CS178 Midterm Exam

Machine Learning & Data Mining: Winter 2012

Thursday February 15th, 2012

Your name: SOLUTIONS

Name of the person in front of you (if any): _____

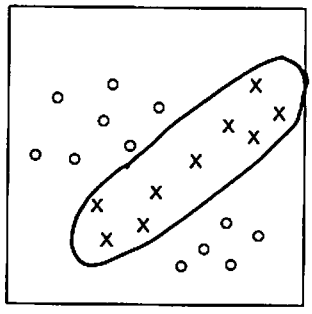
Name of the person to your right (if any): _____

- Total time is 1:15. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly and show all your work.**
- If you need clarification on a problem, please raise your hand and wait for the instructor to come over.
- Turn in any scratch paper with your exam.

(This page intentionally left blank)

Problem 1: (10 points) Separability

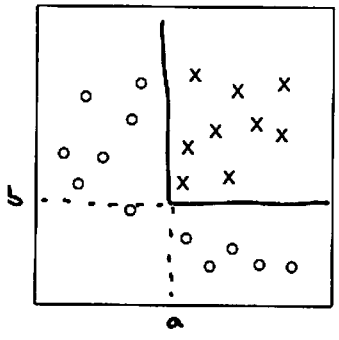
For each of the following examples of training data, sketch a classification boundary that separates the data. State whether or not the data are linearly separable, and if not, give a set of features that would allow the data to be separated.



Not lin seperable

Could separate w/ (for example) quadratic features:

$[1 \; x_1 \; x_2 \; x_1^2 \; x_2^2 \; x_1 x_2]$.



Just to be different...

Also not linearly separable

Can separate with (for example) threshold features

$[1 \; x_1 \geq a \; \; x_2 \geq b]$

where "$x_1 \geq a$" is zero if $x_1 < a$ and 1 otherwise.

Problem 2: (8 points) Under- and Over-fitting

Suppose that I am training a neural network classifier to recognize faces in images. Using cross-validation, we discover that my classifier appears to be overfitting the data. Give two ways I could improve my performance – be specific.

Lots of ways:

(1) Feature selection / dimensionality reduction
(2) Regularization
(3) Early stopping
(4) Get more data
(5) reduce # hidden nodes ...

After following some of your advice, we now think that the resulting classifier is underfitting. Give two ways, **other than** reversing the methods you mentioned above, that we could improve performance; again, be specific.

(1) Reverse unused methods from above; eg.
- add features (Eigenvals) or hidden nodes

(2) Learn a boosted ensemble
(3) add more layers
⋮

Problem 3: (16 points) Regression and Optimization

Suppose that we have training data $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$ and we wish to predict y using the nonlinear regression model:

$$\hat{y}(x) = a + \log(x + b)$$

(a) Write the mean squared error cost function for our predictor.

$$MSE(a,b) = \frac{1}{m} \sum_{i=1}^{m} \left(y^i - a - \log(x^i + b)\right)^2$$

(b) Compute its gradient with respect to the parameters a and b.

$$\frac{\partial MSE}{\partial a} = -\frac{2}{m} \sum_i \left[y^i - (a + \log(x^i + b))\right]$$

$$\frac{\partial MSE}{\partial b} = -\frac{2}{m} \sum_i \left[y^i - (a + \log(x^i + b))\right] \cdot \frac{1}{x^i + b}$$

(c) Give pseudocode to find the optimal values of a and b using gradient descent. Be sure to specify all required aspects of the algorithm, and give enough detail to allow someone to complete the code.

① Initialize a, b to something; set step size α (eg .01)
& tolerance ϵ (eg 1e-4)

② while not done

$\nabla = \left[\frac{\partial MSE}{\partial a} \quad \frac{\partial MSE}{\partial b}\right]$ in part (b)

$a \leftarrow a - \alpha \frac{\partial MSE}{\partial a}$

$b \leftarrow b - \alpha \frac{\partial MSE}{\partial b}$

done = true if (for example) $\|\nabla\| < \epsilon$.

Problem 4: (12 points) Bayes classifiers

Consider the following table of measured data:

x_1	x_2	x_3	y
0	0	0	0
0	0	0	1
0	1	1	0
1	1	0	0
1	1	0	1
1	0	1	1
1	1	1	1

We will use the three observed features x_1, x_2, x_3 to predict class y. In the case of a tie, we will prefer to predict class $y = 0$.

