CS273A Midterm Exam

Solution

Introduction to Machine Learning: Fall 2018

Tuesday November 6th, 2018

Your name:

Row/Seat Number:

Your ID #(e.g., 123456789)

UCINetID (e.g.ucinetid@uci.edu)

- Please put your name and ID **on every page**.
- Total time is 60 minutes. 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**.
- Please ensure your final answer is contained in the space provided. We will not consider or grade anything beyond that space.
- If you need clarification on a problem, please raise your hand and wait for the instructor or TA to come over.
- You may use **one** sheet containing handwritten notes for reference, and a (basic) calculator.
- Turn in your notes and any scratch paper with your exam.

Problems

Total, *(60 points.)*

Name: [ ] ID#: [ ]

Problem 1 Bayes Classifiers, *(10 points.)*

x_1	x_2	y
a	c	0
a	b	0
b	a	0
b	b	0
a	b	1
b	c	1
b	c	1
c	c	1

Consider the table of measured data given at right. We will use the two observed features x_1, x_2 to predict the class y. Each feature can take on one of three values, $x_i \in \{a, b, c\}$.
In the case of a tie, we will prefer to predict class $y = 0$.

(1) Write down the probabilities learned by a naïve Bayes classifier: *(4 points.)*

$p(y=0)$: $\frac{1}{2}$ $p(y=1)$: $\frac{1}{2}$

$p(x_1 = a \,|\, y=0)$: $\frac{1}{2}$ $p(x_1 = a \,|\, y=1)$: $\frac{1}{4}$

$p(x_1 = b \,|\, y=0)$: $\frac{1}{2}$ $p(x_1 = b \,|\, y=1)$: $\frac{1}{2}$

$p(x_1 = c \,|\, y=0)$: 0 $p(x_1 = c \,|\, y=1)$: $\frac{1}{4}$

$p(x_2 = a \,|\, y=0)$: $\frac{1}{4}$ $p(x_2 = a \,|\, y=1)$: 0

$p(x_2 = b \,|\, y=0)$: $\frac{1}{2}$ $p(x_2 = b \,|\, y=1)$: $\frac{1}{4}$

$p(x_2 = c \,|\, y=0)$: $\frac{1}{4}$ $p(x_2 = c \,|\, y=1)$: $\frac{3}{4}$

(2) Using your naïve Bayes model, compute: *(3 points.)*

$p(y=1|x_1=b, x_2=c)$:

$$\frac{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4}}{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4}+\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}} = \frac{3}{4}$$

$p(y=0|x_1=b, x_2=c)$:

$$\frac{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{4}+\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{4}} = \frac{1}{4}$$

(3) Compute the probabilities $p(y=1|x_1=b, x_2=c)$ and $p(y=0|x_1=b, x_2=c)$ for a joint Bayes model trained on the same data. *(3 points.)*

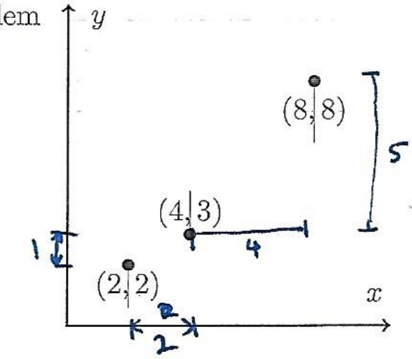
$p(y=1 \,|\, x_1=b, x_2=c) = \frac{2}{2} = 1$

$p(y=0 \,|\, x_1=b, x_2=c) = 0$

Name: ______ ID#: ______

Problem 2 Linear and Nearest Neighbor Regression, *(10 points.)*

Consider the data points shown at right, for a regression problem to predict y given a scalar feature x.



