DSA 6100: Statistical Methods for Data Science and Analytics

Final Exam, 2019

NOTES:

1. It is a take home exam.
2. Write your answer clearly on this exam paper. DO NOT use your own paper.
3. The exam is worth 40 points.
4. Partial credit may be given for partial answers if possible.
5. There are 5 pages in this exam paper.
6. Upload your R code for specified questions.

I have neither given nor received aid on this examination.

Name (print): Igor Ostaptchenko

Signature, Date: 2019-04-24

1. A given data points are generated by the following process:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3 + \alpha_4 X^4 + \varepsilon$$

where X is continues random variable and ε is a random noise. We have fit the data by the following two models:

Model 1: $Y = aX + b + \varepsilon$

Model 2: $Y = w_0 + w_1 X + \ldots + w_8 X^8 + \varepsilon$

Please indicate the following sentences are false or true. Please explain your reasons.

a) With a fixed number of training set, Model 1 has a high bias compare with Model 2

True. The Model 1 is a linear model with two degrees of freedom and high bias while Model 2 is more flexible and has more variance.

b) With a fixed number of training set, Model 1 has the same variance compare with Model 2

False. The Model 1 is a linear model with two degrees of freedom and high bias while Model 2 is more flexible and has more variance.

c) Model 1 is likely to overfit with 5 training data points

False. The Model 1 gives a smother curve - straight line. It may give large train errors and is "ununder-fitted" compare to Model 2.

d) Training error of Model 1 is likely lower than Model 2

False. The Model 2 is 8 degree polynomial model and will fit the training data assigning lower weights to the members where power of X is more then 4

e) Training residual sum of squares (RSS) for Model 1 is lower than RSS for Model 2

False. The Model 1 being a straight line has higher value of RSS approximating the data of the 4th degree polynomial. The optimal weights found for Model 2 will yield lower RSS then Model 1.

2. The following dataset is classified into two classes.

$(1.5,+1),(3.2,+1),(5.4,-1),(6.2,-),(8.5,-1).$

a) What is the predicted class for a test example at point 4 using k-neatest neighbor (NN) with k=3 (upload your R codes). ***It is -1***

https://github.com/borodark/wsu/blob/master/methods/exam/Q2.R

b) What is the decision boundary associated with this Training set using 3-NN. (upload your R code)

It is -1 when at x>=3.85

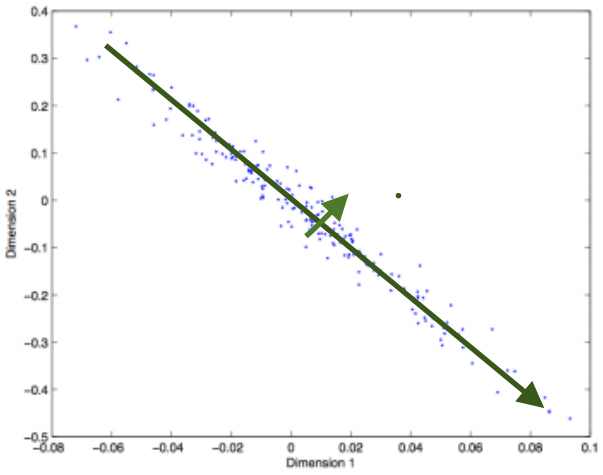
```
> test <- data.frame(x = matrix(c(3.84,3.85)))
> knn3 <-knn(train=train ,test =test, cl=as.factor(y), k=3, prob=TRUE)
> print(knn3)
[1] 1  -1
attr(,"prob")
[1] 0.6666667 0.5000000
```

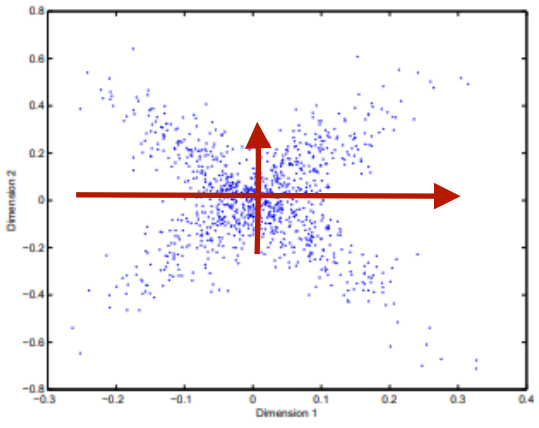
c) Is the Training set linearly separable? Is the accuracy of the 3-NN always 100%? Why?

It is linearly separable. The training set has only 4 pairs. It is only one way this set can be separated into [1 vs 3], hanse the one is measured agains all the others just one time and gives 100%

3.

