EXAM: WINTER 2017
CIS 472/572
INSTRUCTOR: DANIEL LOWD

March 1, 2017

The exam is closed book and open notes (1 page, handwritten except with prior permission). 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. Undergraduates only: You may skip one set of questions (either Problem 1, 2, 3, 4, or 5A and 5B). If you do so, your grade will be your average score from the other questions. Please write down on the front of your test which problem you are choosing to skip.

NAME EXAMPLE SOLUTIONS

- Problem 1: _____
- Problem 2: _____
- Problem 3: _____
- Problem 4: _____
- Problem 5: _____

- TOTAL: _____

PROBLEM 1: TRUE/FALSE QUESTIONS (12 points)

1. (3 points) Classifier A has 90% accuracy on the training set and 75% accuracy on the test set. Classifier B has 78% accuracy on both the training and test sets. Therefore, we can conclude that classifier A is better than classifier B (because it has better mean accuracy). True or (False) **Explain.**

Classifier B is likely to be better because it has better test set performance, which is a better measure of ability to generalize to unseen instances.

2. (3 points) Training a kernelized model (such as kernelized pereptron, logistic regression, or SVM) is equivalent to training a linear model with an expanded set of features. (True) or False. **Explain.**

The "kernel trick" computes a high-dimensional dot product implicitly. This is equivalent to expanding the feature space to many (possibly infinite) dimensions and learning a linear model explicitly.

3. (3 points) Consider two logistic regression models, trained on the same dataset with gradient descent until convergence, but initialized with different initial random weights. Given a sufficiently small learning rate and convergence threshold, both models will have the same accuracy on the test set. (True) or False. **Explain.**

Logistic regression is a convex optimization problem. With small enough steps, trained for long enough, it should converge arbitrarily close to the unique global optimum.

4. (3 points) A non-noisy dataset with n data points can be perfectly represented with no training error by a decision tree with n leaves. (True) of False. **Explain.**

Starting at the root, split on any attribute where some of the points have different values, and proceed recursively until each leaf has one example. Assign the training data point's label to the leaf containing it. This requires n leaves (at most) and makes zero errors on training data.

PROBLEM 2: PERCEPTRON UPDATES (12 points)

You are training a classifier to determine if a web page is about animals or sports.

Word	cute X_1	cat X_2	duck X_3	football X_4	basketball X_5	Topic Y
1.	1	1	0	0	0	1
2.	1	0	1	0	0	1
3.	0	0	2	1	0	-1
4.	1	0	-2	0	1	-1

1. (6 points) Show the weights and bias (w and b) obtained by running perceptron algorithm on this dataset for one iteration. (Here, "one iteration" means going over all of the examples once, in the order shown above.)

Initially, $\vec{w} = \vec{0}$, $b = 0$. $y^{(1)}(w \cdot x^{(1)} + b) = 0 \not> 0$. So: update w+b!

$\vec{w}' = \langle 1, 1, 0, 0, 0 \rangle$ $b' = 1$

$y^{(2)}(w' \cdot x^{(2)} + b') = 2 > 0 \rightarrow$ no change

$(w' \cdot x^{(3)} + b')\, y^{(3)} = -1 \not> 0$. Update! Add $x^{(3)} \cdot y^{(3)}, y^{(3)}$ to w, b.

$w'' = \langle 1, 1, -2, -1, 0 \rangle$ $b = 0$

$y^{(4)}(w'' \cdot x^{(4)} + b'') = -5 \not> 0$. Update! Add $x^{(4)} \cdot y^{(4)}, y^{(4)}$ to w, b.

$w = \langle 0, 1, 0, -1, -1 \rangle$ $b = -1$ ← Final parameters

2. (6 points) Show the weights and bias (α and b) obtained by running the kernelized perceptron algorithm on this dataset with a quadratic kernel.

Initially, $\vec{\alpha} = \vec{0}$, $b = 0$. First pred: $\sum_i \alpha_i K(x^{(i)}, x^{(1)}) y^{(i)} + b = 0 \not> 0$

$\vec{\alpha}' = \langle 1, 0, 0, 0 \rangle$ $b = 1$

Next activation: $1 \cdot (1 + \underbrace{x^{(1)} \cdot x^{(2)}}_{1})^2 \cdot 1 + 1 = 5 > 0 \rightarrow$ no change

Next activation: $1 \cdot (1 + \underbrace{x^{(1)} \cdot x^{(3)}}_{0})^2 \cdot (-1) + (1)(-1) = -2 \not> 0 \rightarrow$ increment α_3

$\vec{\alpha}'' = \langle 1, 0, 1, 0 \rangle$ $b = 0$

Next activation: $\left((1 + \underbrace{x^{(1)} \cdot x^{(4)}}_{1})^2 + -1 \cdot (1 + \underbrace{x^{(3)} \cdot x^{(4)}}_{-4})^2 + 0\right) \cdot (-1) =$

$\left(2^2 - (-3)^2\right) \cdot (-1) = (4 - 9) \cdot (-1) = (-5) \cdot (-1) = 5 > 0$

→ No change.

