CIS 490 MACHINE LEARNING HOMEWORK 2

(Full Score: **45** + 4 bonus points)

October 13, 2017 Fall

Group number:

Group leader:

Group members and student IDs:

(Please mark “G” after names for Graduate Students in your group)

INSTRUCTOR: Dr. Julia Hua Fang

**Please write your answers for each question:**

1. **4 points.**
2. use the multiple linear regression model to fit the “Demand estimation” data below (you can copy and paste the data into a file and save it as, e.g., \*.txt or \*.xlsx); compute SSE, MSE; and predict the number of weekly riders of a new city, given its known values on X1 = $35, X2=$1,600,000, X3 = $12,000 and X4= $ 110.

1 Bonus point: Plot a graph where you have a best fit line for this dataset.

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_Reg.R” (e.g., Group1\_Reg.R) or “Group#\_Reg.m” (e.g., “Group1\_Reg.m”).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Y** | **X1** | **X2** | **X3** | **X4** |
| **City** | **Number of weekly riders** | **Price per week** | **Population of city** | **Monthly income of riders** | **Average parking rates per month** |
| 1 | 192,000 | $15 | 1,800,000 | $5,800 | $50 |
| 2 | 190,400 | $15 | 1,790,000 | $6,200 | $50 |
| 3 | 191,200 | $15 | 1,780,000 | $6,400 | $60 |
| 4 | 177,600 | $25 | 1,778,000 | $6,500 | $60 |
| 5 | 176,800 | $25 | 1,750,000 | $6,550 | $60 |
| 6 | 178,400 | $25 | 1,740,000 | $6,580 | $70 |
| 7 | 180,800 | $25 | 1,725,000 | $8,200 | $75 |
| 8 | 175,200 | $30 | 1,725,000 | $8,600 | $75 |
| 9 | 174,400 | $30 | 1,720,000 | $8,800 | $75 |
| 10 | 173,920 | $30 | 1,705,000 | $9,200 | $80 |
| 11 | 172,800 | $30 | 1,710,000 | $9,630 | $80 |
| 12 | 163,200 | $40 | 1,700,000 | $10,570 | $80 |
| 13 | 161,600 | $40 | 1,695,000 | $11,330 | $85 |
| 14 | 161,600 | $40 | 1,695,000 | $11,600 | $100 |
| 15 | 160,800 | $40 | 1,690,000 | $11,800 | $105 |
| 16 | 159,200 | $40 | 1,630,000 | $11,830 | $105 |
| 17 | 148,800 | $65 | 1,640,000 | $12,650 | $105 |
| 18 | 115,696 | $102 | 1,635,000 | $13,000 | $110 |
| 19 | 147,200 | $75 | 1,630,000 | $13,224 | $125 |
| 20 | 150,400 | $75 | 1,620,000 | $13,766 | $130 |
| 21 | 152,000 | $75 | 1,615,000 | $14,010 | $150 |
| 22 | 136,000 | $80 | 1,605,000 | $14,468 | $155 |
| 23 | 126,240 | $86 | 1,590,000 | $15,000 | $165 |
| 24 | 123,888 | $98 | 1,595,000 | $15,200 | $175 |
| 25 | 126,080 | $87 | 1,590,000 | $15,600 | $175 |
| 26 | 151,680 | $77 | 1,600,000 | $16,000 | $190 |
| 27 | 152,800 | $63 | 1,610,000 | $16,200 | $200 |

1. **8 points**

Use logistic regression model to fit the “Graduate Admission” data, in “logistic.txt” (see the data file under Homework2 in Week 6 at myCourses)

Scenario: A researcher is interested in how three variables, (X1=GRE, Graduate Record Exam scores; X2=GPA, grade point average; and X3: Rank, prestige of the undergraduate institution) effect admission into graduate school (the response variable, Y, admit/don’t admit, is a binary variable).

1. If a new student has X1 = 720; X2=3.8; X3= 2, do you admit or not admit him? (1 point)
2. Fill in the number of students in the confusion matrix (2 points)

|  |  |  |
| --- | --- | --- |
|  | Predicted |  |
| Observed/True (Y) | 0= don’t admit | 1= Admit |
| 0= don’t admit |  |  |
| 1= Admit |  |  |

1. Compute misclassification errors, sensitivity and specificity. (3 points)

Please attach your R or Matlab code and output in your word file as Appendices. Please also save and name your scripts as “Group#\_Logit.R” or “Group#\_Logit.m” and any related R or Matlab output files with the same file names, e.g “Group#\_Logit.xxx”.

1. What’s relationship among probability, odds and log odds? Use an example to demonstrate their relationship. What’s the range for probability, odds and log odds? (2 points)
2. 8 **points**
3. What is Bias-variance trade off? (0.5 point)
4. What is overfitting? (0.5 point)
5. What are the two common types of regularization in linear regression? (0.5 point)
6. Write out the objective functions for these two regularization types. Please identify which term is for SSR and which term is for shrinkage penalty term in each function. What is λ called in these regularized regressions? (2 points)
7. Write out the corollaries of Ridge Regression. (0.5 point)
8. what is the major difference between the two regularization types? (hint: why one of them is preferred in real world practice) (1 point)
9. Download credit data from under Homework2 in Week 6 at myCourses and use Lasso regression with cross-validation to select a parsimonious model (Occam’s Razor), ie. what attributes will enter into your final Lasso regression model and what optimal value will you pick for λ? Please replicate the following two graphs. (3 points)

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_Lasso.R” (e.g., Group1\_Lasso.R) or “Group#\_Lasso.m” (e.g., “Group1\_Lasso.m”).

