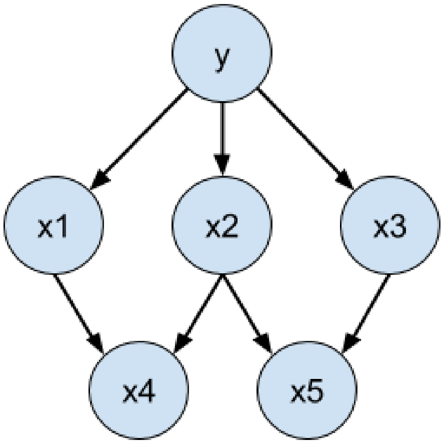


Figure 2: Option B

**Question:** Write down the factored joint probability distribution represented by each model.

- Model A:
  $p(y, x_1, x_2, x_3, x_4, x_5) = p(y)p(x_1|y)p(x_2|y)p(x_3|y)p(x_4|y)p(x_5|y)$

- Model B:
  $p(y, x_1, x_2, x_3, x_4, x_5) = p(y)p(x_1|y)p(x_2|y)p(x_3|y)p(x_4|x_1, x_2)p(x_5|x_2, x_3)$

iii. **(4 points)** We are interested in figuring out the *number of parameters* needed to represent each of the models in Figures 1 and 2 above. Note that in the context of graphical models, a parameter is a probability value for a given variable assignment (e.g. $\Pr(X_1 = 0, X_2 = 1 | X_3 = 1)$ is a single parameter). **Compute** the minimum number of parameters required to represent each model. Explain as needed.

- Model A: $p(y)$ requires 3 parameters (1 for each label, except the last can be computed in terms of the other three). Each feature requires 4 parameters (1 for each possible label for $y$, since the other probability value can be computed in terms of the first) Thus, a total of$3 + 5 \times 4 = 23$ parameters are required.

- Model B: $p(y)$ requires 3 parameters (same reason as before). $x_1$, $x_2$, and $x_3$ require 4 parameters each (same reason as before) $x_4$ and $x_5$ require 4 parameters each (since there are four possible assignments to their parents) Thus, a total of $3 + 5 \times 4 = 23$ parameters are required here as well.

iv. **(3 points)** After staring at your data for a few hours, you realize that the features $x_4$ and $x_5$ are not conditionally independent given the label $y$. Given this piece of information, which of the two Bayesian networks is a better choice for this problem? Explain your answer. Option B is better - Option A encodes that $x_4$ and $x_5$ are conditionally independent given the label, which the data implies is not true. However, Option B does not encode this assumption; therefore, (given no further information), it can better represent the data.

**Short Answer Questions** [24 points]

(a) **(8 points)** For the purpose of this question, consider the AdaBoost algorithm. Let $D_t$ be the probability distribution in the $t$th round of Adaboost, $h_t$ be the weak learning hypothesis learned in the $t$th round, and $\epsilon_t$ its error. Now, fill in the blanks to complete the algorithm:

$D_1(i) = 1/\mathrm{m}$

Given $D_t$ and $h_t$:

$D_{t+1}(i) = \underset{\{a \mid b \mid c \mid d\}}{\underline{\qquad \text{a} \qquad}}$ if $y_i \neg = h_t(x_i)$

$D_{t+1}(i) = \underset{\{a \mid b \mid c \mid d\}}{\underline{\qquad \text{c} \qquad}}$ if $y_i = h_t(x_i)$

where $z_t = \underset{\{e \mid f \mid g\}}{\underline{\qquad \text{f} \qquad}}$

and

where $\alpha_t = \underset{\{h \mid i\}}{\underline{\qquad \text{i} \qquad}}$

Options:
a.) $\frac{D_t(i)}{z_t} \times e^{\alpha_t}$ b.) $\frac{D_{t+1}(i)}{z_t} \times e^{\alpha_t}$ c.) $\frac{D_t(i)}{z_t} \times e^{-\alpha_t}$

d.) $\frac{D_{t+1}(i)}{z_t} \times e^{-\alpha_t}$ e.) $\sum_i D_t(i) \exp(\alpha_t y_i h_t(x_i))$

f.) $\sum_i D_t(i) \exp(-\alpha_t y_i h_t(x_i))$ g.) $\sum_t D_t(i) \exp(-\alpha_t y_i h_t(x_i))$

h.) $1/2\ ln\{\epsilon_t/(1-\epsilon_t)\}$ i.) $1/2\ ln\{(1-\epsilon_t)/\epsilon_t\}$

(b) **(8 points)** Given the instance space $X = \mathbb{R}^2$, consider the hypothesis class

$$\mathcal{H} = \{h(x_1, x_2) = (x_1 - a)^2 + (x_2 - b)^2 \leq r^2 : a, b \in \mathbb{R}, r \in \mathbb{R}_+\}$$

That is, each $h \in \mathcal{H}$ is a circle with radius $r$ and center $(a, b)$ whose interior is labeled as positive and whose exterior is labeled as negative.

