Name+Email ID Brian Bertness 1478201
bert0078@umn.edu

CSCI 5521: Introduction to Machine Learning (Spring 2024)[1]

Midterm Exam

Due on Gradescope by 01:00 pm, Mar 22nd

Instructions:

- This test has 4 questions, 100+2 points, including one extra credit problem worth 2 points.
- Please **write your name & ID on your submission pages**.
- **For full credit, show how you arrive at your answers**.
- You have **24 hours** to complete and submit this test to gradescope.

1. (**30 points**) In I-III, fill in the correct option(s) in the following table (it is not necessary to explain).

(I)	(II)	(III)
b, e	a, c, d	a, b

I. Select all the option(s) that correspond to supervised-learning algorithms:

(a) Principal component analysis No Labels
(b) Linear discriminant analysis Uses Labels, separates classes.
Labels Separate Classes
(c) k-means for clustering No Labels - clustering pg 11
(d) Nonparametric classification with a kernel estimator Density Estimation pg 11
(e) Linear discrimination Uses label pg

II. Which of the following option(s) help reduce overfitting in classification?

(a) Adding training data when performing classification More training data less chance of overfit
(b) Adding test data when performing classification Does nothing for underfit
(c) Performing dimensionality reduction on all data before running a classifier Simpler Model
(d) Reducing the number of the parameters in the classifier Simpler Model
(e) Increasing the number of categories (e.g., from binary classification with $K = 2$ to multi-class classification with $K > 2$) when performing classification Higher k is more complex model!

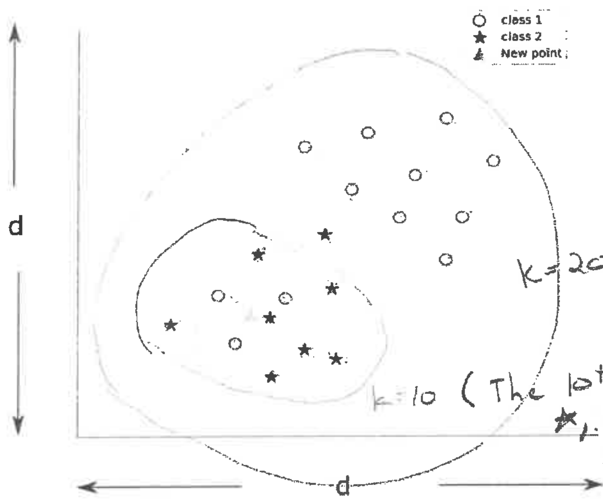
III. Select all the true statement(s) below:

(a) In the training stage of an unsupervised classification task, the model takes in unlabeled data and outputs the model.
(b) In the testing stage of a supervised classification task, the model takes in unlabeled data and outputs the label.
It is Principal Component Analysis, Not Principled Component Analysis
(c) Principled component analysis and linear discriminant analysis are different methods for dimensionality reduction, and therefore must suggest different dimensions for projection.
(d) An objective function is always one to be ~~minimized~~ optimized. Could be maximized or Minimized
(e) Both gradient descent and ~~EM~~ find global optimum.

[1] Instructor: Catherine Qi Zhao. TA: Xianyu Chen, Amarachi Nzeukwu, James Yang, Yifeng Zhang. Email: csci5521.s2024@gmail.com

Name+Email ID Brian Bertness 1478201 bert0078@umn.edu

2. (**24 points**) Given a set of data points $\{x^t\}$ each shown in the figure, find the label of a new data point x using different non-parametric estimators / classifications as specified below.



k=10 (The 10th neighbor will be a ★, tough to tell which of the far away stars are closest.

(a) Write down the label of the new data point x with k nearest neighbor estimator when $k = 10$. Briefly explain the reason. When k=10 we have Δ close to 3 circles (o) and 7 stars (★)

Since 7 > 3 we classify Δ as a Star (★)

(b) Write down the label of the new data point x with k nearest neighbor estimator when $k = 20$. Briefly explain the reason. When k = 20 we have Δ close to 12 circles (o) and 8 stars (★).

Since 12 > 8 we classify Δ as a circle (o).

(c) Assume a uniform kernel function:

$$K(x, x^t) = \begin{cases} \frac{1}{\pi d^2}, & ||x - x^t||_2 \leq d \\ 0, & \text{otherwise} \end{cases}$$

Write down the label of the new data point x with kernel estimator box. Briefly explain the reason.
As stated this is a uniform kernel estimator. So, for any distance within the area of the box the weight of each point is the same. Thus, as in parts a and b above, the classification would not change. That is, when k=10 class = ★ and when k=20 class = O. kernel estimator with uniform kernel is a lot like knn.

