10-601 Machine Learning, Midterm Exam

Instructors: Tom Mitchell, Ziv Bar-Joseph

Wednesday 12th December, 2012

There are 9 questions, for a total of 100 points.
This exam has 20 pages, make sure you have all pages before you begin.
This exam is open book, open notes, but *no computers or other electronic devices*.

This exam is challenging, but don't worry because we will grade on a curve. Work efficiently.

Good luck!

Name: _____

Andrew ID: _____

Question	Points	Score
Short Answers	11	
GMM - Gamma Mixture Model	10	
Decision trees and Hierarchical clustering	8	
D-separation	9	
HMM	12	
Markov Decision Process	12	
SVM	12	
Boosting	14	
Model Selection	12	
Total:	100	

Question 1. **Short Answers**

(a) [3 points] For data D and hypothesis H, say whether or not the following equations must always be true.

- $\sum_h P(H = h|D = d) = 1$ ... is this always true?

Solution:
yes

- $\sum_h P(D = d|H = h) = 1$ ... is this always true?

Solution:
no

- $\sum_h P(D = d|H = h)P(H = h) = 1$ ... is this always true?

Solution:
no

(b) [2 points] For the following equations, describe the relationship between them. Write one of four answers:

(1) "=" (2) "$\leq$" (3) "$\geq$" (4) "(depends)"

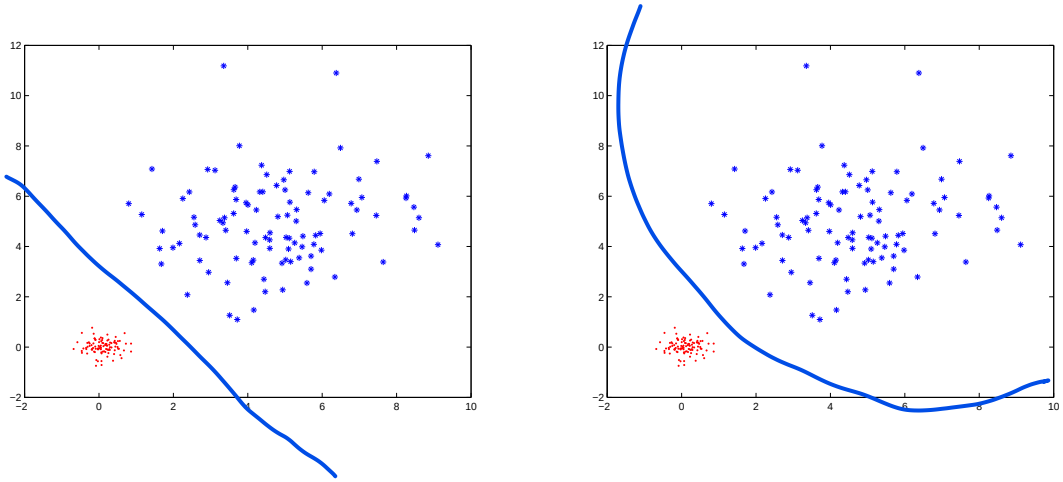
Choose the most specific relation that always holds; "(depends)" is the least specific. Assume all probabilities are non-zero.

$P(H = h\|D = d)$	$P(H = h)$
$P(H = h\|D = d)$	$P(D = d\|H = h)P(H = h)$

Solution:
$P(H|D)$ (DEPENDS) $P(H)$
$P(H|D) \geq P(D|H)P(H)$ .. this is the numerator in Bayes Rule, have to divide by the normalizer $P(D)$, which is less than 1. Tricky... $P(H|D) = P(D|H)P(H)/P(D) > P(D|H)P(H)$.

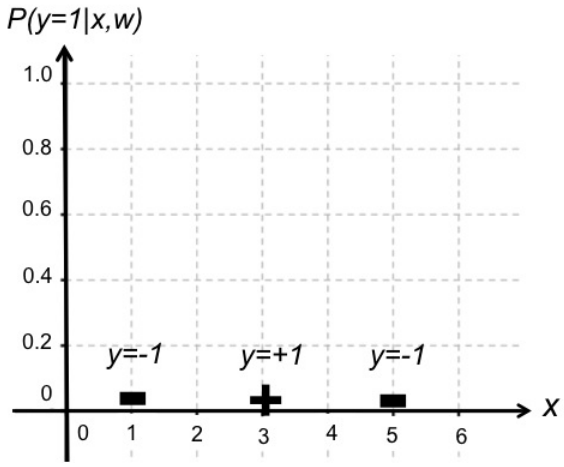
(c) [2 points] Suppose you are training Gaussian Naive Bayes (GNB) on the training set shown below. The dataset satisfies Gaussian Naive Bayes assumptions. Assume that the variance is independent of instances but dependent on classes, i.e. $\sigma_{ik} = \sigma_k$ where i indexes instances $X^{(i)}$ and $k \in 1, 2$ indexes classes. Draw the decision boundaries when you train GNB

a. using the **same** variance for both classes, $\sigma_1 = \sigma_2$
b. using separate variance for each class $\sigma_1 \neq \sigma_2$



