CSE 546 Final Exam, Autumn 2013

1. Personal info:
   - Name:
   - Student ID:
   - E-mail address:
2. There should be 15 numbered pages in this exam (including this cover sheet).
3. You can use any material you brought: any book, class notes, your print outs of class materials that are on the class website, including my annotated slides and relevant readings. You cannot use materials brought by other students. Calculators are not necessary. Laptops, PDAs, phones and Internet access are not allowed.
4. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
5. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
6. You have 1 hour and 50 minutes.
7. Good luck!

Question	Topic	Max	Score
1	k-Means	10	
2	Expectation Maximization	12	
3	Kernel Regression and k-NN	14	
4	Support Vector Machines	12	
5	LOOCV	12	
6	Markov Decision Processes	14	
7	Learning Theory	12	
8	Bayesian Networks	14	
	TOTAL	100	

1 k-Means [10 points]

Consider the following clustering method called Leader Clustering. It receives two parameters: an integer k and a real number t. Similar to k-means, it starts by selecting k instances (which will be called leaders) and assigns each training instances to the cluster of the closest leader. During the assignment step, however, if the distance of a training instance to its closest leader is greater than the input threshold t, then this training instance becomes a new leader. During the same assignment step, remaining points can be assigned to these new leaders. After all the training instances have been assigned to a leader's cluster, new leaders are calculated by averaging each cluster. The process is then repeated until the cluster assignments do not change.

1. [6 points] Given a dataset and a value k, let t vary from 0 to a very large value. When does Leader Clustering produce more, the same number, or fewer clusters than k-means, assuming that the k initial centers are the same for both? When will the clusterings produced by Leader Clustering and k-means be identical?

2. [4 points] Which of the two methods, k-means or Leader Clustering, will be best at dealing with outliers (data instances that are "far away" or very different to the other instances in the dataset)? Explain.

2 Expectation Maximization [12 points]

Let's say that in a certain course, the probability of each student getting a given grade is:

- $P(A) = \frac{1}{2}$
- $P(B) = \mu$
- $P(C) = 2\mu$
- $P(D) = \frac{1}{2} - 3\mu$

We then observe some data. Let a, b, c and d be the number of students who got an A, B, C or D, respectively.

1. [8 points] What is the Maximum Likelihood Estimator for μ, given a, b, c and d?

Now let's say that we observe some new data, but we don't observe a or b. Instead, we observe h, which is the number of students who got either an A or a B. So we don't know a or b but we know that $h = a + b$. We still observe c and d as before. We now intend to use the Expectation Maximization algorithm to find an estimate of μ.

2. [2 points] **Expectation step**: Given $\hat{\mu}$, a current estimate of μ, what are the expected values of a and b?

3. [2 points] **Maximizaton step**: Given $\hat{a}$ and $\hat{b}$, the expected values of a and b, what is the MLE of μ?

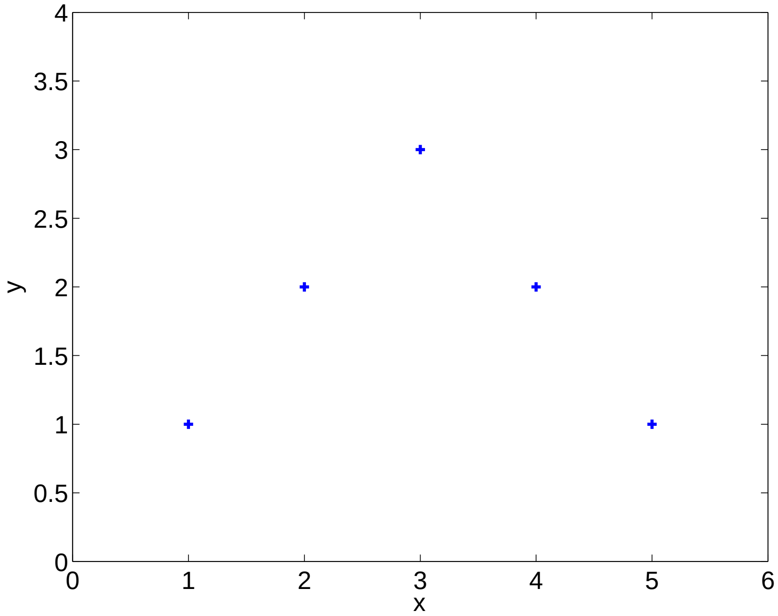
3 Kernel Regression and k-NN [12 points]

