COMP 478 - Sample Questions - Midterm

General ML Questions

1. Suppose you train a classifier and test it on a validation set. It gets 80% classification accuracy on the training set and 20% classification accuracy on the validation set. From what problem is your model most likely suffering: Underfitting or Overfitting?

   which could reasonably be expected to improve your classifier's performance on the validation set: Add extra features or remove some features?

   Collect more training data or throw out some training data?

   Assuming features are outcome counts (k is the Laplace smoothing parameter controlling the number of extra times you "pretend" to have seen an outcome in the training data): Increase k or Decrease k? (assuming $k > 0$ currently)

Parameter Estimation Methods

1. Identical twins are rare, but just how unlikely are they? With the help of the sociology department, you have a representative sample of twins to help you answer the question. The twins data gives the following observations (a twin refers to one pair of two people):

   m_i = number of identical male twins and f_i = number of identical female twins
   m_f = number of fraternal male twins and f_f = number of fraternal female twins
   b = number of fraternal opposite gender twins

   To model this data, we choose these distributions and parameters:

   Twins are identical with probability θ
   Given identical twins, the twins are male with probability p.

   Given fraternal twins, the probability of male twins is q^2, probability of female twins is $(1-q)^2$ and probability of oppsite gender twins is $2q(1-q)$.

   (a) Write expressions for the likelihood and the log-likelihood of the data as functions of the parameters θ, p, and q for the observations m_i, f_i, m_f , f_f , b.

   (b) What are the maximum likelihood estimates for θ, p and q?

2. Consider the geometric distribution, which has P(X = k) = (1 - $\theta)^{k-1}\theta$. Assume in our training data X took on the values 4, 2, 7, and 9.

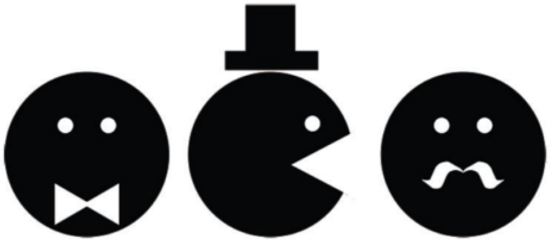
   (a) Write an expression for the log-likelihood of the data as a function of the parameter θ.

   (b) What is the value of θ that maximizes the log-likelihood, i.e., what is the maximum likelihood estimate for θ?

Naive Bayes

1. Pacman and Mrs. Pacman have been searching for each other in the Maze. Mrs. Pacman has been pregnant with a baby, and just this morning she has given birth to Pacbaby (Congratulations, Pacmans!). Because Pacbaby was born before Pacman and Mrs. Pacman reunited in the maze, he has never met his father. Naturally, Mrs. Pacman wants to teach Pacbaby to recognize his father, using a set of Polaroids of Pacman. She also has several pictures of ghosts to use as negative examples. Because the polaroids are black and white, and were taken from strange angles, Mrs. Pacman has decided to teach Pacbaby to identify Pacman based on more salient features: the presence of a bowtie (b), hat (h), or mustache (m).

   The following table summarizes the content of the Polaroids. Each binary feature is represented as 1 (meaning the feature is present) or 0 (meaning it is absent). The subject y of the photo is encoded as +1 for Pacman or 1 for ghost.



(m)	(b)	(h)	Subject (y)
0	0	0	+1
1	0	0	+1
1	1	0	+1
0	1	1	+1
1	0	1	−1
1	1	1	−1

   Suppose Pacbaby has a Naive Bayes based brain:

   (a) Write the Naive Bayes classification rule for this problem (i.e. write a formula which given a data point $x = (m, b, h)$ returns the most likely subject y). Write the formula in terms of conditional and prior probabilities. Be explicit about which parameters are involved, but you do not need to estimate them yet.

   (b) Assuming no smoothing, give estimates for the parameters of the classification rule based on the Polaroids.

	$y = +1$	$y = -1$
$P(y)$		

P	$y = +1$	$y = -1$
$P(m = 1\|y)$		
$P(b = 1\|y)$		
$P(h = 1\|y)$		

   (c) Suppose a character comes by wearing a hat but without a mustache or bowtie. What would happen if Pacbaby had to guess the identity of the character?

