Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.S064 Introduction to Machine Learning

Final exam, Spring 2013

Your name & ID: _____

- This is a closed book exam
- You are permitted one sheet of typed or handwritten notes. Typed notes should be 11pt or larger.
- You do not need nor are permitted to use calculators or other electronic devices
- The value of each question – number of points awarded for full credit – is shown in parenthesis
- The problems are not necessarily in any order of difficulty. We recommend that you read through all the problems first, then do the problems in whatever order suits you best.
- Record all your answers in the places provided

Good luck!

Problem	1.1	2.1	3.1	4.1	5.1	Total
Max score	17	17	14	17	14	79
Your score						

(1.1) Consider a classification problem where we are given a training set of n examples and labels $S_n = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, n\}$, where $x^{(i)} \in \mathcal{R}^2$ and $y^{(i)} \in \{-1, 1\}$. Assume a different dataset for parts (a) and (b).

(a) **(5 points)** Suppose for a moment that we are able to find a linear classifier with parameters θ', θ'_0 such that $y^{(i)}(\theta' \cdot x^{(i)} + \theta'_0) > 0$, $i = 1, \ldots, n$. Let $\hat{\theta}, \hat{\theta}_0$ be the parameters of the maximum margin linear classifier, if it exists, obtained by *minimizing*

$$\frac{1}{2}\|\theta\|^2 \text{ subject to } y^{(i)}(\theta \cdot x^{(i)} + \theta_0) \geq 1, \; i = 1, \ldots, n \tag{1}$$

Which of the following statements are correct?

(X) Eq.(1) has a solution if and only if the training examples S_n are linearly separable

(X) The training examples S_n are linearly separable under our assumptions.

( ) $y^{(i)}(\theta' \cdot x^{(i)} + \theta'_0) \leq y^{(i)}(\hat{\theta} \cdot x^{(i)} + \hat{\theta}_0)$, for all $i = 1, \ldots, n$

( ) $y^{(i)}(\theta' \cdot x^{(i)} + \theta'_0) \geq y^{(i)}(\hat{\theta} \cdot x^{(i)} + \hat{\theta}_0)$, for all $i = 1, \ldots, n$

( ) $\|\theta'\| \geq \|\hat{\theta}\|$

Grading note: -2pts for each incorrect answer with a floor at 0pts

(b) **(12 points)** Here we estimate support vector machines based on a separate set of n training examples (see figures below). After trying out several methods, we generated six plots of $\hat{\theta} \cdot \phi(x) + \hat{\theta}_0 = 0$ (solid) and $\hat{\theta} \cdot \phi(x) + \hat{\theta}_0 = 1$ (dashed), $\hat{\theta} \cdot \phi(x) + \hat{\theta}_0 = -1$ (dashed), where $\hat{\theta}, \hat{\theta}_0$ are the estimated primal parameters. Each plot was generated by optimizing the kernel version. In other words, we *maximized*

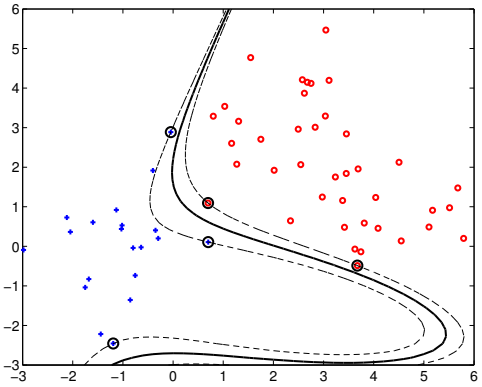
$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)}), \quad \text{subject to [constraints on } \alpha_i] \tag{2}$$

with respect to α_i, $i = 1, \ldots, n$. Each classifier was defined by a different choice of the kernel and the constraints:

