CS 484: Introduction to Machine Learning

Spring 2021 Assignment 4 Answer Key

Question 1 (40 Points)

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge. The data contain transaction history for customers that ended up purchasing a policy. For each Customer ID, we know the quote history and the purchased coverage options.

The data is available on the Blackboard as **Purchase_Likelihood.csv**.

1. It contains 665,249 observations on 97,009 unique Customer ID.
2. The nominal target variable is **insurance** that has these categories 0, 1, and 2
3. The nominal features are (categories are inside the parentheses):
   a. **group_size**. *How many people covered under the policy (1, 2, 3, or 4)*?
   b. **homeowner**. *Whether the customer owns a home or not (0 = No, 1 = Yes)*?
   c. **married_couple**. *Does the customer group contain a married couple (0 = No, 1 = Yes)*?

You will train a Naïve Bayes model without any smoothing using all the observations in the **Purchase_Likelihood.csv**. In other words, the Laplace/Lidstone alpha is zero. Please answer the following questions based on your model.

a) (5 points) Show in a table the frequency counts and the Class Probabilities of the target variable.

insurance	0	1	2
Frequency Count	143,691	426,067	95,491
Class Probability	0.2159958	0.6404624	0.1435417

b) (5 points) Show the crosstabulation table of the target variable by the feature **group_size**. The table contains the frequency counts.

group_size	insurance		
	0	1	2
1	115,460	329552	74293
2	25,728	91065	19600
3	2,282	5069	1505
4	221	381	93

c) (5 points) Show the crosstabulation table of the target variable by the feature **homeowner**. The table contains the frequency counts.

homeowner	insurance		
	0	1	2
0	78,659	183,130	46,734
1	65,032	242,937	48,757

d) (5 points) Show the crosstabulation table of the target variable by the feature **married_couple**. The table contains the frequency counts.

Married_couple	insurance		
	0	1	2
0	117,110	333,272	75,310
1	26,581	92,795	20,181

e) (5 points) Calculate the Cramer's V statistics for the above three crosstabulations tables. Based on the Cramer's V statistics, which feature has the strongest association with the target insurance?

Feature	Cramer's V
group_size	0.0271020
homeowner	0.0970864
married_couple	0.0324216

Homeowner has the largest Cramer's v

f) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for insurance = 0, 1, 2 based on the Naïve Bayes model that includes features group_size, homeowner, and married_couple. List your answers in a table with proper labeling.

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.2697219	0.5801334	0.1501447
1	0	1	0.2327892	0.6142186	0.1529922
1	1	0	0.1940379	0.6696590	0.1363031

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	1	1	0.1649350	0.6982780	0.1367869
2	0	0	0.2311433	0.6165185	0.1523382
2	0	1	0.1980156	0.6479068	0.1540776
2	1	0	0.1636275	0.7002878	0.1360847
2	1	1	0.1382742	0.7259550	0.1357709
3	0	0	0.3082194	0.5159242	0.1758564
3	0	1	0.2683111	0.5509509	0.1807380
3	1	0	0.2269718	0.6096118	0.1634164
3	1	1	0.1943695	0.6404098	0.1652207
4	0	0	0.3754904	0.4878101	0.1366995
4	0	1	0.3307434	0.5270983	0.1421583
4	1	0	0.2821727	0.5881965	0.1296309
4	1	1	0.2439303	0.6237660	0.1323037

group_size	homeowner	married_couple	Prob(insurance = 0)	Prob(insurance = 1)	Prob(insurance = 2)
1	0	0	0.227037	0.627593	0.145370
1	0	1	0.214391	0.637467	0.148142
1	1	0	0.205588	0.654128	0.140284
1	1	1	0.193842	0.663414	0.142744
2	0	0	0.238441	0.614462	0.147097
2	0	1	0.225342	0.624635	0.150024
2	1	0	0.216281	0.641528	0.142192
2	1	1	0.204079	0.651128	0.144794
3	0	0	0.250201	0.601084	0.148715
3	0	1	0.236653	0.611546	0.151801
3	1	0	0.227342	0.628652	0.144006
3	1	1	0.214684	0.638559	0.146756
4	0	0	0.262308	0.587475	0.150218
4	0	1	0.248318	0.598215	0.153467
4	1	0	0.238767	0.615513	0.145720
4	1	1	0.225656	0.625720	0.148624

g) (5 points) Based on your model, determine the value combination of group_size, homeowner, and married_couple that will yield the maximum value for this odds Prob(insurance = 1) / Prob(insurance = 2)? What is that maximum odds value?

