EXAM: CIS 472/572, WINTER 2015
INSTRUCTOR: DANIEL LOWD

February 25, 2015

The exam is closed book. You are allowed one page of notes. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. There are 6 problems. Each problem is worth 6 points total.
Undergraduates only: You may skip one problem. Your grade will be computed from the 5 problems you answer. Please write down on the front of your test which problem you are choosing to skip.

NAME Answer Key

- Problem 1: ______
- Problem 2: ______
- Problem 3: ______
- Problem 4: ______
- Problem 5: ______
- Problem 6: ______

- TOTAL: ______

PROBLEM 1A: KERNELS

Consider a kernelized SVM with the following instances and weights:

weight (α_i)	instance (x_i)	label (y_i)
1.0	(0,0,0)	+1
1.0	(1,1,1)	-1

What is the predicted label for the instance $x = (-1, -1, -1)$ under each kernel? (NOTE: Be careful with positive and negative signs! I recommend showing your work in order to have a chance at partial credit.)

1. (2 points) Linear kernel, $K(x, x') = x \cdot x'$.

$1 \cdot 1 \cdot (0+0+0) + (-1 \cdot 1 \cdot (-1 + -1 + -1)) = 0 + 3 = 3$

$\rightarrow$ predicted label is +1

2. (2 points) Cubic kernel, $K(x, x') = (1 + x \cdot x')^3$

$1 \cdot 1 \cdot (1 + (0+0+0))^3 + -1 \cdot 1 \cdot (1 + -3)^3 = 1 + -1 \cdot -8 = 9$

$\rightarrow$ predicted label is +1

PROBLEM 1B: PERCEPTRON UPDATES

Consider a perceptron with $b = -2$ (bias), $w_1 = -1$, $w_2 = 3$, and $w_3 = 2$.

1. (2 points) Show the updated weights after applying the perceptron update with the following example: $x = (1, 0, 1)$, $y = -1$.

$w_1 = -1$

$w_2 = 3$

$w_3 = 2$

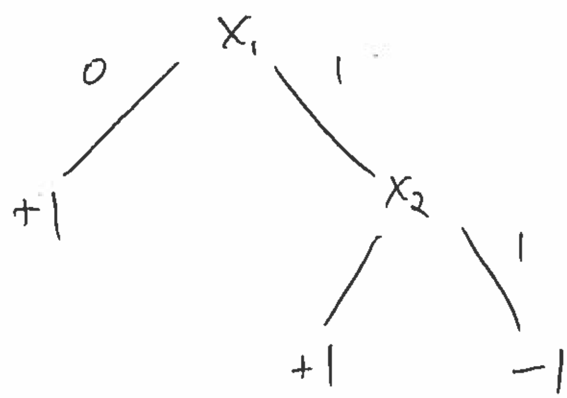
$b = -2$

$\rightarrow$ unchanged because prediction is correct with current weights

PROBLEM 2: REPRESENTATION

Let x_1 and x_2 be two binary-valued attributes, which can take on values of 0 or 1. Consider the NAND function, $y = \neg(x_1 \wedge x_2)$, which is -1 if both $x_1 = 1$ and $x_2 = 1$ and +1 otherwise.

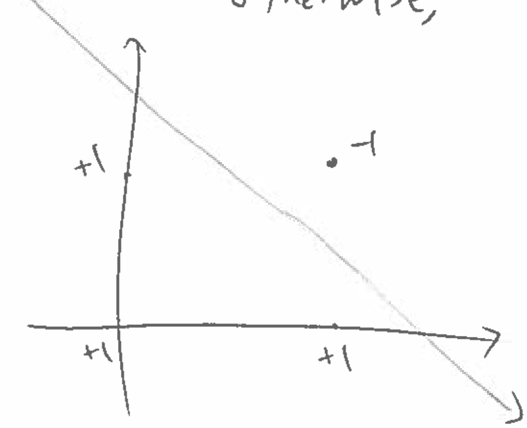
1. (3 points) Draw a decision tree that represents this function.



2. (3 points) Specify parameters for a linear classifier that represents this function. (You do not need to find a maximum margin separator – any separator will do. Do not create any additional attributes.)

$W_1 = -1$
$W_2 = -1$
$b = 1.5$

When $X_1 = 1$ and $X_2 = 1$,
$-1 \cdot 1 + -1 \cdot 1 + 1.5 = 1.5 - 2 = -0.5 < 0$
otherwise, score or activation is > 0.



PROBLEM 3: LEARNING POWER

Which of the following classifiers will have zero training error (under 0/1 loss) on the following dataset? Give a short explanation (one sentence) for each answer.

X1	X2	Category
1	1	0
1	0	1
0	1	1
0	0	0

1. (2 points) 3-nearest neighbor

No — the majority label will always be wrong, since the two closest points (other than the identical training example) have a different label.

2. (2 points) Logistic regression

No — logistic regression is a linear classifier that cannot perfectly represent XOR.

3. (2 points) 2-layer neural network with only one node in the hidden layer

No — only one node in the hidden layer means that the network can only copy or reverse the input, which is a linear classifier. You need multiple hidden units to represent XOR.

PROBLEM 4: NAIVE BAYES

(**NOTE:** This problem could take longer than the others due to the amount of arithmetic required. Consider saving it for when you've answered any other questions that you can easily answer.)

Suppose you are given the following set of training examples:

X_1	X_2	Y
0	1	1
0	0	0
1	0	0

1. (3 points) Use this data to compute the MAP parameter estimates for a naive Bayes model, using a Beta(2,2) prior for all parameters. (Leave your answers as conditional probabilities; do not convert them to linear weights.)

i.e., add 1 to all counts.