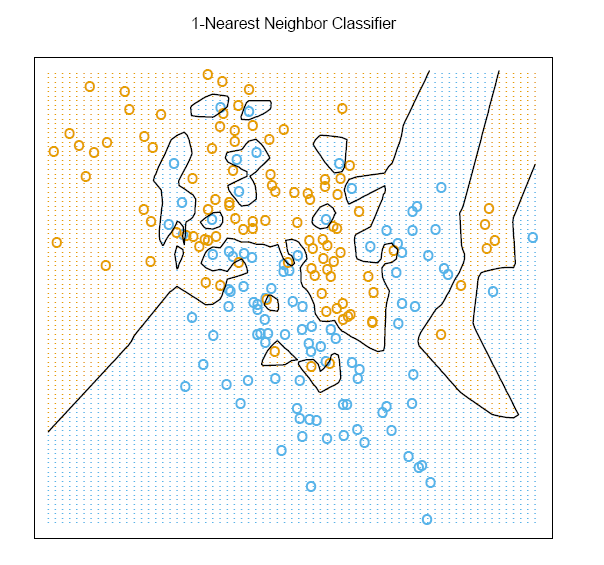
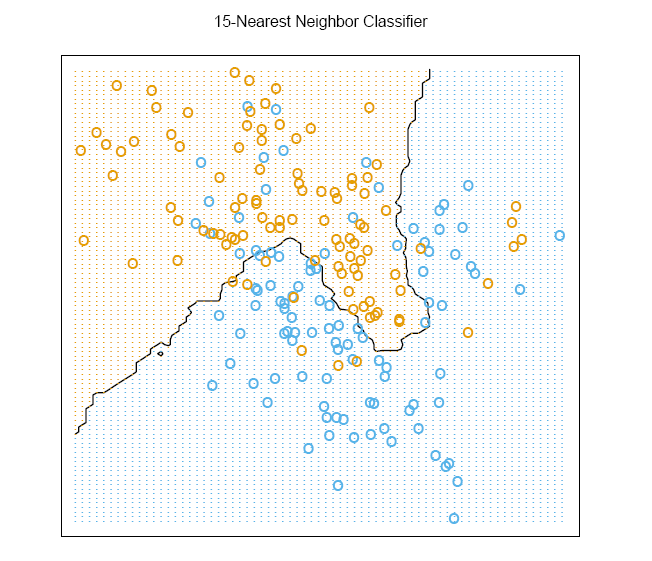
 

1 Bonus point: What validation indices can be used to pick the optimal λ in regularized linear regression?

1. **7 points**
2. What are the differences between parametric and non-parametric methods? (1.5 point)
3. What is called curse of dimensionality? (0.5 point)
4. Why should we consider rescaling the attributes when using KNN? (0.5 point)
5. What options can we use to determine the class from a nearest neighbor list? (0.5 point)
6. (c) What method can we use to choose the number of k for nearest neighbor classification (0.5 point)

(d) Given data below, compute the distance between nominal attribute values: d(Married,Divorced). Please show you calculation steps below. (1 point)



1. KNN is a local or global method? (0.5 point)
2. Which decision boundary in the graph seems to derive from more nearest neighbors? (0.5 point)

(i) (ii)

(g) Please download IRIS data from UCI machine learning repository (<https://archive.ics.uci.edu/ml/datasets.html> ), and run KNN on IRIS data, compute the misclassification error rates. (1.5 points)

Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_KNN.R” (e.g., Group1\_KNN.R) or “Group#\_KNN.m” (e.g., “Group1\_IRIS.m”). ( points)

2 Bonus Points: Using the same IRIS data, use cross validation to choose the optimal number of nearest neighbors for KNN. Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_KNNCV.R” (e.g., Group1\_KNNCV.R) or “Group#\_KNNCV.m” (e.g., “Group1\_KNNCV.m”).

1. **6 points**
2. What are the differences between Frequentist and Bayesian approaches? (2 point)

(b) What are the four terms called in the following equation from Bayesian classification approach. (1 point)



(c) Given the evidence/observed data below



what is the class label for a new animal with the following attribute values? Please show your calculation steps using Naïve Bayes Formulas. (3 points)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Give Birth | Can fly | Live in water | Have Legs | Class |
| Yes | no | yes | no | ? |

1. 8 **points**
2. What is the relationship between LDA and Naïve Bayes? ( points)
3. What is the relationship between LDA and Logistic regression? ( points)
4. When do we use Logistic regression, LDA or Naïve Bayes? ( points)
5. Please download IRIS data from UCI machine learning repository (<https://archive.ics.uci.edu/ml/datasets.html> ), compare the misclassification error rates, false positive rates and false negative rates between Naïve Bayes and LDA. Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_IRIS.R” (e.g., Group1\_IRIS.R) or “Group#\_IRIS.m” (e.g., “Group1\_IRIS.m”). ( points)

1. **4 points**

Generate ROC curve using the following tables. Please attach your R or Matlab code and output in your word file as Appendices. Please (1) attach your R or Matlab codes and output in your word file as Appendices; (2) save and name your original scripts as “Group#\_ROC.R” (e.g., Group1\_ROC.R) or “Group#\_ROC.m” (e.g., “Group1\_ROC.m”).

|  |  |  |
| --- | --- | --- |
| **New measure** | **Disease** | **Non-Disease** |
| 5 or less | 18 | 1 |
| 5.1 - 7 | 7 | 17 |
| v7.1 - 9 | 4 | 36 |
| 9 or more | 3 | 39 |
| **Totals:** | 32 | 93 |
| **Cut point** | Sensitivity | Specificity |
| **5** | 0.56 | 0.99 |
| **7** | 0.78 | 0.81 |
| **9** | 0.91 | 0.42 |