Assignment 5

Travis Pressler
CSCI B490
Indiana University

November, 2014

Problem 1:

My three data sets are **iris.data**, which is concerned with classification of flowers based on their physical measurements; **sperm.data**, which is concerned with classification of sperm donors into abnormal and normal sperm types; and **haberman.data**, which is concerned with predicting cancer deaths based on the attributes of patients.

Part a:

Run **Problem1a.m** to see the results.

Part b:

Here are some sample runs of 10-fold cross-validation on the three data sets:

Table 1: Problem 1b run data:

Data Set	Algorithm	Performance Metric	Test #1	Test #2	Test #3	Test Avg.
iris.data	Linear Regression	Accuracy	0.933333	0.953333	0.926667	0.937778
		AUC	0.920000	0.940000	0.930000	0.930000
	K Nearest Neighbors	Accuracy	0.960000	0.953333	0.940000	0.951111
		AUC	0.955000	0.940000	0.930000	0.941667
sperm.data	Linear Regression	Accuracy	0.750000	0.770000	0.790000	0.770000
		AUC	0.606061	0.617424	0.628788	0.617424
	K Nearest Neighbors	Accuracy	0.750000	0.780000	0.780000	0.770000
		AUC	0.426136	0.551136	0.515152	0.497475
haberman.data	Linear Regression	Accuracy	0.754902	0.745098	0.751634	0.750545
		AUC	0.584444	0.565926	0.574321	0.574897
	K Nearest Neighbors	Accuracy	0.715686	0.728758	0.715686	0.720043
		AUC	0.569630	0.578519	0.585432	0.577860

The iris.data data set has fantastic classification results for linear regression and K Nearest Neighbors. The sperm.data data set has positive results but much smaller Area Under the Curve than the Iris data set, and its KNN AUC is poor. The Haberman data set performs adequately for both linear regression and K Nearest Neighbors

Part c:

Run **Problem1c.m** to see the results. Here are some sample runs of 10-fold cross-validation on the three data sets:

Table 2: Problem 1c Run Data

Data Set	Algorithm	Performance Metric	Test #1	Test #2	Test #3	Test Avg.
iris.data	Neural Network	Accuracy	0.966670	0.960000	0.973333	0.966668
		AUC	0.946667	0.960000	0.970000	0.958889
	Decision Trees	Accuracy	0.946667	0.946667	0.946667	0.946667
		AUC	0.935000	0.940000	0.935000	0.936667
sperm.data	Neural Network	Accuracy	0.860000	0.840000	0.840000	0.846667
		AUC	0.632576	0.621212	0.549242	0.60101
	Decision Trees	Accuracy	0.840000	0.860000	0.860000	0.853333
		AUC	0.549242	0.560606	0.560606	0.556818
haberman.data	Neural Network	Accuracy	0.735294	0.754902	0.748366	0.746187
		AUC	0.575062	0.612099	0.595802	0.594321
	Decision Trees	Accuracy	0.663399	0.683007	0.673203	0.673203
		AUC	0.538025	0.563210	0.556543	0.552593

The Iris data set has excellent classification results for both neural networks and decision trees. The sperm data set has a good AUC for neural networks but lackluster results for decision trees. The haberman data set performs similar to the sperm data set when both algorithms are applied.

Problem 2:

Part a:

Run **Problem2a_P1.m** to see the results of z-score normalization on Linear Regression and K Nearest Neighbors. Run **Problem2a_P2.m** to see the results of z-score normalization on Neural Networks and Decision Trees.

Part b:

Run **Problem2b_P1.m** to see the results of min-max normalization on Linear Regression and K Nearest Neighbors. Run **Problem2b_P2.m** to see the results of min-max normalization on Neural Networks and Decision Trees.

Problem 3:

Run **Problem3_P1** to see single-feature analysis for linear regression and K Nearest Neighbors. Run Run **Problem3_P2** to see single-feature analysis

for neural networks and decision trees.

I predict that the most important features of Haberman's Survival Data Set *(haberman.data)* will be Attribute 3 (Number of positive axillary nodes detected), then Attribute 2 (Patient's year of operation).

- Top features for Linear Regression: (#1 and #2=0.5)(#3=0.572099)
- Top features for K Nearest Neighbors:(#1=.453086)(#2=.479012)(#3=.549877)
- Top features for Neural Networks: (#1 and #2=0.5)(#3=0.53358)
- Top features for Decision Trees: (#1=0.485185)(#2=0.5)(#3=0.555556)

I was correct in my predictions, but these AUC's are all very close to 0.5, and therefore don't indicate a very strong classification.

I predict that the most important features of the Iris Data Set *(iris.data)* will be Attribute 4 (Petal width), then Attribute 3 (Petal length).

- Top features for Linear Regression: (#2=.485)(#1=.8)(#3=.95)(#4=.95)
- Top features for K Nearest Neighbors: (#2=.485)(#1=.675)(#3=.89)(#4=.945)
- Top features for Neural Networks: (#1=.5)(#2=.52)(#3=.76)(#4=.94)
- Top features for Decision Trees: (#2=.475)(#1=.51)(#3=.76)(#4=0.945)

It's pretty clear from these results that the order of importance is 4 to 1 (descending). This corresponds to my prediction because I guessed correctly that petals would have more evolutionary pressure to differentiate themselves than would sepals.

I predict that the most important features of the Fertility Data Set *(sperm.data)* will be Attribute 8 (Smoking Habit), then Attribute 7 (Frequency of Alcohol Consumption), then Attribute 4(Accident or Serious Trauma) .

- Top features for Linear Regression: (all attributes = 0.5)
- Top features for K Nearest Neighbors: (#7 and #9=.477273)(#2=.482955)(all other attributes=0.5)
- Top features for Neural Networks: (#9=0.494318)(#1=.496212)(#2, #3, #4,#5, #6, #7=0.5)

- Top features for Decision Trees: (#9=0.494318)(#1,#2, #3, #4, #5,#6, #7=0.5)

It appears that this may not be the best data set to attempt classification using these algorithms. This may indicate that these attributes are not correlated with the class.