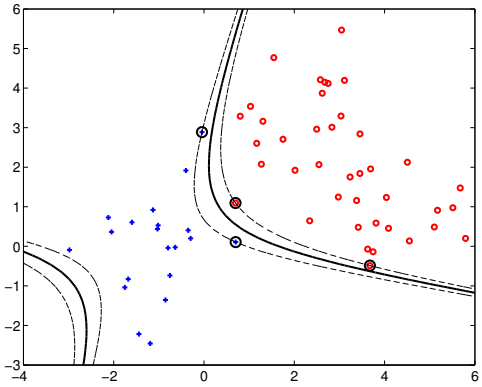
$(K1)$ $K_1(x, x') = (1 + x \cdot x'/2)$,
$(K2)$ $K_2(x, x') = (1 + x \cdot x'/2)^2$,
$(K3)$ $K_3(x, x') = (1 + x \cdot x'/2)^3$,
(Kg) $K_g(x, x') = \exp(-\|x - x'\|^2/2)$

$(C1)$ $0 \leq \alpha_i \leq 0.1$, $i = 1, \ldots, n$
$(C2)$ $\alpha_i \geq 0$, $i = 1, \ldots, n$

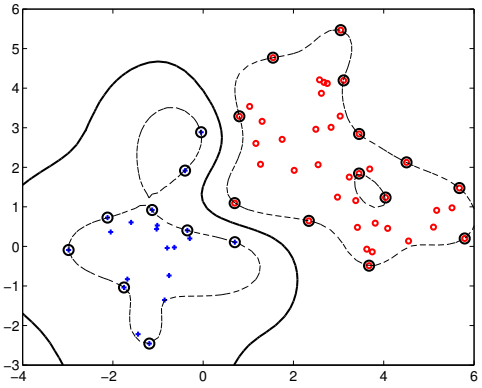
Under each figure, please write a pair of the kernel and the constraints, e.g., $(K1, C2)$ specifying the method that could have generated the plot. Each method can be assigned to *at most one plot*.



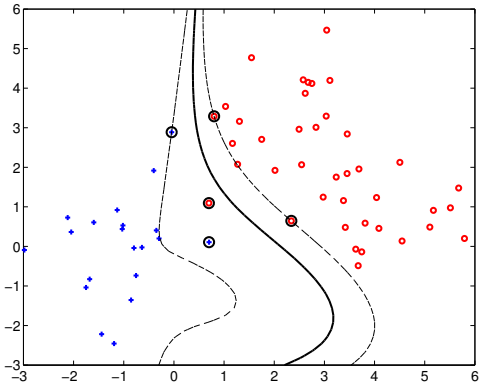
$(K, C) = (K3, C2)$



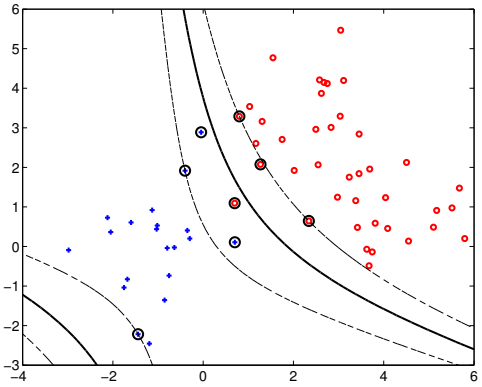
$(K, C) = (K2, C2)$



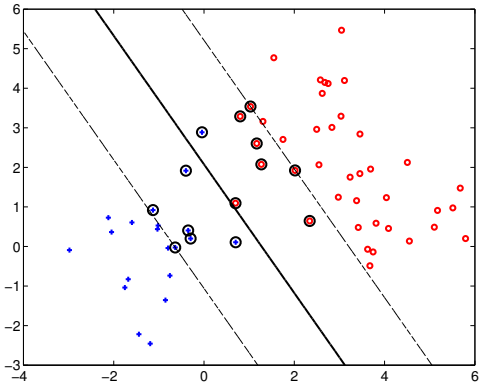
$(K, C) = (Kg, C2)$



$(K, C) = (K3, C1)$



$(K, C) = (K2, C1)$



$(K, C) = (K1, C1)$

(2.1) Consider estimating a mixture of two spherical Gaussians over points $x \in \mathcal{R}^2$. In other words, we consider

$$P(x;\theta) = \sum_{y=1}^{2} p_y N(x;\mu^{(y)},\sigma_y^2) = p_1 N(x;\mu^{(1)},\sigma_1^2) + p_2 N(x;\mu^{(2)},\sigma_2^2)$$