Solution:
The decision boundary for part a will be linear, and part b will be quadratic.

(d) [2 points] Assume that we have two possible conditional distributions ($P(y = 1|x, w)$) obtained by training a logistic regression on the dataset shown in the figure below:

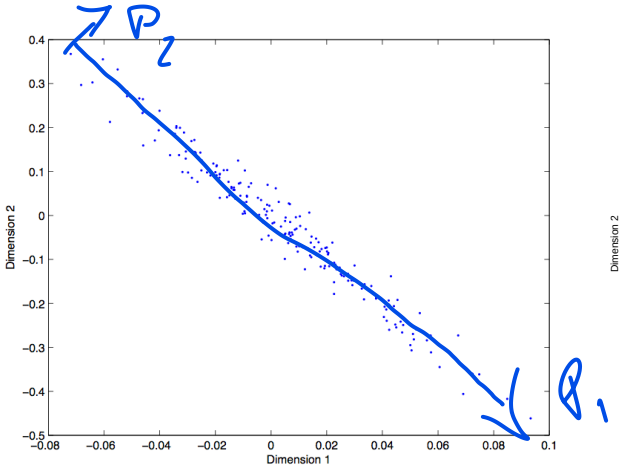


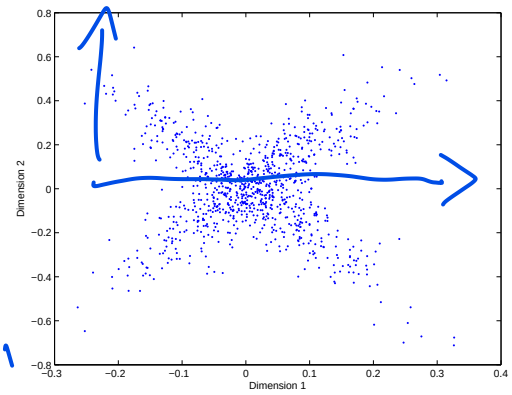
In the first case, the value of $P(y = 1|x, w)$ is equal to 1/3 for all the data points. In the second case, $P(y = 1|x, w)$ is equal to zero for $x = 1$ and is equal to 1 for all other data points. One of these conditional distributions is obtained by finding the maximum likelihood of the parameter w. Which one is the MLE solution? Justify your answer in at most three sentences.

↳ Sum whose $\frac{1}{N} = \frac{1}{3}$

Solution:
The MLE solution is the first case where the value of $P(y = 1|x, w)$ is equal to 1/3 for all the data points.

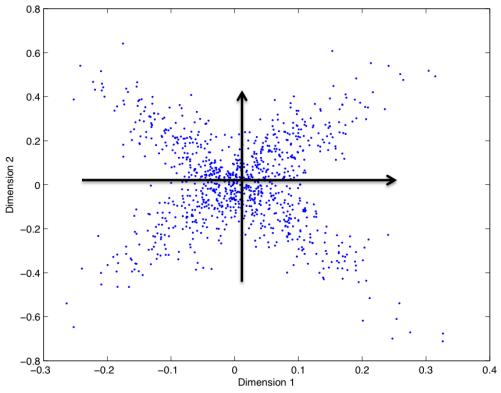
(e) [2 points] Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principle components on each plot.

