10-701 Midterm Exam, Spring 2006

1. Write your name and your email address below.
   - Name:
   - Andrew account:
2. There should be 17 numbered pages in this exam (including this cover sheet).
3. You may use any and all books, papers, and notes that you brought to the exam, but not materials brought by nearby students. Calculators are allowed, but no laptops, PDAs, or Internet access.
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. Note there are extra-credit sub-questions. The grade curve will be made without considering students' extra credit points. The extra credit will then be used to try to bump your grade up without affecting anyone else's grade.
7. You have 80 minutes.
8. Good luck!

Question	Topic	Max. score	Score
1	Short questions	12 + 0.52 extra	
2	Regression	12	
3	k-NN and Cross Validation	16	
4	Decision trees and pruning	20	
5	Learning theory	20 + 6 extra	
6	SVM and slacks	20 + 6 extra	

1 [12 points] Short questions

The following short questions should be answered with at most two sentences, and/or a picture. For the **(true/false)** questions, answer true or false. If you answer true, provide a short justification, if false explain why or provide a small counterexample.

1. [2 points] Discuss whether MAP estimates are less prone to overfitting than MLE.

2. [2 points] **true/false** Consider a classification problem with n attributes. The VC dimension of the corresponding (linear) SVM hypothesis space is larger than that of the corresponding logistic regression hypothesis space.

3. [2 points] Consider a classification problem with two classes and n binary attributes. How many parameters would you need to learn with a Naive Bayes classifier? How many parameters would you need to learn with a Bayes optimal classifier?

4. [2 points] For an SVM, if we remove one of the support vectors from the training set, does the size of the maximum margin decrease, stay the same, or increase for that data set?

5. [2 points] **true/false** In n-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Are the error estimates of the separate folds also independent? So, given that the data in test folds i and j are independent, are e_i and e_j, the error estimates on test folds i and j, also independent?

6. [2 points] **true/false** There is an *a priori* good choice of n for n-fold cross-validation.

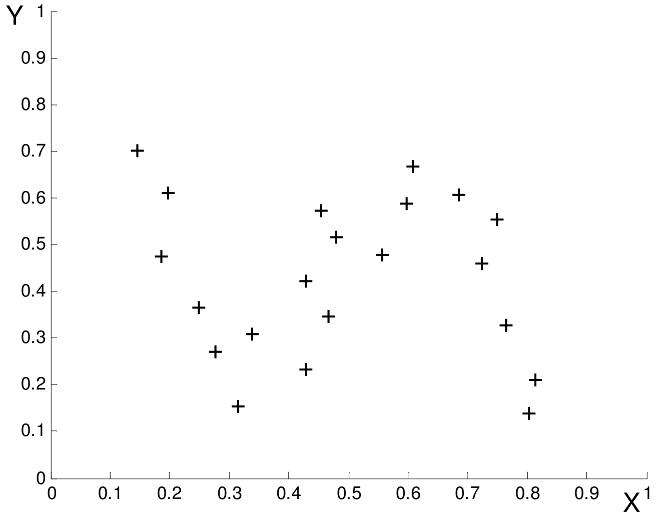
7. [0.52 extra credit points] Which of following songs are hits played by the B-52s:

   - Love Shack
   - Private Idaho
   - Symphony No. 5 in C Minor, Op. 67

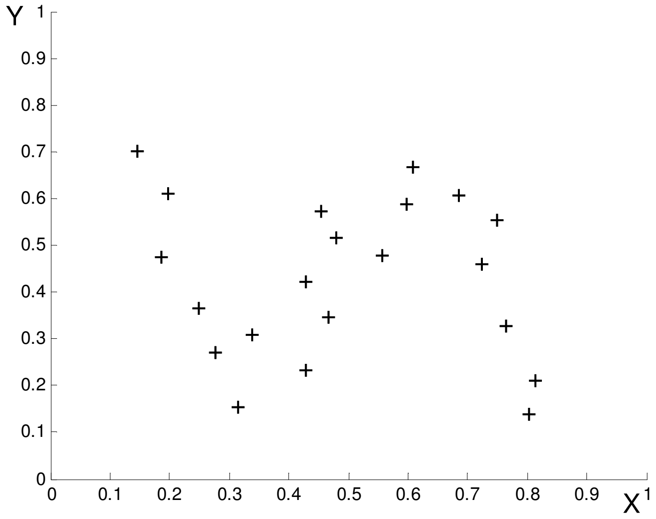
2 [12 points] Regression

For each of the following questions, you are given the same data set. Your task is to fit a smooth function to this data set using several regression techniques. Please answer all questions qualitatively, drawing the functions in the respective figures.

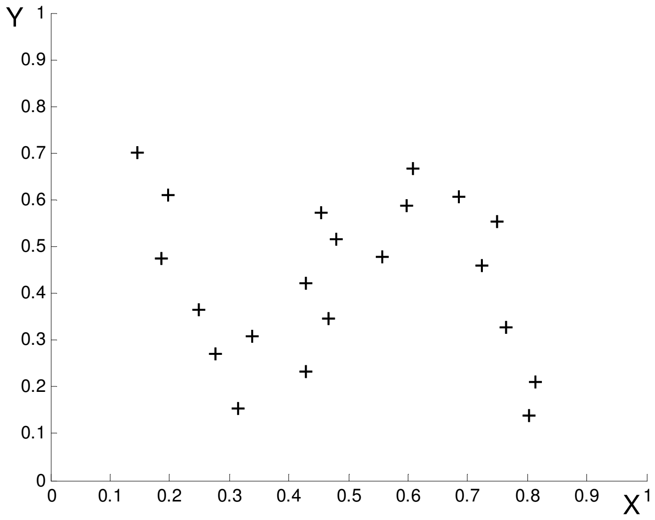
1. [3 points] Show the least squares fit of a linear regression model $Y = aX + b$.



2. [3 points] Show the fit using Kernel regression with Gaussian kernel and an appropriately chosen bandwidth.



3. [3 points] Show the fit using Kernel local linear regression for an appropriately chosen bandwidth.



4. [3 points] Suggest a linear regression model $Y = \sum_i \phi_i(X)$ which fits the data well. Why might you prefer this model to the kernel local linear regression model from part 3)?

3 [16 points] *k*-nearest neighbor and cross-validation

In the following questions you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. Note that a point can be its own neighbor.

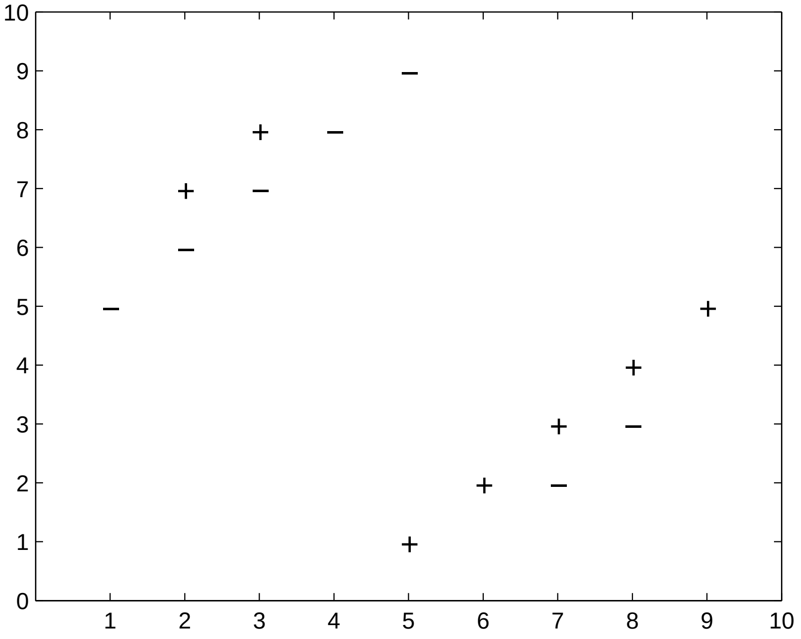


Figure 1: Dataset for KNN binary classification task.