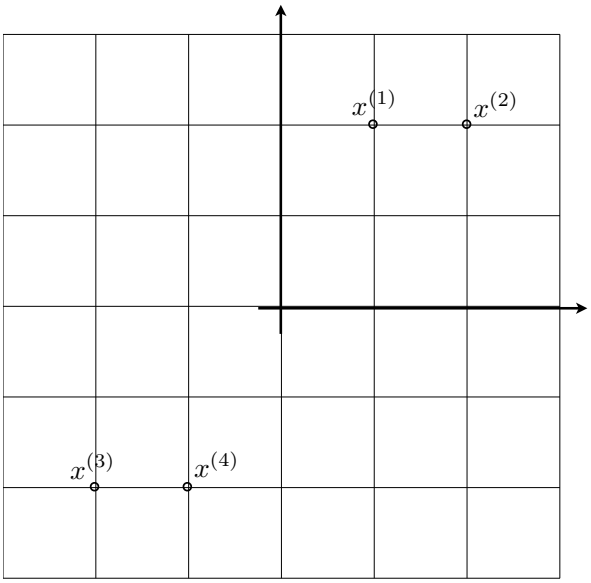
$$\text{where } N(x;\mu,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp\big(-\frac{1}{2\sigma^2}\|x-\mu\|^2\big)$$

We will use $y = 1, 2$ to refer to the mixture components.

(a) **(2 points)** If we had k components instead of just two, how many independent parameters would we have in the mixture model?

k-1 (mixing proportions) + k (variances) + 2k (2dim means) = 4k-1

Our data for estimating the mixture model is quite limited. In fact, we only have four points shown in the figure below: $x^{(1)} = (1,2)^T$, $x^{(2)} = (2,2)^T$, $x^{(3)} = (-2,-2)^T$, $x^{(4)} = (-1,-2)^T$.



We consider two different initializations of the mixture model

$$\text{initialization } \theta^A \quad p_1 = \frac{1}{4}, p_2 = \frac{3}{4}, \quad \mu^{(1)} = \mu^{(2)} = x^{(3)}, \quad \sigma_1^2 = \sigma_2^2 = 1$$

$$\text{initialization } \theta^B \quad p_1 = p_2 = \frac{1}{2}, \quad \mu^{(1)} = \mu^{(2)} = x^{(3)}, \quad \sigma_1^2 = 1, \sigma_2^2 = 4$$

(b) **(3 points)** Consider initialization θ^A. List all the points $x^{(i)}$, $i = 1, \ldots, 4$, for which $P(y = 2|x^{(i)}, \theta^A) > P(y = 1|x^{(i)}, \theta^A)$.

1,2,3,4 because of the higher prior probability for $y = 2$

(c) **(2 points)** Using posterior probabilities $p^A(y|i) = P(y|x^{(i)}, \theta^A)$, write down an expression for the mean $\mu^{(1)}$ after the first M-step of the EM algorithm

$$\hat{\mu}^{(1)} = \frac{\sum_{i=1}^{4} p^A(1|i) x^{(i)}}{\sum_{i=1}^{4} p^A(1|i)}$$

(d) **(4 points)** Assume initialization θ^A. Which of the following statements are true for the means $\mu^{(1)}$ and $\mu^{(2)}$ after the first M-step of the EM algorithm.

- ( ) $\mu^{(1)} = x^{(3)}$, $\mu^{(2)} = (\sum_{i=1}^{4} x^{(i)})/4$
- ( ) $\mu^{(1)} = x^{(3)}$, $\mu^{(2)} = (x^{(1)} + x^{(2)} + x^{(4)})/3$
- (X) $\mu^{(1)} = \mu^{(2)} = (0, 0)^T$
- ( ) $\mu^{(1)} = (x^{(3)} + x^{(4)})/2$, $\mu^{(2)} = (x^{(1)} + x^{(2)})/2$