   (d) Suppose now that Pacbaby performs Laplace smoothing with strength $k = 1$ (on both the prior and classconditional parameters). Re-estimate the parameters. Now how will Pacbaby classify this new character with the hat and without a mustache or bowtie?

	$y = +1$	$y = -1$
$P(y)$		

P	$y = +1$	$y = -1$
$P(m = 1\|y)$		
$P(b = 1\|y)$		
$P(h = 1\|y)$		

2. The Naive Bayes model has been famously used for classifying spam. We will use it in the "bag-of-words" model: Each email has binary label Y which takes values in spam, ham. Each word w of an email, no matter where in the email it occurs, is assumed to have probability $P(W = w|Y)$, where W takes on words in a pre-determined dictionary. Punctuation is ignored.

   (a) You are in possession of a bag of words spam classifier trained on a large corpus of emails. Below is a table of some estimated word probabilities.

W	note	to	self	become	perfect
$P(W \mid Y = \text{spam})$	1/6	1/8	1/4	1/4	1/8
$P(W \mid Y = \text{ham})$	1/8	1/3	1/4	1/12	1/12

   You are given a new email to classify, with only two words: "perfect note":

   Fill in the circles corresponding to all values of P(Y = spam) for which the bag of words with these word probabilities will give "spam" as the most likely label.

   ○ 0
   ○ 0.2
   ○ 0.4
   ○ 0.6
   ○ 0.8
   ○ 1

   (b) You are given only three emails as a training set:

   (Spam) "dear sir, I write to you in hope of recovering my gold watch."

   (Ham) "hey, lunch at 12?"

   (Ham) "fine, watch it tomorrow night."

   Fill in the circles corresponding to values you would estimate for the given probabilities, if you were doing no smoothing. (Using Term Frequency: TF)

$P(W = \textbf{sir} \mid Y = \textbf{spam})$	○ 0	○ 1/10	○ 1/5	○ 1/3	○ 2/3	○ None of the above
$P(W = \textbf{watch} \mid Y = \textbf{ham})$	○ 0	○ 1/10	○ 1/5	○ 1/3	○ 2/3	○ None of the above
$P(W = \textbf{gauntlet} \mid Y = \textbf{ham})$	○ 0	○ 1/10	○ 1/5	○ 1/3	○ 2/3	○ None of the above
$P(Y = \textbf{ham})$	○ 0	○ 1/10	○ 1/5	○ 1/3	○ 2/3	○ None of the above

   (c) You are training with the same emails as in the previous question, but now doing Laplace Smoothing with k = 2. There are V words in the dictionary. Write concise expressions for: (Using TF)

$P(W = \textbf{sir} \mid Y = \textbf{spam})$

$P(W = \textbf{watch} \mid Y = \textbf{ham})$

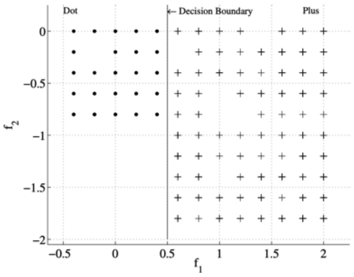
$P(Y = \textbf{ham})$

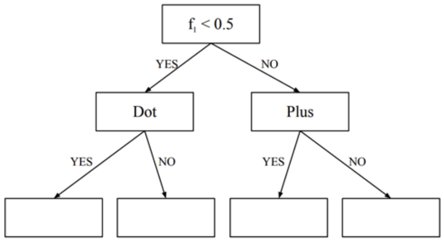
Decision Trees

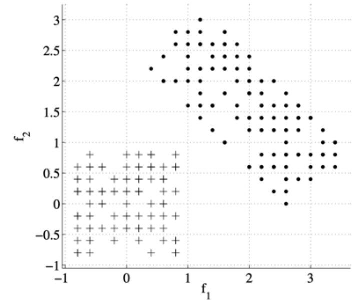
1. You are given points from 2 classes, shown as +'s and ·'s. For each of the following sets of points,

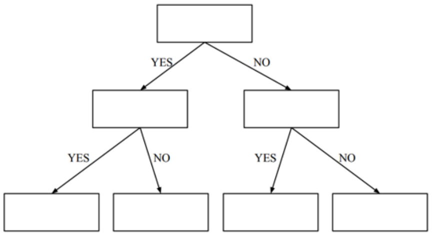
   1. Draw the decision tree of depth at most 2 that can separate the given data completely, by filling in binary predicates (which only involve thresholding of a single variable) in the boxes for the decision trees below. If the data is already separated when you hit a box, simply write the class, and leave the sub-tree hanging from that box empty.