1. [6 points] Sketch the fit Y given X for the dataset given below using kernel regression with a box kernel

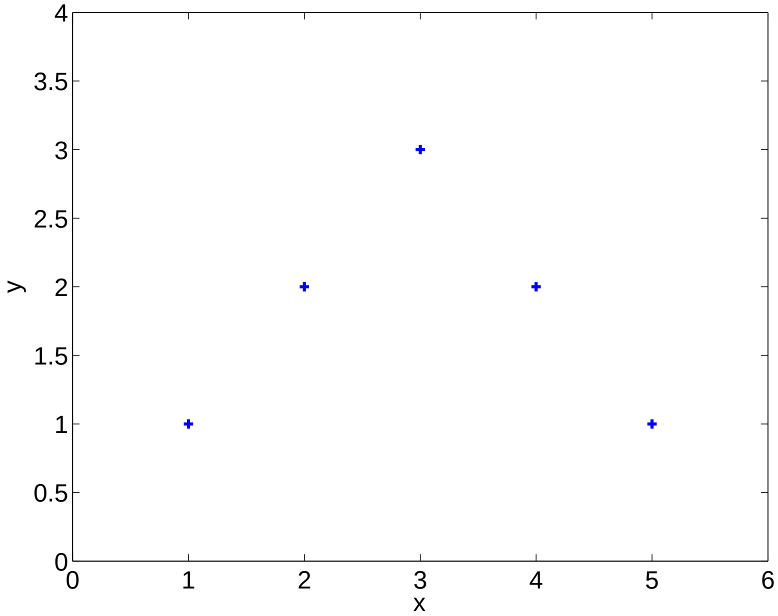
$$K(x_i, x_j) = I(-h \leq x_i - x_j < h) = \begin{cases} 1 & if - h \leq x_i - x_j < h \\ 0 & otherwise \end{cases}$$

for $h = 0.5, 2$.