(1) Compute **training** MSE of a 1-nearest neighbor predictor. *(2 points.)*

0

(2) Compute the **leave-one-out** cross-validation error (MSE) of a 1-nearest neighbor predictor. *(2 points.)*

$$\frac{1}{3}\left((2-3)^2+(3-2)^2+(8-3)^2\right) = \frac{1+1+25}{3} = 9$$

(3) Compute the **leave-one-out** cross-validation error (MSE) of a 2-nearest neighbor predictor. *(3 points.)*

$$\frac{1}{3}\left((2-5.5)^2+(3-5)^2+(8-2.5)^2\right) = \frac{1}{3}\left(3.5^2+4+5.5^2\right)$$
$$= 15.5$$

(4) Compute the **leave-one-out** cross-validation MSE of a linear regressor, $f(x) = \theta_0 + \theta_1 x$. *(3 points.)*

$$\frac{1}{3}\left((2-0.5)^2+(3-4)^2+(8-5)^2\right) = \frac{1}{3}\left(1.5^2+1+9\right)$$
$$= \frac{49}{12} \quad (4.0833\ldots)$$

Name: ID#:

Problem 3 Multiple Choice, *(10 points.)*

Here, assume that we have m data points $y^{(i)}, x^{(i)}$, $i = 1 \ldots m$, each with n features, $x^{(i)} = [x_1^{(i)} \; \ldots \; x_n^{(i)}]$. For each of the choices below, will it likely increase, decrease, or have no effect on overfitting (circle your choice)? If you think it is equally likely to go either way, pick *No Effect*.

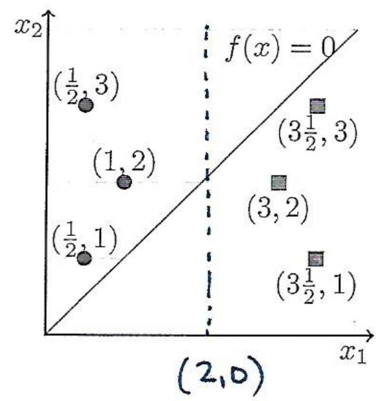
1	Gathering more labeled training data	**Reduce**	**Increase**	**No Effect**
2	For a 3-nearest neighbor classifier, use $2 \times m$ training data by copying (duplicating) each data point.	**Reduce**	**Increase**	**No Effect**
3	For a 3-nearest neighbor classifier, use $2 \times n$ features per data point by copying (duplicating) the features.	**Reduce**	**Increase**	**No Effect**
4	Increasing k for a k-nearest neighbor classifier	**Reduce**	**Increase**	**No Effect**
5	For a linear regressor, use $2 \times m$ training data by adding m all-zero (x and y) data points.	**Reduce**	**Increase**	**No Effect**
6	For a linear regressor, use $2 \times n$ features per data point by adding n all-zero features to each.	**Reduce**	**Increase**	**No Effect**
7	For a linear regressor, use $2 \times n$ features per data point by adding n random values to each.	**Reduce**	**Increase**	**No Effect**
8	Adding another layer to a Neural Network	**Reduce**	**Increase**	**No Effect**
9	Changing the activation function of hidden nodes	**Reduce**	**Increase**	**No Effect**
10	Increasing the `minParents` of a decision tree	**Reduce**	**Increase**	**No Effect**

Name: ______ ID#: ______

Problem 4 Support Vector Machines, *(12 points.)*

Suppose we are learning a linear support vector machine with two real-valued features x_1, x_2 and binary target $y \in \{-1, +1\}$. We observe training data (pictured at right):

x_1	x_2	y
0.5	1	-1
1	2	-1
0.5	3	-1
3	2	+1
3.5	1	+1
3.5	3	+1



Our linear classifier takes the form

$$f(x; w_1, w_2, b) = \text{sign}(w_1 x_1 + w_2 x_2 + b).$$

(1) For given line $x_1 = x_2$ that perfectly separates the data, list the support vectors. *(2 points.)*

$(\frac{1}{2}, 1)$ $\quad$ $(3\frac{1}{2}, 3)$

(2) Derive the parameter values w_1, w_2, b of this $f(x)$. What is the length of the margin? *(4 points.)*

Since $f(x) \Rightarrow x_1 = x_2$, we know $b = 0$ and $w_1 = -w_2$. Let $w_1 = w$, $w_2 = -w$

$f(0,0) = 0$

$\frac{w}{2} - w = -1 \quad \therefore w = 2$

$M = \frac{2}{\sqrt{2^2 + 2^2}} = \frac{1}{\sqrt{2}}$

Verify

$\frac{7w}{2} - 3w = +1 \quad \therefore 7w - 6w = 2 \quad \therefore w = 2$

(3) Consider the *best* linear-SVM classifier; one that separates the data and has the largest margin. Sketch the boundary in the above figure, and list the support vectors here. *(3 points.)*