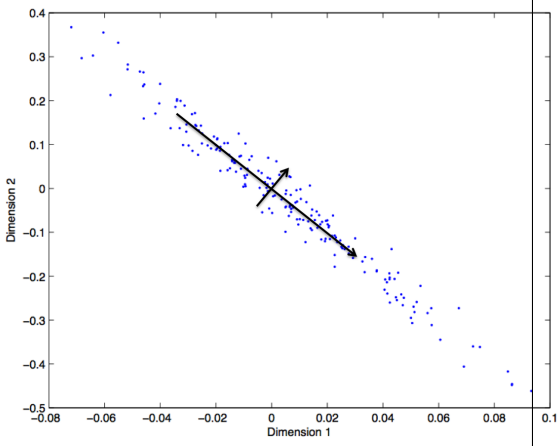




← Min reconstruction error

Solution:





Question 2. **GMM - Gamma Mixture Model**

A Assume each data point $X_i \in \mathbb{R}^+$ $(i = 1...n)$ is drawn from the following process:

$$Z_i \sim Multinomial(\pi_1, \pi_2, ..., \pi_K)$$

$$X_i \sim Gamma(2, \beta_{Z_i})$$

The probability density function of $Gamma(2, \beta)$ is $P(X = x) = \beta^2 x e^{-\beta x}$.

(a) [3 points] Assume $K = 3$ and $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 4$. What's $P(Z = 1|X = 1)$?

Solution:
$P(Z = 1|X = 1) \propto P(X = 1|Z = 1)P(Z = 1) = \pi_1 e^{-1}$
$P(Z = 2|X = 1) \propto P(X = 1|Z = 2)P(Z = 2) = \pi_2 4e^{-2}$
$P(Z = 3|X = 1) \propto P(X = 1|Z = 3)P(Z = 3) = \pi_3 16e^{-4}$

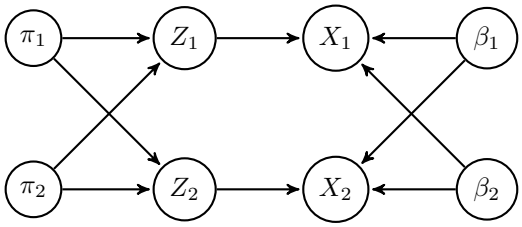
$P(Z = 1|X = 1) = \frac{\pi_1 e^{-1}}{(\pi_1 e^{-1} + \pi_2 4e^{-2} + \pi_3 16e^{-4})}$

(b) [3 points] Describe the E-step. Write an equation for each value being computed.

Solution:
For each $X = x$,

$$P(Z = k|X = x) = \frac{P(X = x|Z = k)P(Z = k)}{\sum_{k'} P(X = x|Z = k')P(Z = k')} = \frac{\beta_k^2 x e^{-\beta_k x}\pi_k}{\sum_{k'} \beta_{k'}^2 x e^{-\beta_{k'} x}\pi_{k'}}$$

(c) [2 points] Here's the Bayes net representation of the Gamma mixture model for $k = n = 2$. Note that we are treating π's and β's as variables – we have priors for them.



Would you say π's are independent given the observations X? Why?

Solution:
No. $\pi_1 \to Z_1 \to \pi_2$ is an active trail since X is given.

(d) For the following parts, choose true or false with an explanation in **one sentence**

i. [1 point] Gamma mixture model can capture overlapping clusters, like Gaussian mixture model.

Solution:
(All or none. 1 pt iff you get the answer and the explanation correct) true. in the e-step it does soft assignment

ii. [1 point] As you increase K, you will **always** get better likelihood of the data.

Solution:
(All or none. 1 pt iff you get the answer and the explanation correct) false. Won't improve after $K > N$

Question 3. **Decision trees and Hierarchical clustering**

Assume we are trying to learn a decision tree. Our input data consists of N samples, each with k attributes ($N \gg k$). We define the depth of a tree as the maximum number of nodes between the root and any of the leaf nodes (including the leaf, not the root).

(a) [2 points] If all attributes are binary, what is the maximal number of leaf (decision) nodes that we can have in a decision tree for this data? What is the maximal possible depth of a decision tree for this data?

Solution:
$2^{(k-1)}$. Each feature can only be used once in each path from root to leaf. The maximum depth is O(k).

(b) [2 points] If all attributes are continuous, what is the maximum number of leaf nodes that we can have in a decision tree for this data? What is the maximal possible depth for a decision tree for this data?

Solution:
Continuous values can be used multiple times, so the maximum number of leaf nodes can be the same as the number of samples, N and the maximal depth can also be N.

(c) [2 points] When using **single link** what is the maximal possible depth of a hierarchical clustering tree for the data in 1? What is the maximal possible depth of such a hierarchical clustering tree for the data in 2?