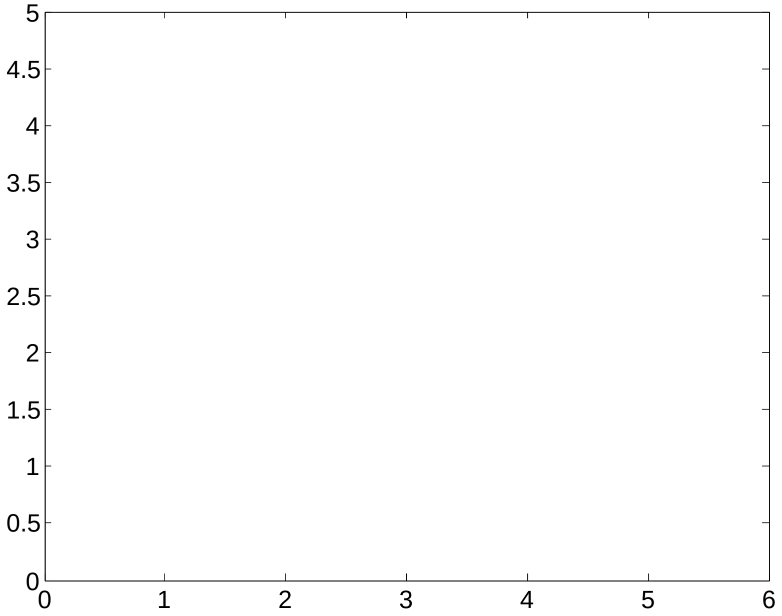
- $h = 0.5$



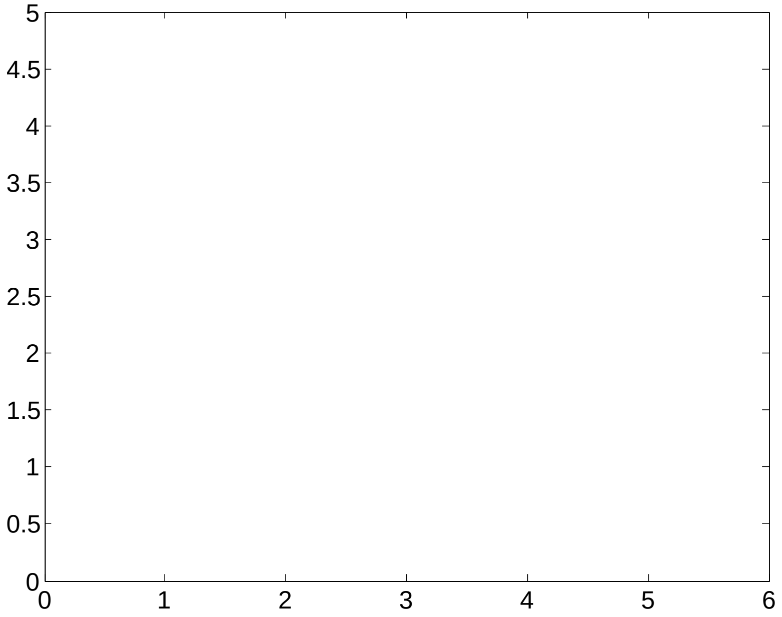
- $h = 2$



2. [4 points] Sketch or describe a dataset where kernel regression with the box kernel above with $h = 0.5$ gives the same regression values as 1-NN but not as 2-NN in the domain $x \in [0, 6]$ below.

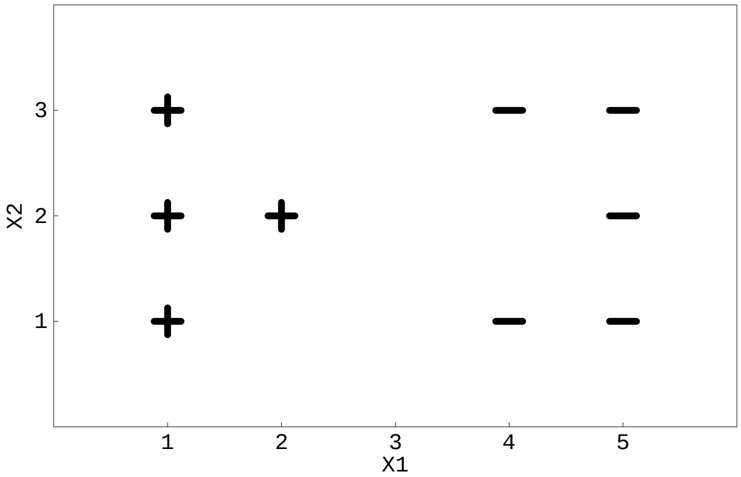


3. [4 points] Sketch or describe a dataset where kernel regression with the box kernel above with $h = 0.5$ gives the same regression values as 2-NN but not as 1-NN in the domain $x \in (0, 6)$ below.



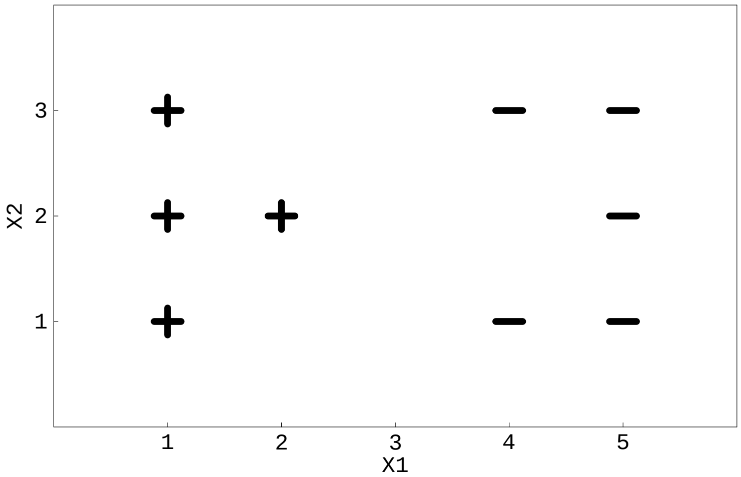
4 Support Vector Machines [12 Points]

1. [2 points] Suppose we are using a linear SVM (i.e., no kernel), with some large C value, and are given the following data set.



Draw the decision boundary of linear SVM. Give a brief explanation.

[3 points] In the following image, circle the points such that by removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample.