1. [3 points] What value of k minimizes the training set error for this dataset? What is the resulting training error?

2. [3 points] Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?

3. [6 points] What value of k minimizes leave-one-out cross-validation error for this dataset? What is the resulting error?

4. [4 points] In Figure 1, sketch the 1-nearest neighbor decision boundary for this dataset.

4 [20] Decision trees and pruning

You get the following data set:

V	W	X	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	0

Your task is to build a decision tree for classifying variable Y. (You can think of the data set as replicated many times, i.e. overfitting is not an issue here).

1. [6 points] Compute the information gains $IG(Y|V)$, $IG(Y|W)$ and $IG(Y|X)$. Remember, information gain is defined as

   $$IG(A|B) = H(A) - \sum_{b \in B} P(B = b) H(A|B = b)$$

   where

   $$H(A) = -\sum_{a \in A} P(A = a) \log_2 P(A = a)$$

   is the entropy of A and

   $$H(A|B = b) = -\sum_{a \in A} P(A = a|B = b) \log_2 P(A = a|B = b)$$

   is conditional entropy of A given B.

   Which attribute would ID3 select first?

2. [3 points] Write down the entire decision tree constructed by ID3, without pruning.

3. [3 points] One idea for pruning would be to start at the root, and prune splits for which the information gain (or some other criterion) is less than some small ε. This is called top-down pruning. What is the decision tree returned for $\varepsilon = 0.0001$? What is the training set error for this tree?

4. [3 points] Another option would be to start at the leaves, and prune subtrees for which the information gain (or some other criterion) of a split is less than some small ε. In this method, no ancestors of children with high information gain will get pruned. This is called bottom-up pruning. What is the tree returned for $\varepsilon = 0.0001$? What is the training set error for this tree?

5. [2 points] Discuss when you would want to choose bottom-up pruning over top-down pruning and vice versa. Compare the classification accuracy and computational complexity of both types of pruning.

6. [3 points] What is the height of the tree returned by ID3 with bottom-up pruning? Can you find a tree with smaller height which also perfectly classifies Y on the training set? What conclusions does that imply about the performance of the ID3 algorithm?

5 [20 + 6 points] Learning theory

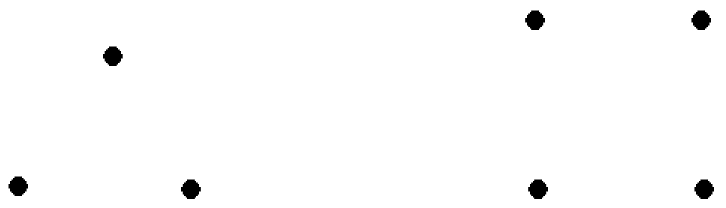
5.1 [8 points] Sample complexity

Consider the following hypothesis class: 3-SAT formulas over n attributes with k clauses. A 3-SAT formula is a conjunction (AND, $\wedge$) of clauses, where each clause is a disjunction (OR, $\vee$) or three attributes, the attributes may appear positively or negated ($\neg$) in a clause, and an attribute may appear in many clauses. Here is an example over 10 attributes, with 5 clauses:

$$(X_1 \vee \neg X_2 \vee X_3) \wedge (\neg X_2 \vee X_4 \vee \neg X_7) \wedge (X_3 \vee \neg X_5 \vee \neg X_9) \wedge (\neg X_7 \vee \neg X_6 \vee \neg X_{10}) \wedge (X_5 \vee X_8 \vee X_{10}).$$

You are hired as a consultant for a new company called FreeSAT.com, who wants to learn 3-SAT formulas from data. They tell you: We are trying to learn 3-SAT formulas for secret widget data, all we can tell you us that true hypothesis is a 3-SAT formula in the hypothesis class, and our top-secret learning algorithm always returns a hypothesis consistent with the input data.