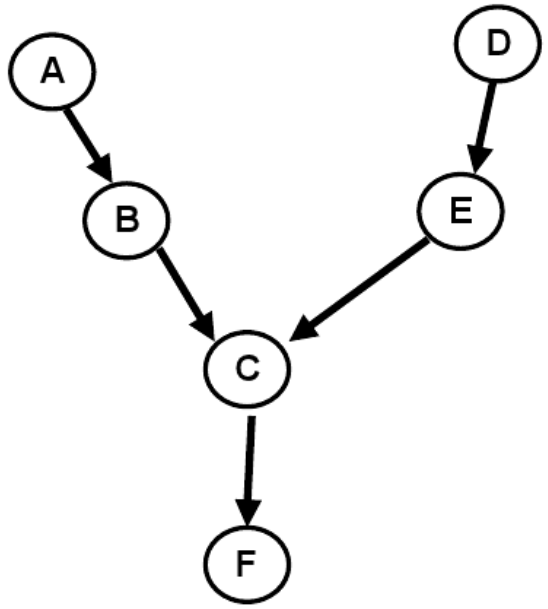
Solution:
When using single link with binary data, we can obtain cases where we are always growing the cluster by 1 node at a time leading to a tree of depth N. This is also clearly the case for continuous values.

(d) [2 points] Would your answers to (3) change if we were using **complete link** instead of **single link**? If so, would it change for both types of data? Briefly explain.

Solution:
While the answer for continuous values remain the same (its easy to design a dataset where each new sample is farther from any of the previous samples) for binary data, if k is small compared to N we will not be able to continue to add one node at a time to the initial cluster and so the depth will change to be lower than N.

Question 4. **D-separation**

Consider the following Bayesian network of 6 variables.



(a) [3 points] Set $X = \{B\}$ and $Y = \{E\}$. Specify two distinct (not-overlapping) sets Z such that: $X \perp\!\!\!\perp Y | Z$ (in other words, X is independent of Y given Z).

Solution:
$Z = \{A\}$ and $Z = \{D\}$

(b) [2 points] Can you find another distinct set for Z (i.e. a set that does not intersect with any of the sets listed in 1)?

Solution:
The empty set $Z = \{\}$

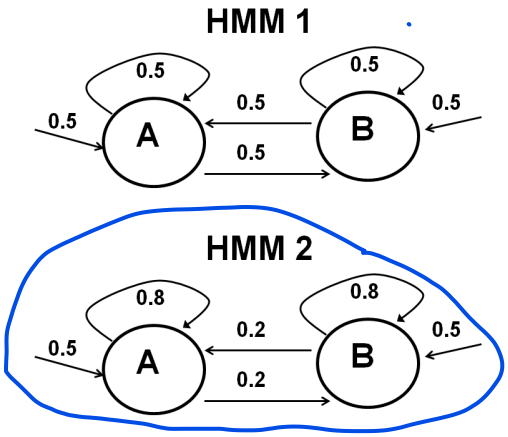
(c) [2 points] How many distinct Z sets can you find if we replace B with A while Y stays the same (in other words, now $X = \{A\}$ and $Y = \{E\}$)? What are they?

Solution:
$Z = \{\}$, $Z = \{B\}$ and $Z = \{D\}$

(d) [2 points] If $W \perp\!\!\!\perp X | Z$ and $X \perp\!\!\!\perp Y | Z$ for some distinct variables W, X, Y, Z, can you say $W \perp\!\!\!\perp Y | Z$? If so, show why. If not, find a counterexample from the graph above.

Solution:
No. $A \perp\!\!\!\perp F | B$ and $D \perp\!\!\!\perp A | B$ but D and F are not independent given B.

Question 5. **HMM**



The figure above presents two HMMs. States are represented by circles and transitions by edges. In both, emissions are deterministic and listed inside the states.

Transition probabilities and starting probabilities are listed next to the relevant edges. For example, in HMM 1 we have a probability of 0.5 to start with the state that emits A and a probability of 0.5 to transition to the state that emits B if we are now in the state that emits A.

In the questions below, O_{100}=A means that the 100th symbol emitted by the HMM is A.

(a) [3 points] What is $P(O_{100} = A, O_{101} = A, O_{102} = A)$ for HMM1?

Solution:
Note that P(O100=A, O101=A, O102=A) = P(O100=A, O101=A, O102=A,S100=A, S101=A, S102=A) since if we are not always in state A we will not be able to emit A. Given the Markov property this can be written as:

P(O100=A, O101=A, O102=A,S100=A, S101=A, S102=A) = P(O100=A|S100=A) P(S100=A) P(O101=A|S101=A) P(S101=A|S100=A) P(O102=A|S102=A) P(S102=A|S101=A)

The emission probabilities in the above equation are all 1. The transitions are all 0.5. So the only question is: What is P(S100=A)? Since the model is fully symmetric, the answer to this is 0.5 and so the total equation evaluates to: 0.5^3

(b) [3 points] What is $P(O_{100} = A, O_{101} = A, O_{102} = A)$ for HMM2?