(e) **(3 points)** Consider now initialization θ^B. List all the points $x^{(i)}$, $i = 1, \ldots, 4$, for which $P(y = 2|x^{(i)}, \theta^B) > P(y = 1|x^{(i)}, \theta^B)$.

points 1 and 2

(f) **(3 points)** Suppose we use initialization θ^B and run the EM algorithm until it converges. What are the resulting values of the means $\mu^{(1)}$ and $\mu^{(2)}$? (we are asking for approximate values that require no numerical calculation)

$$\hat{\mu}^{(1)} \approx (x^{(3)} + x^{(4)})/2 = (-1.5, -2)^T$$
$$\hat{\mu}^{(2)} \approx (x^{(1)} + x^{(2)})/2 = (1.5, 2)^T$$

(3.1) Given that the finals are approaching you are keen on monitoring your health. You always start in perfect health before the finals (state = 1), but the situation may change day by day. You may get a cold (state = 2), or something more serious (state = 3). Inspired by course material, you seek to model your health with an HMM. For this, you need to specify the initial state distribution π, state transition probabilities a, and the output distribution b where there are two possible output symbols, L=low or H=high energy. Based on past observations (and visits to the doctor's office), the parameters of this HMM are given by

$$\pi : \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad a : \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.3 & 0.6 & 0.1 \\ 0 & 0.5 & 0.5 \end{bmatrix} \end{array}, \quad b : \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} L & H \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0.5 & 0.5 \\ 1 & 0 \end{bmatrix} \end{array} \tag{3}$$

where, e.g., $a_{12} = P(Y_{t+1} = 2|Y_t = 1)$ and $b_{1H} = P(X_t = H|Y_t = 1)$.

(a) **(2 points)** What is the probability that you are well (state = 1) for the first two days but nevertheless feel low energy after the first day? In other words, what is $P(Y_1 = 1, Y_2 = 1, X_1 = H, X_2 = L)$?

probability = 0 (cannot generate L from $Y_2 = 1$)

(b) **(4 points)** What is the most likely sequence of states (Y_1, Y_2, Y_3) if you have high energy on the first day but low on the next two? In other words, what is the most likely hidden state sequence given $X_1 = H, X_2 = L, X_3 = L$?

Y_1 can only be 1, Y_2 can only be 2, Y_3 can be 2 or 3
$P(Y = (1, 2, 2), X = (H, L, L)) = \pi_1 \cdot b_{1H} \cdot a_{12} \cdot b_{2L} \cdot a_{22} \cdot b_{2L} = 1 \cdot 1 \cdot 0.5 \cdot 0.5 \cdot 0.6 \cdot 0.5$
$P(Y = (1, 2, 3), X = (H, L, L)) = \pi_1 \cdot b_{1H} \cdot a_{12} \cdot b_{2L} \cdot a_{23} \cdot b_{3L} = 1 \cdot 1 \cdot 0.5 \cdot 0.5 \cdot 0.1 \cdot 1$
so the most likely hidden state sequence is 1,2,2

(c) **(4 points)** What is the posterior probability over your health (state) on the 3rd day based on the same observations? In other words, what is $P(Y_3 = y|X_1 = H, X_2 = L, X_3 = L)$ for $y = 1, 2, 3$.

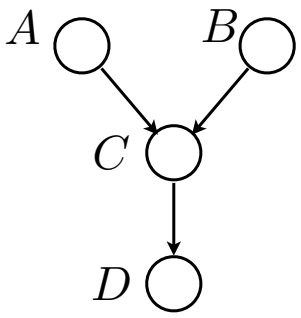
$P(Y_3 = 1|X = (H, L, L)) = 0$
$P(Y_3 = 2|X = (H, L, L)) = 3/4$
$P(Y_3 = 3|X = (H, L, L)) = 1/4$

(d) **(4 points)** What is the most likely hidden state sequence (Y_1, Y_2, Y_3, Y_4) if we observe $X_1 = H, X_2 = L, X_3 = L, X_4 = H$?