(a) Write down the probabilities necessary for a naïve Bayes classifier:

$p(y=1) = 4/7$

$p(x_1=1 \mid y=0) = 1/3$ $\quad$ $p(x_1=1 \mid y=1) = 3/4$

$p(x_2=1 \mid y=0) = 2/3$ $\quad$ $p(x_2=1 \mid y=1) = 1/2$

$p(x_3=1 \mid y=0) = 1/3$ $\quad$ $p(x_3=1 \mid y=1) = 1/2$

(b) Using your naïve Bayes model, what value of y is predicted given observation $(x_1, x_2, x_3) = (000)$.

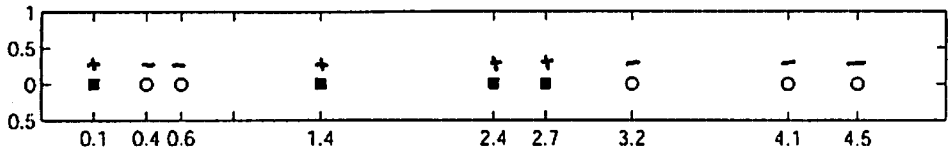
$$p(y=1 \mid 000) = \frac{p(y=1)\, p(x_1 \mid y=1)\, p(x_2 \mid y=1) \cdots}{p(y=1)\, p(x_1 \mid y=1) \cdots + p(y=0)\, p(x_1 \mid y=0) \cdots}$$

$$= \frac{4/7 \cdot 1/4 \cdot 1/2 \cdot 1/2}{4/7 \cdot 1/4 \cdot 1/2 \cdot 1/2 + 3/7 \cdot 2/3 \cdot 1/3 \cdot 2/3} = .36 \leq 1/2 \Rightarrow \text{decide } \hat{y}(000) = 0.$$

(c) Describe a specific prediction problem in which a naïve Bayes classifier might be a good choice, and why.

The classic example is spam filtering. There are many features ($\approx$ 10000 common words in English), so a joint distribution is not feasible, but individual feature models are easy to learn.

Problem 5: (12 points) Cross-validation and Nearest Neighbor



Using the above data with one feature x (whose values are given below each data point) and a class variable $y \in \{-1, +1\}$, with squares indicating $y = +1$ and circles $y = -1$ (the sign is also shown above each data point for redundancy), answer the following:

(a) Compute the leave-one-out cross-validation error of a 1-Nearest-Neighbor classifier. In the case of any ties, select the left-most neighbor at the same distance as the nearest.

x ✓ ✓ x ✓ ✓ x ✓ ✓

⇒ 3/9 error rate.

(b) Compute the leave-one-out cross-validation error for a 3-Nearest-Neighbor classifier.

x x x x ✓ ✓ x ✓ ✓

⇒ 5/9 error rate.

(c) Compute the leave-one-out cross-validation error for a 9-NN classifier.

Need to specify tie-breaker; let's say we prefer "-1"

⇒ x ✓ ✓ x x x ✓ ✓ ✓ ⇒ 4/9

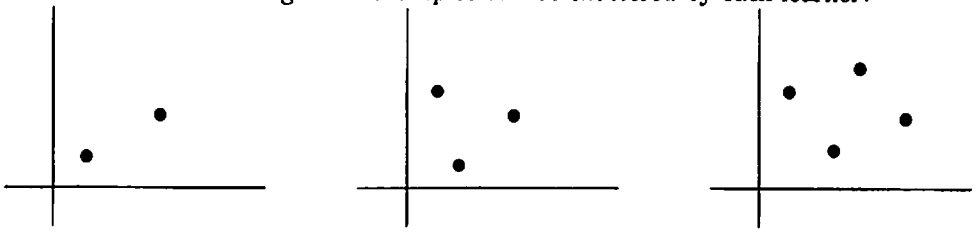
If we prefer "+" instead ⇒ 9/9 errors!

If we prefer the nearest nbr as a tie-breaker, ⇒ 5/9.

Problem 6: (12 points) Shattering and VC Dimension

Suppose that we have two real-valued features x_1 and x_2, and let $T(z) = \begin{cases} +1 & z > 0 \\ -1 & z < 0 \end{cases}$ denote the hard-threshold "sign" function. The variables a, b, and c represent real-valued parameters of our model.

Which of the following three examples can be shattered by each learner?



(#1) (#2) (#3)

(a) $T(a + bx_1)$

⇒ boundaries are vertical lines;

⇒ #1 only

(b) $T(a + bx_1 + cx_2)$

⇒ boundaries are arbitrary lines

⇒ #1, #2.

(c) $T(x_1^2 + a)$ (be careful!)

⇒ decide class -1 inside radius $\sqrt{-a}$

⇒ none - cannot shatter any of the examples.

(d) A Gaussian Bayes classifier with equal covariances

⇒ arbitrary linear decision boundary

⇒ #1, #2.