Solution:
$0.5 * 0.8^2$

(c) [3 points] Let P_1 be: $P_1 = P(O_{100} = A, O_{101} = B, O_{102} = A, O_{103} = B)$ for HMM1 and let P_2 be: $P_2 = P(O_{100} = A, O_{101} = B, O_{102} = A, O_{103} = B)$ for HMM2. Choose the correct answer from the choices below and briefly explain.

1. $P_1 > P_2$
2. $P_2 > P_1$
3. $P_1 = P_2$
4. Impossible to tell the relationship between the two probabilities

Solution:
(a). P1 evaluates to 0.5^4 while P2 is $0.5 * 0.2^4$ so clearly $P_1 > P_2$.

(d) [3 points] Assume you are told that a casino has been using one of the two HMMs to generate streams of letters. You are also told that among the first 1000 letters emitted, 500 are As and 500 are Bs. Which is of the following answers is the most likely (briefly explain):

1. The casino has been using HMM 1
2. The casino has been using HMM 2
3. Impossible to tell

do not have to be in order

Solution:
(c). While we saw in the previous question that it is much more less likely to switch between A and B in HMM2, this is only true if we switch at every step. However, when aggregating over 1000 steps, since the two HMMs are both symmetric, both are likely to generate the *same* number of As and Bs.

Question 6. **Markov Decision Process**

Consider a robot that is moving in an environment. The goal of the robot is to move from an initial point to a destination point as fast as possible. However, the robot has the limitation that if it moves fast, its engine can overheat and stop the robot from moving. The robot can move with two different speeds: *slow* and *fast*. If it moves fast, it gets a reward of 10; if it moves slowly, it gets a reward of 4. We can model this problem as an MDP by having three states: *cool*, *warm*, and *off*. The transitions are shown in below. Assume that the discount factor is 0.9 and also assume that when we reach the state *off*, we remain there without getting any reward.

s	a	s'	$P(s'\|a,s)$
cool	slow	cool	1
cool	fast	cool	1/4
cool	fast	warm	3/4
warm	slow	cool	1/2
warm	slow	warm	1/2
warm	fast	warm	7/8
warm	fast	off	1/8

(a) [2 points] Consider the **conservative** policy when the robot always moves slowly. What is the value of $J^*(cool)$ under the conservative policy? Remember that $J^*(s)$ is the expected discounted sum of rewards when starting at state s

Solution:
$J^*(cool) = 4 + 0.9J^*(cool)$
$J^*(cool) = 40$

(b) [3 points] What is the optimal policy for each state?

Solution:
If in state *cool* then move *fast*. If in state *warm* then move *slow*.

(c) [2 points] Is it possible to change the discount factor to get a different optimal policy? If yes, give such a change so that it results to a **minimum** changes in the optimal policy and if no justify your answer in at most two sentences.

Solution:
Yes, by decreasing the discount factor. For example by choosing the discount factor equal to zero the robot always chooses an action that gives the highest immediate reward.

(d) [2 points] Is it possible to change the immediate reward function so that J^* changes but the optimal policy remains unchanged? If yes, give such a change and if no justify your answer in at most two sentences.

Also eg cool>fast>cool is $J^*(s) = R(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot J^*(s)$

$1/4 \cdot 10 = 2.5$

10

≥ 27.5

Solution:
Yes, for example by multiplying all the rewards by two.

(e) [3 points] One of the important problems in MDPs is to decide what should be the value of the discount factor. For now assume that we don't know the value of discount factor but an expert person tells us that action sequence {*fast, slow, slow*} is preferred to the action sequence { *slow, fast, fast*} if we start from either of states *cool* or *warm*. What does it tell us about the discount factor? What ranges of discount factor is consistent with this preference?

