CS446: Machine Learning **Fall 2012**

Final Exam

December 11^{th}, 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 $f(x_1, \ldots, x_n)$ 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.

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

(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?

(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) size(**C**) (j) size(**H**)
(k) number of examples (l) instance size (m) computation time
(n) linear (o) polynomial (p) exponential

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

for all $f \in$ __________
{**C** | **H**}

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

a hypothesis $g \in$ __________
{**C** | **H**}

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

where the ____________________ is ____________________ in
{one of (k) to (m)} {one of (n) to (p)}

_____, _____, _____, and _____ .
{four of (a) to (j)}

Support Vector Machines [25 points]

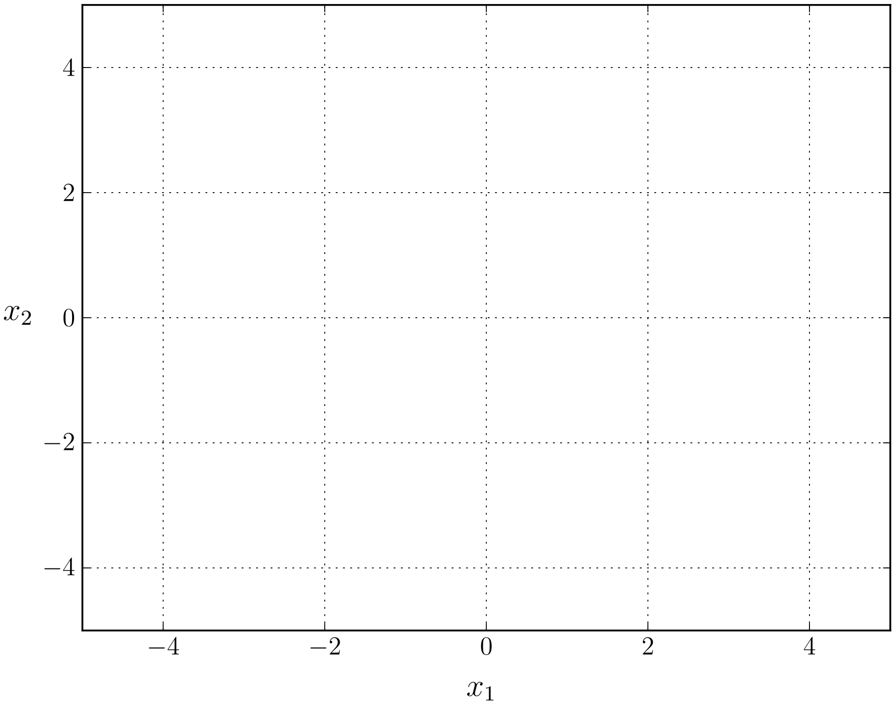
We are given the following set of training examples $D = \{(x_1^{(i)}, x_2^{(i)}, y^{(i)})\}, i = 1, \ldots, m$, where $x_j^{(i)}$ are integer-valued features, and $y^{(i)}$ are binary labels.

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

Our objective is to learn a hyperplane $w_1 x_1 + w_2 x_2 + b = 0$ using the hard-SVM objective:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}\left(w_1^2 + w_2^2\right) \\ \text{subject to} & y^{(i)}\left(w_1 x_1^{(i)} + w_2 x_2^{(i)} + b\right) \geq 1,\ i = 1, \ldots, m. \end{array}$$

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.

2. [**3 points**] Draw the hyperplane produced by the hard-SVM on the grid.

3. [**5 points**] Find the values of $w_1, w_2, b \in \mathbb{R}$ that optimize the hard-SVM objective.

(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.

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

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?

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?

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

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

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

$$\frac{\text{min}}{\text{max}}\underline{\hspace{6cm}} + C\underline{\hspace{6cm}}.$$

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

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

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

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

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.

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

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

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

(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. $P_M(B = 1)$

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

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

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 $D = \{D_1, \ldots, D_m\}$ 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.

#	*IsIrrigated* (X_1)	*IsFertilized* (X_2)	*UsesPesticide* (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) $y_i = \Pr(V = 0 \mid X_i = 0)$

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

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

$\Pr(V = v \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)$

$=$

(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).

(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.

1. [**3 points**] What is the decision rule for the Naïve Bayes classifier trained on S?

   $v_{\text{NB}} =$

2. [**3 points**] Predict the yield for the following farm using the decision rule written earlier.

#	*IsIrrigated* (X_1)	*IsFertilized* (X_2)	*UsesPesticide* (X_3)	**Yield** (V)
9	Yes (1)	Yes (1)	Yes (1)	?

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) $P(X \mid Y=b)$	(5) $P(X \mid Z=b)$	(6) $P(Y \mid X=b)$	(7) $P(Y \mid Z=b)$
(8) $P(Z \mid X=b)$	(9) $P(Z \mid Y=b)$	(10) $P(X \mid Y=b,Z=c)$	(11) 3

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

(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 $\{(y^1, z^1), \ldots, (y^m, z^m)\}$. 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.

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

(d) [**4 points**] Let (x^j, y^j, z^j) represent the completed j^{th} example, $j \in \{1, \ldots, m\}$. 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).

(e) **[9 points]** Maximize LL, and determine update rules for *any two* unknown parameters of your choice (from those you identified in part (a)).

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.