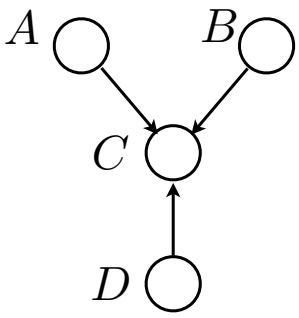
In this case there are only three possible hidden state sequences with non-zero probability: (1,2,2,1), (1,2,2,2), (1,2,3,2)

Most likely ones: (1,2,2,1) and (1,2,2,2)
(3pts for either, 4pts for giving both)

(4.1) Suppose we have four discrete random variables, A, B, C, and D corresponding to levels of traffic congestion in different locations. The random variables each take values in $\{1, 2, 3\}$ denoting low, medium, and high levels of congestion. We entertain here two alternative Bayesian network models over these variables. These are given as graphs G_1 and G_2 below.



graph G_1



graph G_2

(a) **(3 points)** List one independence property that holds for graph G_1 but NOT for graph G_2

Possible answers are
A independent of D given C
B independent of D given C
A,B independent of D given C
A independent of D given C,B
B independent of D given A,C

(b) **(3 points)** List one independence property that holds for graph G_2 but NOT for graph G_1

A independent of D
B independent of D

(c) **(4 points)** How many independent parameters do we need in order to specify the joint distribution associated with graph G_2?

$P(A)$, $P(B)$, $P(D)$: 2 parameters each, 6 in total
$P(C|A, B, D)$: $2 \cdot 3^3 = 54$ paramerers
total = 60

(d) **(3 points)** Suppose we are given a dataset consisting of n complete observations of the four variables (n days worth of traffic assessments). We find ML parameter estimates for each of our alternative models, G_1 and G_2. The resulting maximum log-likelihood values turned out to be practically the same. Hmmm... which model should we choose? Please check one.

(X) G_1 ( ) G_2 ( ) either/both

(e) **(4 points)** Which of the following rationales are correct for answering part (d)?

( ) Since the two models attain the same log-likelihood values, they should be considered equally good

(X) the BIC score for G_1 would be larger than for G_2

( ) the BIC score for G_2 would be larger than for G_1

( ) Because the two models make different independence assumptions about the variables, yet attain the same log-likelihood of the data, we cannot statistically decide between them.

Grading note: Some students may have seen BIC defined as -BIC, indicating that the correct answer would be the third one. Accept the third if you see writing to this effect.

(5.1) Consider the following MDP. It has states {0,1,2,3,4} with 4 as the starting state. In every state, you can take one of two possible actions: walk (W) or jump (J). The Walk action decreases the state by one. The Jump action has probability 0.5 of decreasing the state by two, and probability 0.5 of leaving the state unchanged. Actions will not decrease the state below zero: you will remain in state 0 no matter which action you will take (i.e., state 0 is a terminal state). Jumping in state 1 leads to state 0 with probability 0.5 and state 1 with probability 0.5. This definition leads to the following transition functions:

- For states $k \geq 1$, $T(k, W, k-1) = 1$
- For states $k \geq 2$, $T(k, J, k-2) = T(k, J, k) = 0.5$
- For state $k = 1$, $T(k, J, k-1) = T(k, J, k) = 0.5$

The reward gained when taking an action is the distance travelled squared: $R(s, a, s') = (s - s')^2$. The discount factor is $\gamma = 0.5$.

(a) **(4 points)** Suppose we initialize $Q_0(s, a) = 0$ for all $s \in \{0, 1, 2, 3, 4\}$ and $a \in \{J, W\}$. Evaluate the Q-values $Q_1(s, a)$ after exactly one Q-value iteration.

	s_0	s_1	s_2	s_3	s_4
J	0	0.5	2	2	2
W	0	1	1	1	1

(b) **(4 points)** What is the policy that we would derive from $Q_1(s, a)$?

s_1	s_2	s_3	s_4
W	J	J	J

(c) **(2 points)** What are the values $V_1(s)$ corresponding to $Q_1(s, a)$?

s_0	s_1	s_2	s_3	s_4
0	1	2	2	2

(d) **(4 points)** Will the policy change after the second iteration? If your answer is "yes", briefly describe how

No