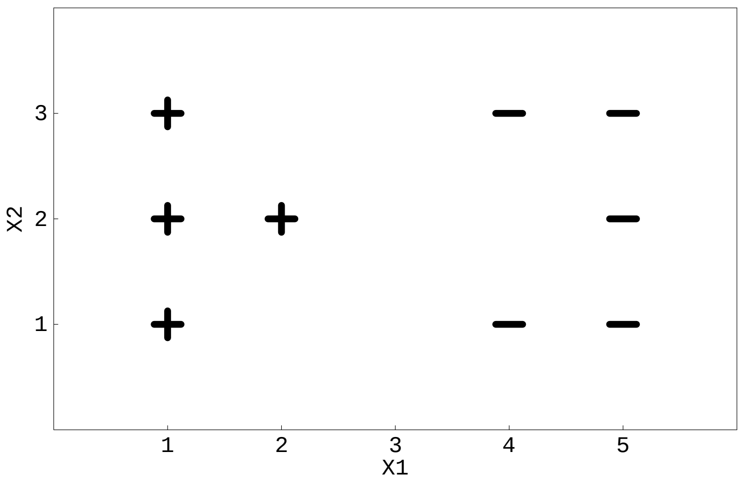


You do not need to provide a formal proof, but give a one or two sentence explanation.

2. [3 points] Suppose instead of SVM, we use regularized logistic regression to learn the classifier. That is,

$$(w, b) = \arg\min_{w \in \mathbb{R}^2, b \in \mathbb{R}} \frac{\|w\|^2}{2} - \sum_i \mathbb{1}[y^{(i)} = 0] \ln \frac{1}{1 + e^{(w \cdot x^{(i)} + b)}} + \mathbb{1}[y^{(i)} = 1] \ln \frac{e^{(w \cdot x^{(i)} + b)}}{1 + e^{(w \cdot x^{(i)} + b)}}.$$

In the following image, circle the points such that by removing that example from the training set and running regularized logistic regression, we would get a different decision boundary than training with regularized logistic regression on the full sample.



You do not need to provide a formal proof, but give a one or two sentence explanation.

3. [4 points] Suppose we have a kernel $K(\cdot, \cdot)$, such that there is an implicit high-dimensional feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$ that satisfies $\forall x, z \in \mathbb{R}^d,\ K(x, z) = \phi(x) \cdot \phi(z)$, where $\phi(x) \cdot \phi(z) = \sum_{i=1}^{D} \phi(x)_i \phi(z)_i$ is the dot product in the D-dimensional space.

   Show how to calculate the Euclidean distance in the D-dimensional space

$$\|\phi(x) - \phi(z)\| = \sqrt{\sum_{i=1}^{D} (\phi(x)_i - \phi(z)_i)^2}$$

   without explicitly calculating the values in the D-dimensional vectors. For this question, you should provide a formal proof.

5 LOOCV [12 points]

5.1 Mean Squared Error

Suppose you have 100 datapoints $\{(x^{(k)}, y^{(k)})\}_{k=1}^{100}$. Your dataset has one input and one output. The kth datapoint is generated as follows:

$$\begin{aligned} x^{(k)} &= k/100 \\ y^{(k)} &\sim Bernoulli(p) \end{aligned}$$

(A random variable with a Bernoulli distribution with parameter p equals 1 with probability p and 0 with probability $1 - p$.) Note that all of the $y^{(k)}$'s are just noise, drawn independently of all other $y^{(k)}$'s. You will consider two learning algorithms:

- **Algorithm NN:** 1-nearest neighbor (with ties broken arbitrarily).
- **Algorithm Zero:** Always predict zero

Hint: Recall that the mean of Bernoulli(p) is p.

1. [1 point] What is the expected Mean Squared Training Error for **Algorithm Zero**?

2. [1 point] What is the expected Mean Squared Training Error for **Algorithm NN**?

3. [1 point] What is the expected Mean Squared LOOCV error for **Algorithm Zero**?

4. [4 points] What is the expected Mean Squared LOOCV error for **Algorithm NN**?

5.2 k-NN

[5 points] Come up with a one-dimensional dataset (with 8 to 20 examples) for which the LOOCV accuracy with 1-nearest neighbor is 100%, but with 3-nearest neighbors is 0%.

6 Markov Decision Processes [14 points]

In the game micro-blackjack, you repeatedly draw a number (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw (a number) or Stop if the sum of the numbers you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your reward equals the sum if the sum is less than 6, or equals 0 if the sum is 6 or higher. When you Draw, you receive no reward for that action and continue to the next round. The game ends only when you Stop. There is no discount ($\gamma = 1$).