We first calculate the odds Prob(insurance = 1) / Prob(insurance = 2). When group_size = 2, homeowner = 1, and married_couple = 1, the odds attains its maximum value of 5.3469.

group_size	homeowner	married_couple	Prob(insurance = 1)/Prob(insurance = 2)
1	0	0	3.8638
1	0	1	4.0147
1	1	0	4.9130
1	1	1	5.1049
2	0	0	4.0470
2	0	1	4.2051
2	1	0	5.1460
2	1	1	5.3469
3	0	0	2.9338
3	0	1	3.0483
3	1	0	3.7304
3	1	1	3.8761
4	0	0	3.5685
4	0	1	3.7078
4	1	0	4.5375
4	1	1	4.7147

Question 2 (60 points)

The **SpiralWithCluster.csv** contains four variables.

Name	Description	Measurement Level	Role
Id	Case Identifier	Nominal	Identifier
X	x-coordinate	Interval	Feature
Y	y-coordinate	Interval	Feature
SpectralCluster	Cluster Identifier	Binary	Target

Please use the Support Vector Machine (SVM) algorithm to classify SpectralCluster. You will use the sklearn.svm.SVC function with the following specifications.

1. The linear kernel
2. The decision function shape is One Over Rest (OVR)
3. No limit on the number of iterations
4. The random seed is 20210325

Please answer the following questions based on your model.

a) (5 points) What is the equation of the separating hyperplane in the Slope-Intercept form? Please state the coefficients up to seven decimal places.

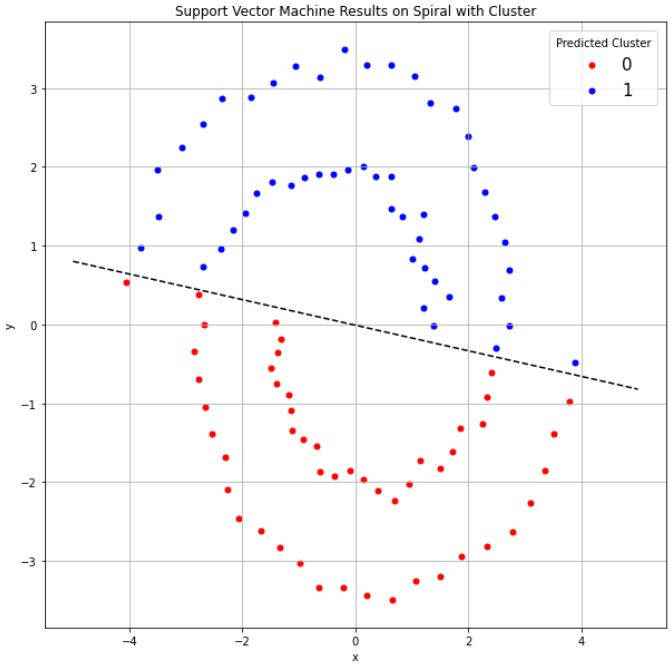
The equation is 0.0033450 + 0.0533351 * X + 0.3286838 * Y = 0.

Or Y = -0.010177 – 0.162269 * X

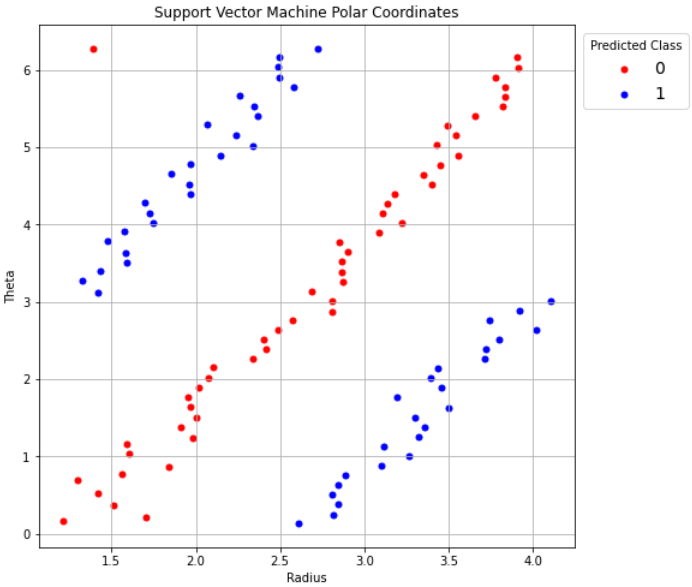
b) (5 points) What is the misclassification rate?