(d) (Extra credit, 2 points) Analyze the case when we use a kernel estimator with a Gaussian kernel (i.e., analyze the changes with the label with respect to different parameters of the Gaussian).

A gaussian kernel is weighted. The further away from the particular data point of interest the less weight it has. So, if we are using a gaussian kernel it will behave differently giving more weight to the closer x^t samples. If I look at the circle above when k=20, for example, a gaussian kernel will classify Δ as a ★ since, even though there are less ★'s than O's the ★'s are much closer.

Brian Bertness 1478201 bert0078@umn.edu

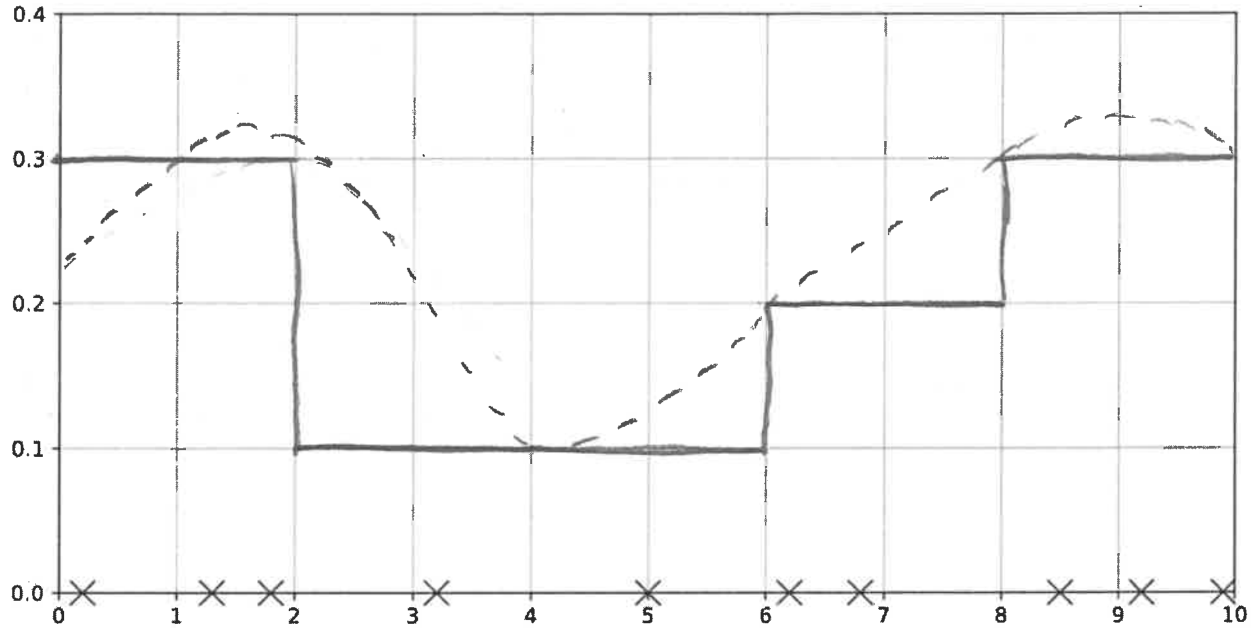
3. Histogram estimator, starting from the origin when $h = 2$. So bins are the intervals:
$(0,2)$, $(2,4)$, $(4,6)$, $(6,8)$, $(8,10)$

There are 10 sample points total so,

interval	# in bin	probability
$(0,2)$	3	$3/10 = .30$
$(2,4)$	1	$1/10 = .10$
$(4,6)$	1	$1/10 = .10$
$(6,8)$	2	$2/10 = .20$
$(8,10)$	3	$3/10 = .30$
	10 total	

Name+Email ID Brian Bertness 1478201 bert~~xx~~78 @umn.edu

3. (**26 points**) Answer the following questions about nonparametric density estimator:



(a) Draw a histogram estimator (start from origin) using $h = 2$ for the following 10 training data points in $\mathbb{R}$: 0.2, 1.3, 1.8, 3.2, 5.0, 6.2, 6.8, 8.5, 9.2, 9.9

(b) Given a test data point $x = 5.5$, what is the predicted density $p(x)$ for the data point?

$p(5.5) = 0.10$

since $5.5 \in (4,6)$

(c) List one possible approach to get a smoother density estimate.

One possible approach would be to use a smooth weight function otherwise known as a kernel function. (pg 192-193) in the book.