$P(Y=1) = \frac{1+1}{3+2} = \frac{2}{5}$

$P(Y=0) = \frac{3}{5}$

$P(X_1=0 / Y_1=1) = \frac{1+1}{1+2} = \frac{2}{3}$

$P(X_1=1 / Y_1=1) = \frac{1}{3}$

$P(X_2=0 / Y_1=1) = \frac{0+1}{1+2} = \frac{1}{3}$

$P(X_2=1 / Y_1=1) = \frac{2}{3}$

$P(X_1=0 / Y_1=0) = \frac{1+1}{2+2} = \frac{1}{2}$

$P(X_1=1 / Y_1=0) = \frac{1}{2}$

$P(X_2=0 / Y_1=0) = \frac{2+1}{2+2} = \frac{3}{4}$

$P(X_2=1 / Y_1=0) = \frac{1}{4}$

2. (3 points) Using your MAP parameter estimates, what label would naive Bayes predict for the instance $x = (1, 1)$? Show at least some of your work.

$P(X_1=1, X_2=1, Y=1) = P(Y=1)P(X_2=1 / Y=1)P(X_1=1 / Y=1)$

$= \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{45}$

$P(X_1=1, X_2=1, Y=0) = P(Y=0)P(X_2=1 / Y=0)P(X_1=1 / Y=0)$

$= \frac{3}{5} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{40}$

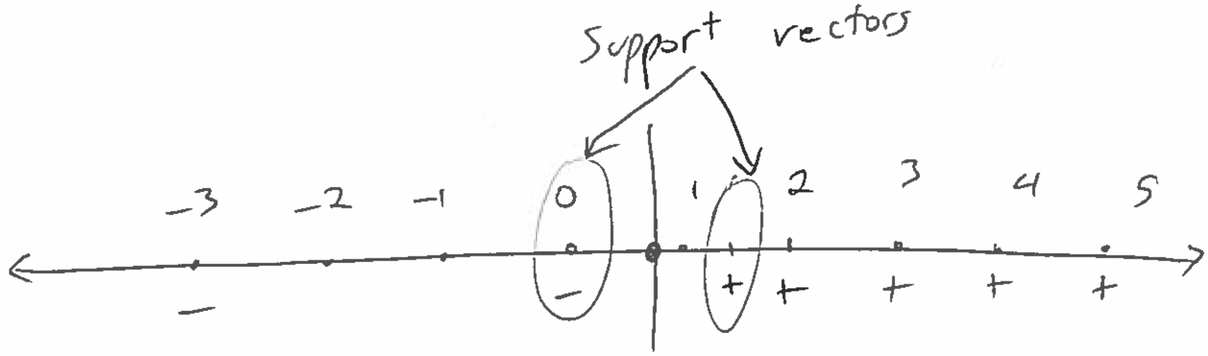
$\boxed{Y = +1 \text{ is more likely according to naive Bayes.}}$

PROBLEM 5: SUPPORT VECTOR MACHINES

Consider the following 1-dimensional data:

x	-3	0	1.5	2	3	4	5
Class	–	–	+	+	+	+	+

1. (3 points) Draw the decision boundary of a (hard-margin) linear support vector machine on this data and identify the support vectors. (The boundary should be a single point on the number line, separating the positive and negative classes.)



Decision boundary is at $x = 0.75$

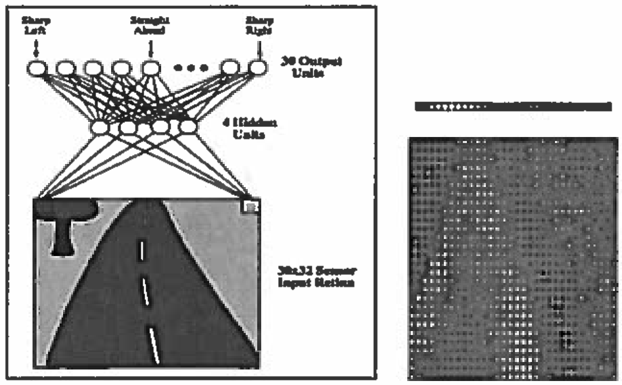
2. (3 points) Calculate the leave one out cross validation error for this SVM on the data set. (That is, how many individual points would be predicted incorrectly if they were removed from the training data?)

Just one — when 0 is removed, boundary moves to -0.75 classifying 0 as positive instead of negative.

Removing 1.5 shifts the boundary to 1, but 1.5 is still correctly classified as positive.

PROBLEM 6: APPLICATIONS

In 1989, 20 years before Google's self-driving car project, researchers at Carnegie Mellon developed an autonomous driving system called ALVINN. ALVINN uses a simple neural network to pick the steering direction based on real-time video input. The video input is a 30x32 pixel camera image and the output is the steering direction. ALVINN was trained using simulated road images along with the recommended steering direction.



1. (2 points) Give one reason why a neural network is a good choice for this problem.

   – Can handle non-linear functions
   (and image processing / computer vision could be very non-linear)
   – Efficient predictions (unlike nearest neighbor)

2. (2 points) Give one reason why a decision tree might be a bad choice for this problem.

   Good driving decisions probably rely on all pixels, requiring an exponentially large tree to test them.

3. (2 points) Give one reason why logistic regression might be a bad choice for this problem.

   Logistic regression is linear, so each pixel has the same influence on the output in all contexts. A linear function is probably too limited to do well on this problem.