The misclassification rate is 0.5.

c) (10 points) Please plot the y-coordinate against the x-coordinate in a scatterplot. Please color-code the points using the predicted SpectralCluster (0 = Red and 1 = Blue). Besides, plot the hyperplane as a dotted line to the graph. To obtain the full credits, you should properly label the axes, the legend, and the chart title. Also, please add grid lines to the axes.



d) (10 points) Please express the data in polar coordinates. Please plot the theta-coordinate against the radius-coordinate in a scatterplot. Please color-code the points using the SpectralCluster variable (0 = Red and 1 = Blue). To obtain the full credits, you should properly label the axes, the legend, and the chart title. Also, please add grid lines to the axes.



e) (10 points) You should expect to see three distinct strips of points and a single point. Since the SpectralCluster variable has two values, you will create another variable, named Group, and use it as

the new target variable. Use your method to generate this Group variable. However, the Group variable must have four values. Value 0 for the single point on the upper left corner of the chart in (d), values 1, 2, and 3 for the next three strips of points.

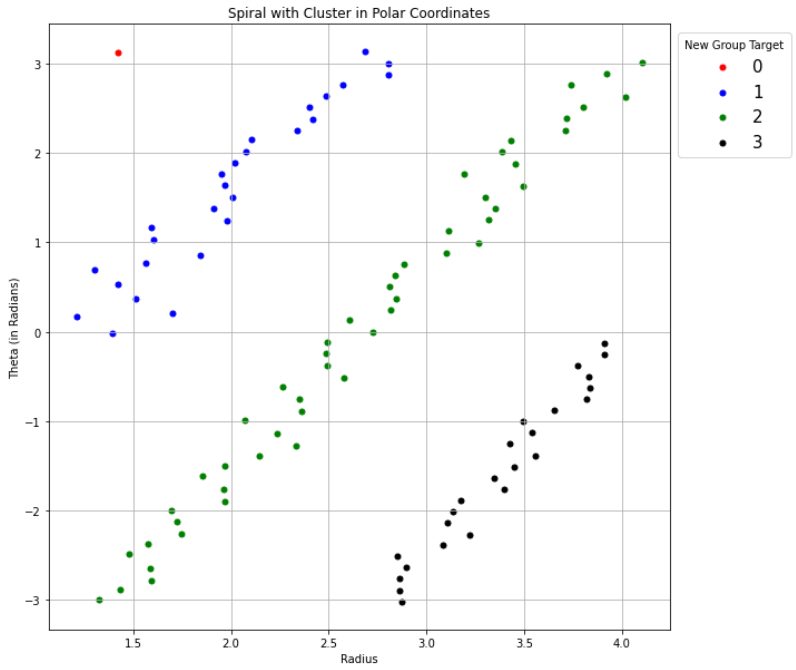
Please plot the theta-coordinate against the radius-coordinate in a scatterplot. Please color-code the points using the new Group target variable (0 = Red, 1 = Blue, 2 = Green, 3 = Black). To obtain the full credits, you should properly label the axes, the legend, and the chart title. Also, grid lines should be added to the axes.

There are many ways to create the new Group target variable. One way to do it is below.

```
trainData['radius'] = numpy.sqrt(trainData['x']**2 + trainData['y']**2)
trainData['theta'] = numpy.arctan2(trainData['y'], trainData['x'])

g1 = numpy.where(2*trainData['radius'] - trainData['theta'] > 1, 1, 0)
g2 = numpy.where(2*trainData['radius'] - trainData['theta'] > 4, 1, 0)
g3 = numpy.where(2*trainData['radius'] - trainData['theta'] > 7, 1, 0)

trainData['Group'] = g1 + g2 + g3
```



f) (10 points) Since the graph in (e) has four separate but neighboring segments, we will apply the Support Vector Machine algorithm differently. Instead of applying SVM once on a multi-class target variable, you will SVM three times, each on a pair of groups.