Solution:
The discounted sum of future rewards using discount factor λ is calculated by: $r + r(\lambda) + r(\lambda^2) + \ldots$.
So by solving the below equation, we would be able to find a range for discount factor λ:
$10 + 4\lambda + 4\lambda^2 > 4 + 10\lambda + 10\lambda^2$

↳ actually needs 7/8 probs inside

$10 + \frac{7}{8}(4\lambda + 4\lambda^2) > 4 + 10\lambda + \frac{7}{8}10\lambda^2$

Question 7. **SVM**

(a) Kernels

i. [4 points] In class we learnt that SVM can be used to classify linearly inseparable data by transforming it to a higher dimensional space with a kernel $K(x,z) = \phi(x)^T\phi(z)$, where $\phi(x)$ is a feature mapping. Let K_1 and K_2 be $R^n \times R^n$ kernels, K_3 be a $R^d \times R^d$ kernel and $c \in R^+$ be a positive constant. $\phi_1 : R^n \to R^d$, $\phi_2 : R^n \to R^d$, and $\phi_3 : R^d \to R^d$ are feature mappings of K_1, K_2 and K_3 respectively. Explain how to use ϕ_1 and ϕ_2 to obtain the following kernels.

a. $K(x,z) = cK_1(x,z)$

b. $K(x,z) = K_1(x,z)K_2(x,z)$

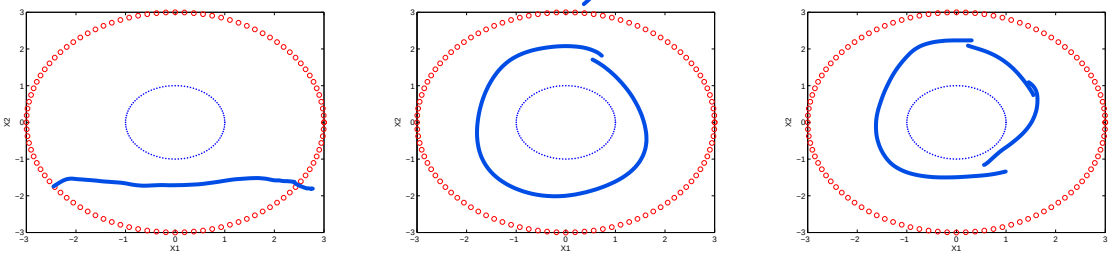
Solution:
a. $\phi(x) = \sqrt{(c)}\phi_1(x)$
b. $\phi(x) = \phi_1(x)\phi_2(x)$

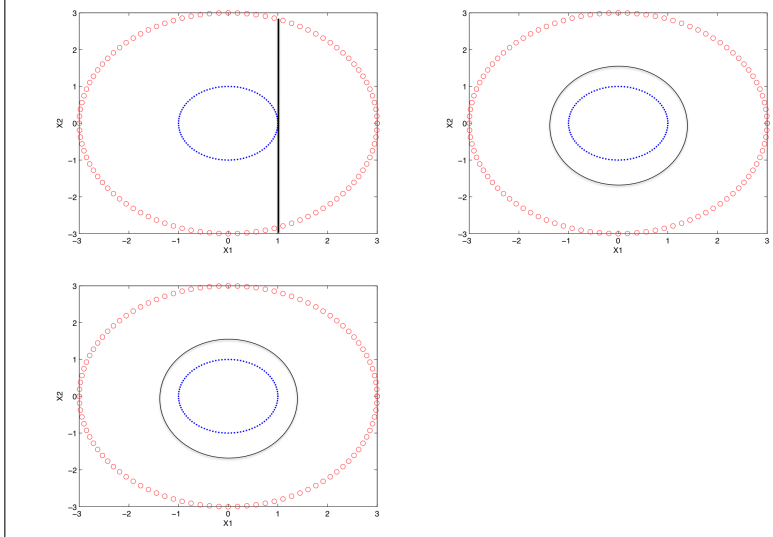
ii. [2 points] One of the most commonly used kernels in SVM is the Gaussian RBF kernel: $k(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma}\right)$. Suppose we have three points, z_1, z_2, and x. z_1 is geometrically very close to x, and z_2 is geometrically far away from x. What is the value of $k(z_1, x)$ and $k(z_2, x)$?. Choose one of the following:
a. $k(z_1, x)$ will be close to 1 and $k(z_2, x)$ will be close to 0.
b. $k(z_1, x)$ will be close to 0 and $k(z_2, x)$ will be close to 1.
c. $k(z_1, x)$ will be close to c_1, $c_1 \gg 1$ and $k(z_2, x)$ will be close to c_2, $c_2 \ll 0$, where $c_1, c_2 \in R$
d. $k(z_1, x)$ will be close to c_1, $c_1 \ll 0$ and $k(z_2, x)$ will be close to c_2, $c_2 \gg 1$, where $c_1, c_2 \in R$

Solution:
Correct answer is a, RBF kernel generates a "bump" around the center x. For points z_1 close to the center of the bump, $K(z_1, x)$ will be close to 1, for points away from the center of the bump $K(z_2, x)$ will be close to 0.