Final weights:

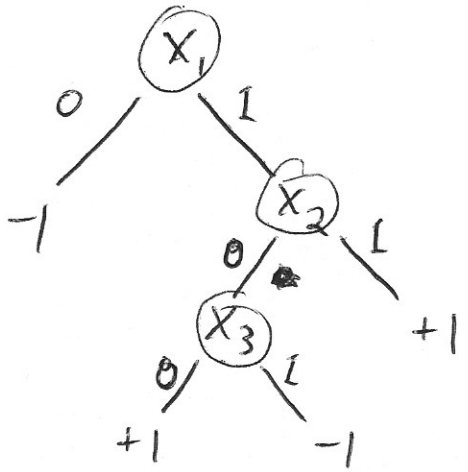
$\vec{\alpha} = \langle 1, 0, 1, 0 \rangle$ $b = 0$ ↖ white board

(or: $\vec{\alpha} = \langle 1, 0, -1, 0 \rangle$ $b = 0$) ↑ CIML

PROBLEM 3: REPRESENTATION (12 points)

Consider the function (x_1 AND (x_2 OR $\neg x_3$)), which is true if $x_1 = 1$ and either $x_2 = 1$ or $x_3 = 0$ (or both). For the inputs (x_1, x_2, x_3), you may assume that each is either 0 (representing false) or 1 (representing true). For the label (y), represent true as 1 and false as -1.

1. (4 points) Draw a decision tree that represents this function.



2. (4 points) Specify weights for a linear classifier that represents this function.

$w = \langle 10, 1, -1 \rangle \quad b = -9.5$

$\uparrow$

$w\vec{x} + b > 0$ when $x_1 = 1$ except for $x_2 = 0$ and $x_3 = 1$,

in which case: $10 \cdot 1 + 1 \cdot 0 + -1 \cdot 1 - 9.5 = -0.5 < 0$.

3. (4 points) Define a set of points to represent this function with a 1-nearest neighbor classifier. Your set of points does not need to be minimal.

Complete set of points:

X_1	X_2	X_3	Y
0	0	0	-1
0	0	1	-1
0	1	0	-1
0	1	1	-1
1	0	0	+1
1	0	1	-1
1	1	0	+1
1	1	1	+1

PROBLEM 4: OVERFITTING (12 points)

For each of the following model classes, briefly describe one method for reducing overfitting (1-2 sentences). In some cases, there may be more than one correct answer.

1. (2 points) Decision tree

Limit depth of decision tree.
(Or: set minimum mutual information to some higher threshold, or adjust some other split criterion.)

2. (2 points) Nearest neighbor

Increasing k smooths predictions by averaging over more neighbors, reducing overfitting.

3. (2 points) Perceptron

Stop early, after a certain number of iterations or when accuracy on validation data gets worse.

4. (2 points) Logistic regression

Increase weight of L_1 and/or L_2 regularization to penalize large attribute weights.

5. (2 points) Neural network with one hidden layer

- Reduce number of hidden units
- Early stopping (as in perceptron)
- (Stronger) L_2 regularization (as in LR)

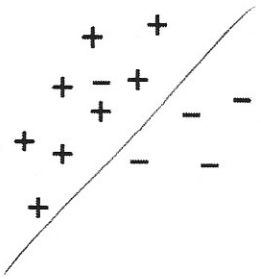
6. (2 points) Support vector machine with a polynomial kernel

- Reduce degree of polynomial kernel (e.g., quadratic instead of cubic)
- Reduce C, so that margin violations are penalized less relative to the size of the margin.

PROBLEM 5A: CLASSIFIER CHOICE (6 points)

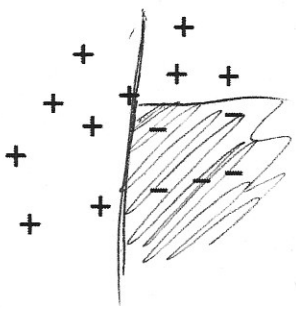
For each of the following 2D datasets, which classifier would you choose and why?

1. (2 points) Dataset 1



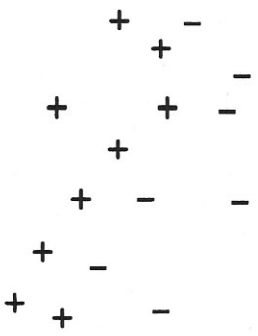
Linear SVM.
Decision boundary looks roughly linear except for 1 point that could be noise.

2. (2 points) Dataset 2



Decision tree.
Negative instances can be partitioned off by setting threshold on each dimension, making it easy to represent as a tree.

3. (2 points) Dataset 3



SVM with polynomial kernel or K-NN (k=1).
Decision boundary is not straight, making it hard for linear model or decision tree to represent well.

PROBLEM 5B: APPLICATIONS (6 points)

Suppose a book publisher wants to predict which people will purchase a new book about cats, based on which other books they've purchased in the past. They send review copies to 1000 people and collect information about which people liked the book and which other books they've purchased. Your task is to suggest a machine learning method to build a predictive model using this data. **For full credit, your answers must be specific to this problem, not simply listing the common advantages and disadvantages of each method.**

1. (2 points) Give one reason why nearest neighbor might be a bad choice for this problem.

Number of other books could be very large, and irrelevant purchases could lead to a poor distance function and poor predictions from dissimilar (but close) neighbors.

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

Decision trees require a number of leaves that is exponential in the number of attributes tested along each path. Since there are many books that could be relevant, testing many would require a large model and a lot of training data.

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

Book preferences might not be linear — some purchases could increase or decrease probability of liking a book, depending on context.

Also, a probability of purchasing might be more valuable than just a binary output, so logistic regression might be a better linear model.