Here is your job: we give you an upper bound $\epsilon > 0$ on the amount of true error we are willing to accept. We know that this machine learning stuff can be kind of flaky and the hypothesis you provide may not always be good, but it can only be bad with probability at most $\delta > 0$. We really want to know how much data we need. Please provide a bound on the amount of data required to achieve this goal. Try to make your bound as tight as possible. Justify your answer.



(a) *three points* (b) *four points*

Figure 2: Figures for Question 5.2.

5.2 [12 points] VC dimension

Consider the hypothesis class of rectangles, where everything inside the rectangle is labeled as positive: A rectangle is defined by the bottom left corner (x_1, y_1) and the top right corner (x_2, y_2), where $x_2 > x_1$ and $y_2 > y_1$. A point (x, y) is labeled as positive if and only if $x_1 \leq x \leq x_2$ and $y_1 \leq y \leq y_2$. In this question, you will determine the VC dimension of this hypothesis class.

1. [3 points] Consider the three points in Figure 2(a). Show that rectangles can shatter these three points.

2. [3 points] Consider the four points in Figure 2(b). Show that rectangles cannot shatter these four points.

3. [3 points] The VC dimension of a hypothesis space is defined in terms of the largest number of input points that can be shattered, where the "hypothesis" gets to pick the locations, and an opponent gets to pick the labels. Thus, even though you showed in Item 2 that rectangles cannot shatter the four points in Figure 2(b), the VC dimension of rectangles is actually equal to 4. Prove that rectangles have VC dimension of at least 4 by showing the position of four points that can be shattered by rectangles. Justify your answer.

4. [3 points] So far, you have proved that the VC dimension of rectangles is at least 4. Prove that the VC dimension is exactly 4 by showing that there is no set of 5 points which can be shattered by rectangles.

5. **Extra credit:** [6 points] Now consider *signed* rectangles, where, in addition to defining the corners, you get to define whether everything inside the rectangle is labeled as positive or as negative. What is the VC dimension of this hypothesis class?

   Prove tight upper and lower bounds: if your answer is k, show that you can shatter k points and also show that $k + 1$ points can not be shattered.

6 [20 + 6 points] SVM and slacks

Consider a simple classification problem: there is one feature x with values in $\mathbb{R}$, and class y can be 1 or -1. You have 2 data points:

$$(x_1, y_1) = (1, 1)$$
$$(x_2, y_2) = (-1, -1).$$

(a) [4 points] For this problem write down the QP problem for an SVM with slack variables and Hinge loss. Denote the weight for the slack variables C, and let the equation of the decision boundary be

$$wx + b = 0.$$

(b) [6 points] It turns out that optimal w is

$$w^* = \min(C, 1).$$

Find the optimal b as a function of C. *Hint: for some values of C there will be an interval of optimal b's that are equally good.*

(c) [4 points] Suppose that $C < 1$ and you have chosen a hyperplane

$$xw^* + b^* = 0, \text{such that } b^* = 0$$

as a solution. Now a third point, $(x_3, 1)$, is added to your dataset. Show that if $x_3 > \frac{1}{C}$, then the old parameters (w^*, b^*) achieve the same value of the objective function

$$w^2 + \sum_i C\xi_i$$

for the 3-point dataset as they did for a 2-point dataset.

(d) [6 points] Now in the same situation as in part 5c., assume $x_3 \in [1, \frac{1}{C}]$. Show that there exists a b_3^* such that (w^*, b_3^*) achieve the same value of the objective function for the 3-point dataset as (w^*, b^*) achieve for the 2-point dataset. *Hint: Consider* b_3^* *such that the positive canonical hyperplane contains* x_3.

(e) **Extra credit:** [6 points] Solve the QP problem that you wrote in part 1 for the optimal w. Show that the optimal w is

$$w^* = \min(C, 1).$$

Hint: Pay attention to which constraints will be tight. It is useful to temporarily denote $\xi_1 + \xi_2$ with t. Solve the constraints for t and plug into the objective. Do a case analysis of when the constraint for t in terms of C will be tight.