3.1 Given the following datasets represented in graph, draw the first and second principle components on each plot





3.2 Given the air pollution table (available in Canvas) which represents measurements of air pollution variables recorded at 12 noon on 42 different days at a location in Los Angeles, please answer the following questions.

a) A principal component analysis is performed on the observed data. Based on the output, what percentage of the total variation is explained by the first principal component? (upload your R code)

b) How many principal components are needed to obtain a good approximation of the data? Why? (upload your R code)

c) Provide an interpretation of the first principal component.

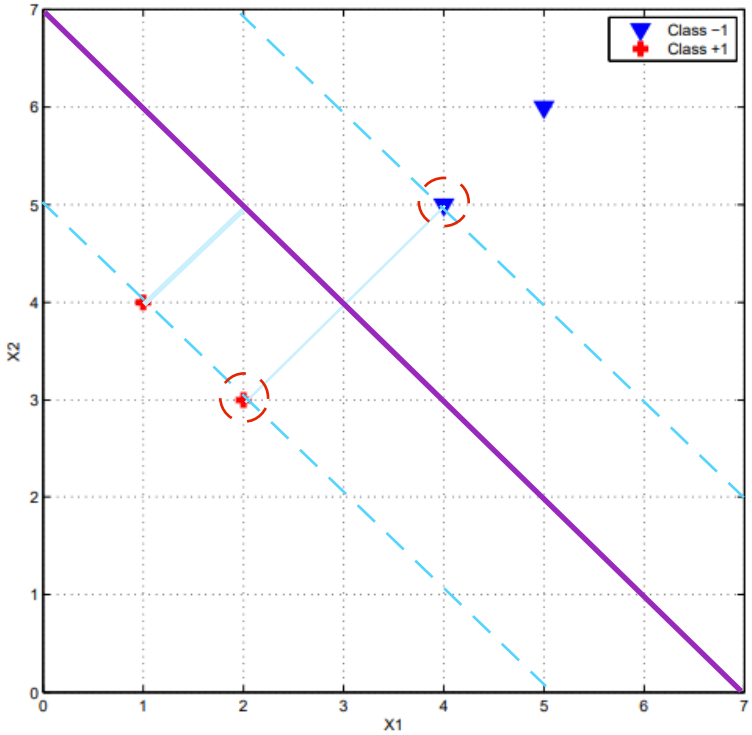
d) Now, a principal component analysis is performed using the correlation matrix. How many components are needed to obtain a good approximation of the data? (upload your R code)

e) Provide interpretations for the first two principal components.

4) Let a configuration of the k-means algorithm correspond to the k way partition generated by the clustering at the end of each iteration. Is it possible for the k-means algorithm to revisit a configuration? Justify how your answer proves that the k-means algorithm converges in a finite number of steps.

Because the k means algorithm converges when number of partitions cease to change in successive iterations, the k -the number of partitions has to change after every iteration. Eventually the k means algorithm will run out of configurations, and converge. The mean squared error monotonically decreases hence it is impossible to revisit a configuration. The maximum number of iterations corresponds to the number of k way partitions possible on a set of n objects : S(n,k) where S are Stirling numbers of the 2nd kind.

5) Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 4 points shown in Figure bellow. This dataset consists of two examples with class label -1 (denoted with plus), and two examples with class label +1 (denoted with triangles).



a) Find the weight vector w and bias b. What's the equation corresponding to the decision boundary?

w = [-1, -1]
b = 7

x2 = 7 - 1 * x1

b) Circle the support vectors and draw the decision boundary

(2,3) and (4,5)

https://github.com/borodark/wsu/blob/master/methods/exam/Q5.R

6)

6-1) For a data set with p features, of which q will eventually enter the model, stepwise feature selection will test approximately how many models? *1 + p(p + 1)/2*

6-2) Suppose we have a regularized linear regression model: $\arg\min_w \| Y - Xw \|_2^2 + \lambda \| w \|_p^p$.

a) What is the effect of increasing λ on bias and variance?

Increasing λ will cause the variance to shrink faster penalizing assigning significant weights to more features. It has less effect on the bias

b) What is the effect of increasing p on bias and variance ($p \geq 1$) if the weights are all larger than 1

Increasing p in given circumstances will cause the variance to shrink faster but again not affecting the bias with the same degree

6-3) Please answer the following true/false questions:

a) The Linear Discriminant Analysis (LDA) classifier computes the direction maximizing the ratio of between-class variance over within-class variance.

True

b) Nearest neighbors is a parametric method.

False, K-nn do not have fixed numbers of parameters in the model.

c) K-means is a supervised method.

False, K-mean is an unsupervised learning technique (no dependent variable)

d) A cubic spline with Knots has K + 5 degrees of freedom

False, it is K+4