(d) Draw an approximate curve when the kernel is used. Discuss the difference with and without kernel used. You do not need to show the calculation.

See line (---) on graph. One difference is the curve is smooth. The other is that that probability is not discrete within the bins any more. Thus, $P(5.5)$ is not going to necessarilly equal 0.10 anymore. You can see how the graph smooths out the probabilities,

Brian Bertness 1478201 bert0078@umn.edu

4. We are given:

$$E(w, w_0 | X) = -\sum_t \left(r^t \log y^t + (1-r^t)\log(1-y^t)\right)$$
$$y^t = \tanh(w^T x^t + w_0)$$

Let $\alpha^t = w^T x^t + w_0$

Then $y^t = \tanh(\alpha^t)$

We are given that if $y = \tan(\alpha) \Rightarrow \frac{\partial y}{\partial \alpha} = 1 - y^2$

(2) Thus $\frac{\partial y^t}{\partial \alpha^t} = 1 - (y^t)^2$

Now, $w_j = w_j + \Delta w_j$ $\underset{j \neq 0}{}$ and $\Delta w_j = -\eta \frac{\partial E}{\partial w_j}$.

So we find $\frac{\partial E}{\partial w_j}$.

$$\frac{\partial E}{\partial w_j} = \sum_t \frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial \alpha^t} \frac{\partial \alpha^t}{\partial w_j}$$

We find each partial derivative in turn

(1) $$\frac{\partial E}{\partial y^t} = -\sum_t \frac{r^t}{y^t} - \frac{1-r^t}{1-y^t}$$

$$= -\sum_t \frac{r^t - \cancel{r^t y^t} - y^t + \cancel{y^t r^t}}{y^t(1-y^t)}$$

$$= -\sum_t \frac{r^t - y^t}{y^t(1-y^t)}$$

Brian Bertness 1478201 bert0678@umn.edu

(2) $$\frac{\partial y^t}{\partial \alpha^t} = 1 - (y^t)^2$$

(3) $$\frac{\partial \alpha^t}{\partial w_j} = \frac{\partial \alpha^t}{\partial w_j}\left[w^T x^t + w_0 \right]$$

$$= \frac{\partial \alpha^t}{\partial w_j}\left[(w_1 x_1 + w_2 x_2 + \cdots + w_t x_t) + w_0 \right]$$

$$= x_j^t$$

Putting (1), (2), and (3) together we get:

$$\frac{\partial E}{\partial w_j} = -\sum_t \frac{r^t - y^t}{y^t(1-y^t)} \cdot 1 - (y^t)^2 \cdot x_j^t$$

Note: $1-(y^t)^2 = 1^2 - (y^t)^2 = (1+y^t)(1-y^t)$

$$= -\sum_t \frac{(r^t - y^t)(1+y^t)\cancel{(1-y^t)}}{y^t\cancel{(1-y^t)}} \cdot x_j^t$$

$$= -\sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} \cdot x_j^t$$

Thus, $w_j = w_j + \Delta w_j$

$$= w_j + -\eta \frac{\partial E}{\partial w_j} = \boxed{w_j + \eta \sum_t \frac{(r^t - y^t)(1+y^t)}{y^t} \cdot x_j^t}$$

Brian Bertness 1478201 bert0678@umn.edu

Now, $w_0 = w_0 + \Delta w_0$ where $\Delta w_0 = -\eta \frac{\partial E}{\partial w_0}$

And $\frac{\partial E}{\partial w_0} = \Sigma_t \frac{\partial E}{\partial y^t} \cdot \frac{\partial y^t}{\partial \alpha^t} \cdot \frac{\partial \alpha^t}{\partial w_0}$

We have calculated $\frac{\partial E}{\partial y^t}$ and $\frac{\partial y^t}{\partial \alpha^t}$ already

(4) $$\frac{\partial \alpha^t}{\partial w_0} = \frac{\partial \alpha^t}{\partial w_0}\left[w^T x^t + w_0\right]$$

$$= \frac{\partial \alpha^t}{\partial w_0}\left[(w_1 x_1 + w_2 x_2 + \cdots + w_t x_t) + w_0\right]$$

$$= 1$$

Putting (1), (2), and (4) together gives

$$\frac{\partial E}{\partial w_0} = -\Sigma_t \frac{(r^t - y^t)(1+y^t)}{y^t} \cdot 1$$

Thus, $w_0 = w_0 + \Delta w_0$

$$= w_0 + -\eta \frac{\partial E}{\partial w_0} = \boxed{w_0 + \eta \Sigma_t \frac{(r^t - y^t)(1+y^t)}{y^t}}$$