$(1, 2)$ and $(3, 2)$

(4) Derive the parameter values w_1, w_2, b of this $f(x)$. What is the length of the margin? *(3 points.)*

① $w_1 + 2w_2 + b = -1$

② $3w_1 + 2w_2 + b = +1$

② - ① $\quad 2w_1 = 2 \quad \therefore w_1 = 1$

Putting $(2,0)$, $2 + b = 0 \quad \therefore b = -2$

Putting $w_1 = 1$ & $b = -2$ in ①

$1 + 2w_2 - 2 = -1$

$\therefore w_2 = 0$

$M = \frac{2}{\sqrt{1^2 + 0^2}} = 2$

Name: ID#:

Problem 5 Decision Trees, *(8 points.)*

x_1	x_2	y
a	c	0
b	a	1
b	b	1
c	d	2
d	a	1
d	b	1
d	d	3
d	c	3

Consider the table of measured data given at right. We will use a decision tree to predict the outcome y (one of four classes) using two features, x_1, x_2, where each can take one of four values: a, b, c, d. In the case of ties, we prefer to use the feature with the smaller index (x_1 over x_2, etc.) and prefer to predict class 0 over 1, 1 over 2, etc. You may find the following values useful (**do not** leave logs unexpanded):

$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2$
$\log_2(5) = 2.32 \quad \log_2(6) = 2.59 \quad \log_2(7) = 2.81 \quad \log_2(8) = 3$

(1) What is the entropy of y? *(2 points.)*

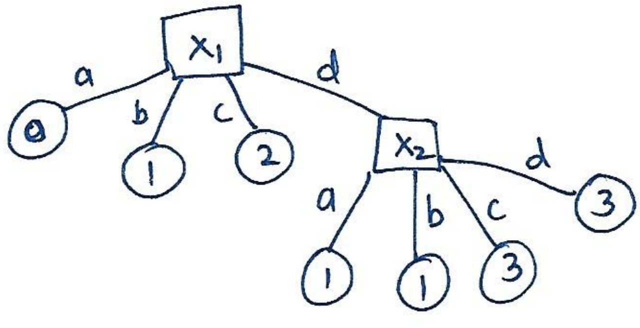
$$H(y) = -\frac{1}{8}\log\frac{1}{8} - \frac{1}{2}\log\frac{1}{2} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{4}\log\frac{1}{4}$$
$$= \frac{3}{8} + \frac{4}{8} + \frac{3}{8} + \frac{4}{8} = \frac{14}{8} = \frac{7}{4} = 1.75$$

(2) What is the information gain of x_1 and x_2? *(4 points.)*

$$H(Y|X_1) = \frac{1}{8}(0) + \frac{1}{4}(0) + \frac{1}{8}(0) + \frac{1}{2}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) = \frac{1}{2}$$
$$H(Y|X_2) = \frac{1}{4}(0) + \frac{1}{4}(0) + \frac{1}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{4}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)$$
$$= \frac{1}{2}$$

$$IG(X_1) = H(Y) - H(Y|X_1) = 1.25$$
$$IG(X_2) = H(Y) - H(Y|X_2) = 1.25$$

(3) Based on the information gain computed in (2), build the complete decision tree learned on this data. *(2 points.)*



Name: ____________ ID#: ____________

Problem 6 VC-Dimensionality, *(10 points.)*

We will be considering a family of "box-shaped" classifiers on a two-dimensional feature space (x_1, x_2), such that the region inside the box is clasified as $+1$.