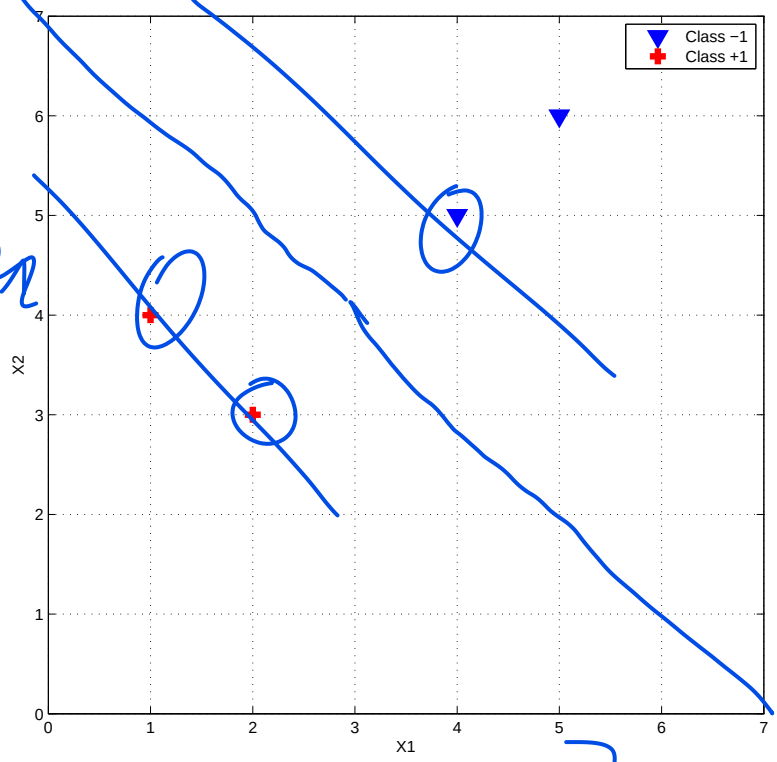
iii. [3 points] You are given the following 3 plots, which illustrates a dataset with two classes. Draw the decision boundary when you train an SVM classifier with linear, polynomial (order 2) and RBF kernels respectively. Classes have equal number of instances.

Solution:


X1
X2


X1
X2

(b) [3 points] Hard Margin SVM



Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 4 points shown in Figure 2. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles).

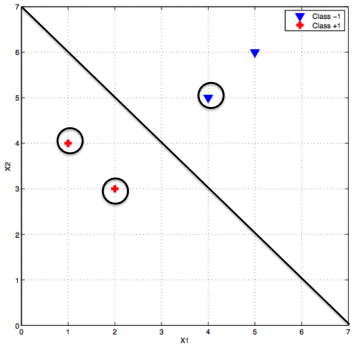
i. Find the weight vector **w** and bias b. What's the equation corresponding to the decision boundary?

Solution:
SVM tries to maximize the margin between two classes. Therefore, the optimal decision boundary is diagonal and it crosses the point (3,4). It is perpendicular to the line between support vectors (4,5) and (2,3), hence it is slope is m = -1. Thus the line equation is $(x_2 - 4) = -1(x_1 - 3) = x_1 + x_2 = 7$. From this equation, we can deduce that the weight vector has to be of the form (w_1, w_2), where $w_1 = w_2$. It also has to satisfy the following equations:
$2w_1 + 3w_2 + b = 1$ and
$4w_1 + 5w_2 + b = -1$

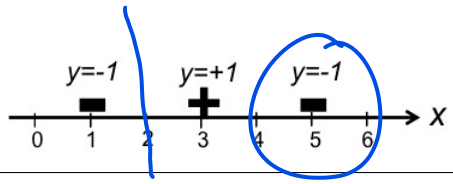
Hence $w_1 = w_2 = -1/2$ and $b = 7/2$

ii. Circle the support vectors and draw the decision boundary.

Solution:



Question 8. **Boosting**



Solution:

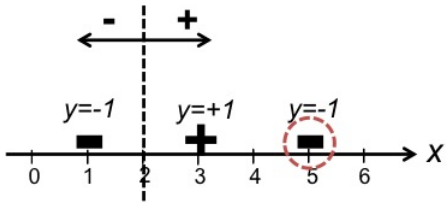


Figure 1: Sample training data for boosting algorithm.

