CS 178 Midterm Exam

Machine Learning and Data Mining: Fall 2018

Monday November 5th, 2018

Your name: Solutions

Row/Seat Number: Lecture

Your ID #(e.g., 123456789)

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

- Please put your name and ID **on every page.**
- Total time is 50 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.**
- 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, *(42 points.)*

Name: Solutions. ID#:

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.)*

$p(y=0)$: $1/2$ $\qquad$ $p(y=1)$: $1/2$

$p(x_1 = a \,|\, y=0)$: $1/2$ $\qquad$ $p(x_1 = a \,|\, y=1)$: $1/4$

$p(x_1 = b \,|\, y=0)$: $1/2$ $\qquad$ $p(x_1 = b \,|\, y=1)$: $1/2$

$p(x_1 = c \,|\, y=0)$: $\emptyset$ $\qquad$ $p(x_1 = c \,|\, y=1)$: $1/4$

$p(x_2 = a \,|\, y=0)$: $1/4$ $\qquad$ $p(x_2 = a \,|\, y=1)$: $\emptyset$

$p(x_2 = b \,|\, y=0)$: $1/2$ $\qquad$ $p(x_2 = b \,|\, y=1)$: $1/4$

$p(x_2 = c \,|\, y=0)$: $1/4$ $\qquad$ $p(x_2 = c \,|\, y=1)$: $3/4$

(2) Using your naïve Bayes model, compute: *(3 points.)*
$p(y=1|x_1=b, x_2=c)$: $\qquad$ $p(y=0|x_1=b, x_2=c)$:

Calculate $p(y=0, x=bc) = Q_1 = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4}$ and $p(y=1, x=bc) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{4}$

Then $p(y=0 | bc) = \frac{Q_0}{Q_1+Q_0} = \frac{1}{1+3} = 1/4$ and $p(y=1 | bc) = \frac{Q_1}{Q_1+Q_0} = 3/4$.

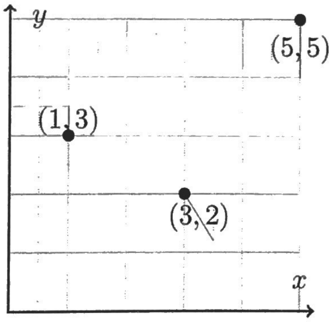
(3) Compute the probabilities $p(y=1|x_1=b, x_2=c)$ and $p(y=0|x_1=b, x_2=c)$ for a joint Bayes model trained on the same data. *(3 points.)*

By inspection, $p(y=1 | bc) = 1$
$p(y=0 | bc) = \emptyset$.

Name: Solutions ID#:

Linear and Nearest Neighbor Regression, *(12 points.)*

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



In the case of ties, ~~we~~ prefer to use the data to the left (smaller values of x).

(1) Compute **training** MSE of a 1-nearest neighbor predictor. *(3 points.)*

ϕ.

(2) Compute the **leave-one-out** cross-validation error (MSE) of a 1-nearest neighbor predictor. *(3 points.)*

Leave out	predict	error
(1,3)	2	1
(3,2)	3	1
(5,5)	2	3

MSE

$\Rightarrow \frac{1}{3}(1^2 + 1^2 + 3^2) = 11/3$.

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

Leave out	predict	error
(1,3)	$3\frac{1}{2}$	$\frac{1}{2}$
(3,2)	4	2
(5,5)	$2\frac{1}{2}$	$2\frac{1}{2}$

MSE

$\Rightarrow \frac{1}{3}\left((\frac{1}{2})^2 + 2^2 + (2\frac{1}{2})^2\right) = 7/2$

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

Leave out	predict	error
(1,3)	-1	4
(3,2)	4	2
(5,5)	1	4

MSE

$\Rightarrow \frac{1}{3}\left[4^2 + 2^2 + 4^2\right] = 12$.

Name: Solutions ID#:

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.*

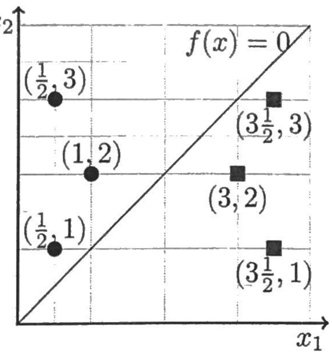
1	Gathering more labeled training data	**Reduce**	**Increase**	**No Effect**
2	For a 3-nearest neighbor classifier, use $2 \times m$ training data by copying (duplicating) each data point. Effectively decreases $K=3$ to $K=1$	**Reduce**	**Increase**	**No Effect**
3	For a 3-nearest neighbor classifier, use $2 \times n$ features per data point by copying (duplicating) the features. Doesn't change distances	**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. Regularizes a bit - sol'n likely to pass through $(0,0) \Rightarrow \theta_0 \approx \emptyset$.	**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. Those features cannot be used by the model	**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	Switching from soft to hard margin SVMs	**Reduce**	**Increase**	**No Effect**

Name: Solutions ID#:

Support Vector Machines, *(10 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.)*

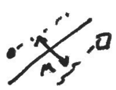
By inspection, SVs are $(1/2, 1)$ and $(3 1/2, 3)$

(closest to decision boundary)

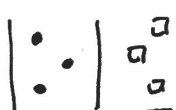
(2) Derive the parameter values w_1, w_2, b of this $f(x)$ using the support vectors. What is the length of the margin? *(3 points.)*

$(0,0)$ on bdry $\Rightarrow$ $w_1 \cdot 0 + w_2 \cdot 0 + b = 0$ $\Rightarrow$ $b = 0$.

SVs $\Rightarrow$ $w_1 \cdot 1/2 + w_2 \cdot 1 + b = -1$, $w_1 \cdot 3 1/2 + w_2 \cdot 3 + b = +1$ $\Rightarrow$ $w_1 = 2$, $w_2 = -2$

Margin  $= \frac{\sqrt{2}}{2}$ by inspection, or formula $\frac{2}{\sqrt{w^T \cdot w}} = \frac{2}{\sqrt{8}} = 1/\sqrt{2}$

(3) Consider the *best* linear-SVM classifier; one that separates the data and has the largest margin. Sketch the boundary i~~n the above fi~~gure, and list the support vectors here. *(2 points.)*



SVs : $(1,2)$ and $(3,2)$.

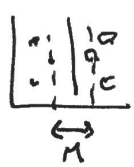
(4) Derive the parameter values w_1, w_2, b of this $f(x)$ using these support vectors. What is the length of the margin? *(3 points.)*

Then, $w_1 \cdot 2 + w_2 \cdot$ (anything) $+ b = 0$.

$w_1 \cdot 1 + w_2 \cdot$ (") $+ b = -1$ $\Rightarrow$

$w_1 \cdot 3 + w_2 \cdot$ (") $+ b = +1$

$w_1 = 1$, $w_2 = 0$, $b = -2$



Margin $M = 2$ by inspection, or $\frac{2}{\sqrt{w^T w}} = \frac{2}{1} = 2$.