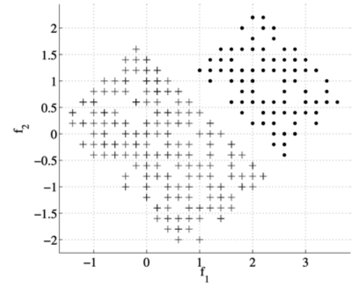
   2. Draw the corresponding decision boundaries on the scatter plot, and write the class labels for each of the resulting bins somewhere inside the resulting bins. If the data can not be separated completely by a depth 2 decision tree, simply cross out the tree template. We solve the first part as an example.

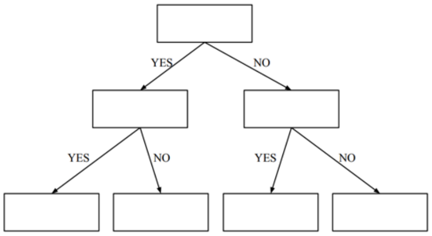


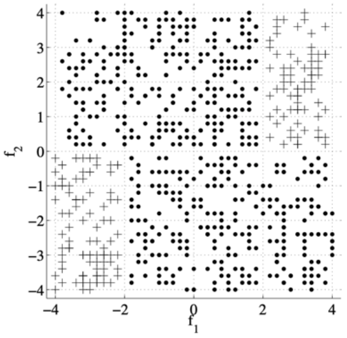


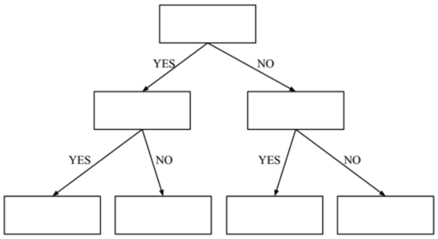












2. Draw a dataset (with + and - samples) for which the decision tree with depth two makes no error while the decision tree with depth one has only 50% accuracy.

K Nearest Neighbours

1. For the given dataset, what is LOOCV accuracy for 1-NN using Manhattan distance function?

f_1	f_2	f_3	Y
1	-1	-1	+
0	0	-1	+
2	-2	0	+
1	1	0	-
1	0	1	-
1	0	-1	-

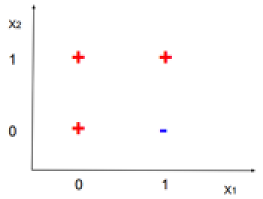
Logistic Regression

1. (True or False?) Logistic regression cannot be kernelized.

2. (True or False?) Sample stochastic gradient descent performs less computation per update than batch gradient descent.

3. (True or False?) A Logistic Regression with L1 regularization will have lower training loss/ higher log likelihood than the same Logistic regression without regularization.

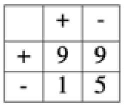
4. A set of reasonably clean sample records was extracted by Barry Becker from the 1994 Census database. We are interested in predicting whether a person makes over 50K a year. For simplicity suppose we model the two features with two boolean variables $X1, X2 \in \{0, 1\}$ and label $Y \in \{0, 1\}$ where Y = 1 indicates a person makes over 50K. In Figure 1 we show three positive samples ("+" for Y = 1) and one negative samples ("-" for Y = 0). Please complete the following questions.



(a) For predicting samples in Figure 1, which model is better: Logistic Regression or Linear Regression. Please explain why.

(b) Is there any logistic regression classifier using X1 and X2 that can perfectly classify the examples in Figure 1? How about if we change label of point (0,1) from "+" to "-"?

(c) If we train a classifier based on data in Figure 1, and then we apply that classifier on a testing dataset. The testing confusion matrix is given by:



	+	-
+	9	9
-	1	5

What is the precision and recall of that classifier?

Questions Credit: Tom Mitchell, Ziv Bar-Joseph, Jonathan Shewchuk, Brian D Ziebart, Daisy Zhe Wang, http://ai.berkeley.edu