In this problem, we study how boosting algorithm performs on a very simple classification problem shown in Figure 1. We use decision stump for each weak hypothesis h_i. Decision stump classifier chooses a constant value c and classifies all points where $x > c$ as one class and other points where $x \leq c$ as the other class.

(a) [2 points] What is the initial weight that is assigned to each data point?

Solution:
$\frac{1}{3}$

(b) [2 points] Show the decision boundary for the first decision stump (indicate the positive and negative side of the decision boundary).

Solution:
One possible solution is shown in the figure.

(c) [3 points] Circle the point whose weight increases in the boosting process.

Solution:
One possible solution is shown in the figure.

(d) [3 points] Write down the weight that is assigned to each data point after the first iteration of boosting algorithm.

Solution:
$\epsilon_t = \frac{1}{3}$
$\alpha_t = \frac{1}{2}\ln(2) = 0.3465$
For data points that are classified correctly $D_2(i) = \frac{1/3*\exp(-0.3465)}{Z_2} \approx 0.25$ and for the data point that is classified incorrectly $D_2(i) = \frac{1/3*\exp(0.3465)}{Z_2} \approx 0.5$ where Z_2 is the normalization factor.

(e) [3 points] Can boosting algorithm perfectly classify all the training examples? If no, briefly explain why. If yes, what is the minimum number of iteration?

Solution:
No, since the data is not linearly separable.

(f) [1 point] **True/False** The training error of boosting classifier (combination of all the weak classifier) monotonically decreases as the number of iterations in the boosting algorithm increases. Justify your answer in at most two sentences.

Solution:
False, boosting minimizes loss function: $\sum_{i=1}^{m} exp(-y_i f(x_i))$ which doesn't necessary mean that the training error monotonically decrease. Please look at slides 14-18 `http://www.cs.cmu.edu/~tom/10601_fall2012/slides/boosting.pdf`.

Question 9. **Model Selection**

(a) [2 points] Consider learning a classifier in a situation with 1000 features total. 50 of them are truly informative about class. Another 50 features are direct copies of the first 50 features. The final 900 features are not informative.

Assume there is enough data to reliably assess how useful features are, and the feature selection methods are using good thresholds.

- How many features will be selected by mutual information filtering?

Solution:
about 100

- How many features will be selected by a wrapper method?

Solution:
about 50

(b) Consider k-fold cross-validation. Let's consider the tradeoffs of larger or smaller k (the number of folds). For each, please select one of the multiple choice options.

i. [2 points] With a higher number of folds, the estimated error will be, on average,

- (a) Higher.
- (b) Lower.
- (c) Same.
- (d) Can't tell.

Extreme Leave one out
Train on all except 1

Solution:
Lower (because more training data)

(c) [8 points] Nearly all the algorithms we have learned about in this course have a tuning parameter for regularization that adjusts the bias/variance tradeoff, and can be used to protect against overfitting. More regularization tends to cause less overfitting.

For each of the following algorithms, we point out one such tuning parameter. If you increase the parameter, does it lead to MORE or LESS regularization? (In other words, MORE bias (and less variance), or LESS bias (and more variance)?) For every blank, please write MORE or LESS.

Naive Bayes: MAP estimation of binary features' $p(X\|Y)$, using a $Beta(\alpha, \alpha)$ prior.	Higher α means…	6.	regularization
Logistic regression, linear regression, or a neural network with a $\lambda \sum_j w_j^2$ penalty in the objective	Higher λ means…	MORE	regularization
Bayesian learning for real-valued parameter θ, given a prior $p(\theta)$, which might a wide or narrow shape. (For example, a high vs. low variance gaussian prior.)	Higher width of the prior distribution means…	1.	regularization
Neural Network: number of hidden units, n	Higher n means…	LESS	regularization
Feature selection with mutual information scoring: Include a feature in the model only if its MI(feat, class) is higher than a threshold t.	Higher t means…	7.	regularization
Decision tree: n, an upper limit on number of nodes in the tree.	Higher n means…	LESS	regularization
Boosting: number of iterations, n	Higher n means…	LESS	regularization
Dimension reduction as preprocessing: Instead of using all features, reduce the training data down to k dimensions with PCA, and use the PCA projections as the only features.	Higher k means…	LESS	regularization

Solution:

- NB α: more
- λ L2 penalty: more
- Bayesian prior width: less
- Num. hidden units: less
- MI threshold: more
- Num. dtree nodes: less
- Num. boosting iter: less
- Num. PC's: less