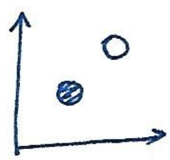
(1) First, consider a simple classifier f_0 that uses a square which has one of the points at the origin, and c as the parameter that defines the edge size, i.e.

$$f_0(x) = \begin{cases} +1 & (0 < x_1 < c) \wedge (0 < x_2 < c) \\ -1 & \text{otherwise} \end{cases}$$

Show that this classifier has a VC-dimensionality of 1. *(2 points.)*

It can shatter one point:

For two points, it cannot shatter because when closer point is -1 & further point is +1



For any two points, you can always create this labelling

∴ VC-dim = 1

(2) Now consider an extension of this classifier with two additional parameters, f_1 that uses a point (a_1, a_2) and c as parameters to describe this square. Specifically:

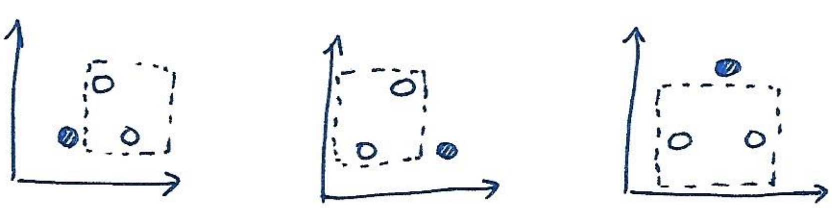
$$f(x) = \begin{cases} +1 & (a_1 < x_1 < a_1 + c) \wedge (a_2 < x_2 < a_2 + c) \\ -1 & \text{otherwise} \end{cases}$$

Show f_1 can shatter 3 points. What does it say about the VC-dimensionality of f? *(2 points.)*

Case when all points are +1 or -1 is easy, since square can surround all or none.

Case where only one point is +1 is easy, since square can surround the point

Interesting case is thus two points being +1.



∴ VC dim ≥ 3

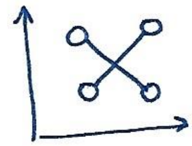
Continued on next page

Name: ID#:

(3) Either show f_1 can shatter 4 points, or argue, informally, why it cannot. What does this say about the VC-dimensionality of f_1? *(4 points.)*

It can't shatter 4 points

For any four points, consider crossing pairs, like:



Clearly, for a square covering the longer line will contain one of the other two.

Thus a labelling of ONLY the two points of the longer crossing line being +1 cannot be ~~sto~~ separated.

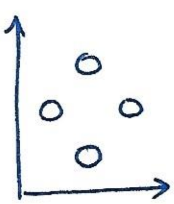
$\therefore$ VC-dim$(f_1) = 3$

(4) Now consider yet another extension f_2 where the region is a rectangle bounded by points (a_1, a_2) and (b_1, b_2), a total of 4 parameters:

$$f_2(x) = \begin{cases} +1 & (a_1 < x_1 < b_1) \wedge (a_2 < x_2 < b_2) \\ -1 & \text{otherwise} \end{cases}$$

Either show f_2 can shatter 4 points, or argue, informally, why it cannot. What does this say about the VC-dimensionality of f_2? *(2 points.)*

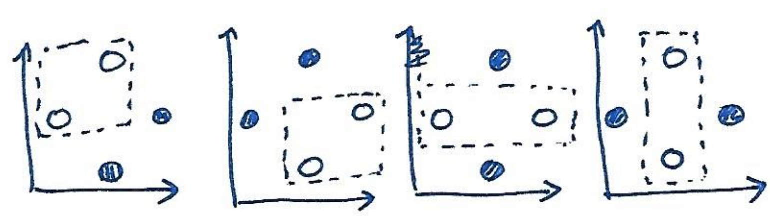
Consider



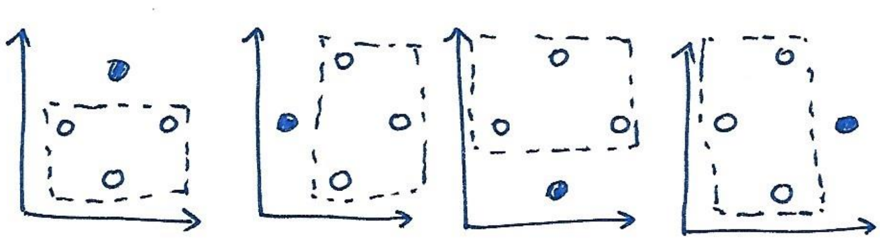
Case where all are +1 or −1 is easy

Case where only one is +1 is easy

For two points +1 :



For three points +1 :



$\therefore$ VC-dim$(f_2) \geqq 4$