Is $\mathcal{H}$ PAC learnable? Explain your answer. (It is sufficient to explain the structure of the argument, without getting to all the technical details.)

Yes. The hypothesis class is PAC learnable because its VC-Dimension is finite.

(c) **(8 points)** **[Support Vector Machine]**

Recall the objective function for soft SVM.

$$\mathbf{min}\ \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{s.t } y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + \theta) \geq 1 - \xi_i, \xi_i \geq 0, \forall(\mathbf{x}^{(i)}, y^{(i)}) \in D \tag{2}$$

where $m$ is the number of examples.

i. State whether the following statements about the SVM formulation above are *correct*. In each case, use one sentence to explain your answer (no need for a mathematical derivation or a proof).

A. When using the value of $C = 0$, we obtain the Hard-SVM objective.

______________________________

*Correct/Incorrect*

Reason: Incorrect. Choosing $C = 0$ leads to a trivial solution for $\vec{w} = 0$.

B. Choosing higher values of $C$ leads to over-fitting the training data.

______________________________

*Correct/Incorrect*

Reason: Correct. Higher value of C leads to more weight to not making mistakes on any training examples, which leads to over-fitting. Alternatively, regularization term gets lesser emphasis in the objective function.

C. The slack variable $\xi_i$ for a data point $x_i$ always takes the value 0 if the data point is correctly classified by the hyper-plane.

______________________

*Correct/Incorrect*

Reason: Incorrect. Data points classified correctly but lying within the margin will also have non-zero slack variable value.

D. The optimal weight vector $\vec{w}$ can be calculated as a linear combination of the training data points. [You need not prove this.]

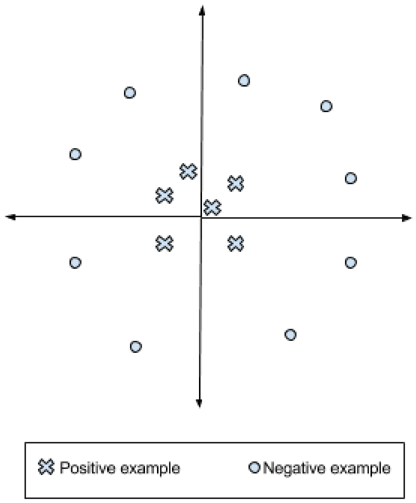
______________________

*Correct/Incorrect*

Reason:

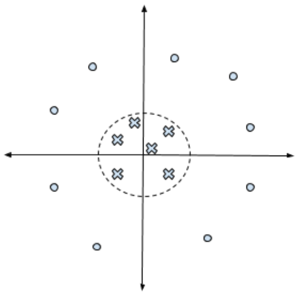
Correct. We can use the dual representation where weight vector $\vec{w} = \sum_i \alpha_i y_i x_i$ where $i$ iterates over each training data points.

ii. **Circles Dataset** Consider the following data set.

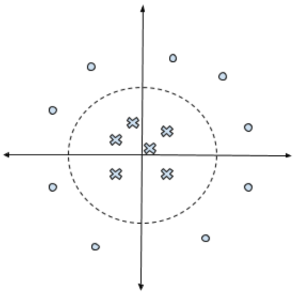


All data points inside a circle of some radius are marked as *positive (+1)* and points outside the circle are marked as *negative (-1)*.

A. Given the data set above that is separable by a circle, explain how Hard-SVM can be used to learn a valid separator in this case.
Basic idea is that we can use the dual form of SVM with a polynomial kernel of degree = 2 to learn a non-linear (may not be a circle necessarily) separator.



Separator (A)



Separator (B)

B. Which of the figures above is more likely to be the separator that would be learned by the Hard-SVM formulation? Justify your choice briefly.
Separator (B) is the more likely for a hard-SVM because SVM is a max-margin classifier.

**Some formulae you may need:**

(a) $\dfrac{d}{dx}e^x = e^x$

(b) $\dfrac{d}{dx}ln(x) = \dfrac{1}{x}$

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

(d) Let $p$ define the probability distribution of a *discrete random variable* $X$, then:

$\mathbb{E}_p[f(X)] = \sum_x p(X = x)f(x)$