1. [1 point] What is the state space for this MDP?

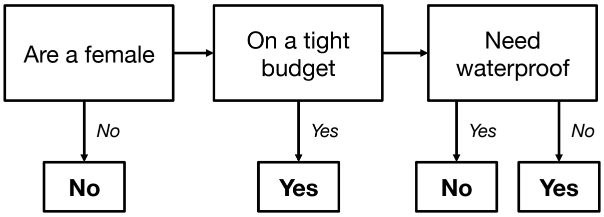
2. [3 points] What is the reward function for this MDP?

3. [7 points] In order to learn the optimal policy for this MDP we must compute the optimal policy value $V^*(x)$, where x is an arbitrary state. In lecture there was given the Value Iteration method for finding $V^*(x)$. Apply one iteration of Value Iteration on this MDP. Let the initial estimate $V^0(x)$ be set to $V^0(x) = \max_a R(x, a)$, where a is an action and $R(x, a)$ is the reward function. Clearly state your answer for $V^1(x)$. You do not have to show basic calculations.

4. [3 points] What is the optimal policy for this MDP? No explanation is necessary.

7 Learning Theory [12 Points]

For this question, we consider the hypothesis space of decision lists. A decision list classifier consists of a series of binary tests. If a test succeeds, a prediction is returned. Otherwise, processing continues with the next test in the list. An example decision list of length 3 is shown below and classifies whether someone should purchase a specific pair of hiking boots.



For this problem, we assume each datapoint consists of d binary attributes. Thus, each decision list represents a function

$$f : \{0,1\}^d \to \{0,1\}$$

1. [8 points] Suppose we consider only decision lists of length k. Derive an upper-bound on the size of the hypothesis space for this problem. Some possible answers include:

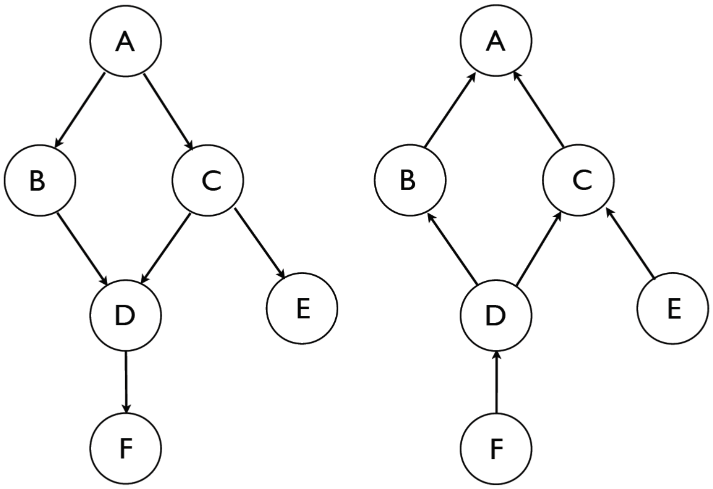
   - $(4d)^k$
   - 2^{2^n}
   - $(2k)^d$
   - $2^{2^k d}$

   Bounds that are overly loose will be given partial credit (by "overly loose," we mean off by a very large amount, so don't worry about this too much). Show your work!

2. [4 points] Assume the true classification function can be expressed by a decision list of length k. Give a PAC bound for the number of examples needed to learn decision lists of length k with d binary attributes to guarantee a generalization error less than ϵ with probability $1 - \delta$. For this problem, assume there is no label noise.

8 Bayes Nets [14 Points]

For this question, consider the following two binary Bayes nets. Each variable may take values from the set $\{0, 1\}$.



(a)

(b)

1. [4 points] Write down two conditional independencies (e.g. $X \perp Y \mid Z$) that hold for both (a) and (b).

2. [4 points] Write down two conditional independencies that hold for one graph but not the other. Specify for which graphs your statements are true and for which graphs they are false.

3. [6 points] For model (a), consider the following marginal distribution $P(A, B, C)$:

A	B	C	P(A, B, C)
0	0	0	3/12
0	0	1	1/12
0	1	0	1/12
0	1	1	1/12
1	0	0	1/12
1	0	1	1/12
1	1	0	1/12
1	1	1	3/12

Based on the above table, list the probabilities $P(A)$, $P(B \mid A)$, and $P(C \mid A)$ associated with the Bayes net. If this is not possible, explain why not.