SVM 0: Group 0 versus Group 1

SVM 1: Group 1 versus Group 2

SVM 2: Group 2 versus Group 3

Please give the equations of the three hyperplanes.

SVM 0: 1. 46912508 + 0. 93378415*radius – 0. 45380249*theta = 0

SVM 1: – 0. 87689426 + 1. 89209533*radius – 0. 89613249*theta = 0

SVM 2: – 4. 13284488 + 2. 01258355*radius – 0. 83756164*theta = 0

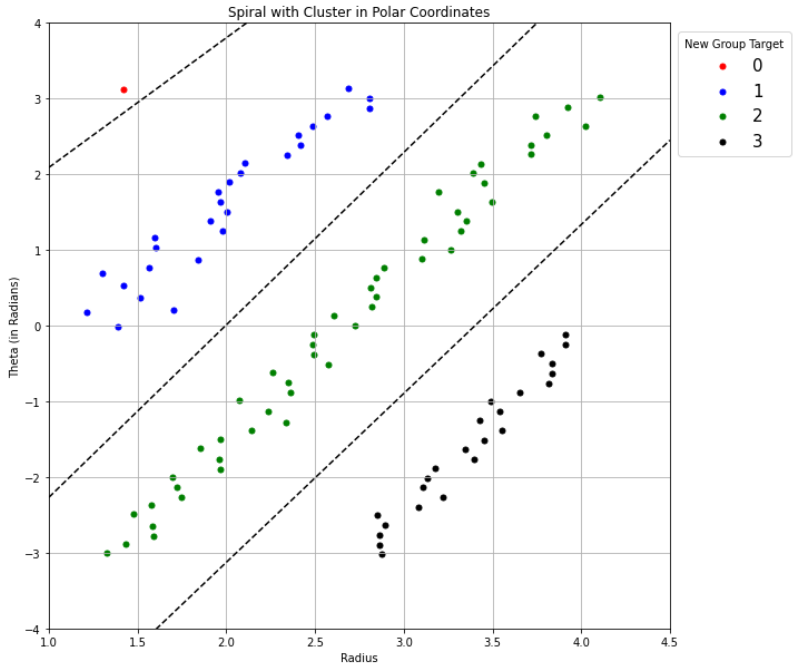
Or

Theta = 2.058*radius + 3.237

Theta = 2.111*radius – 0.9786

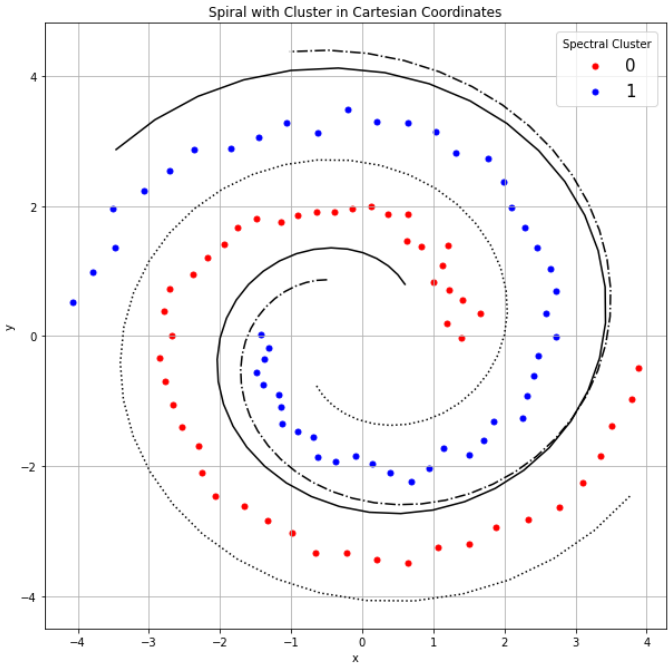
Theta = 2.403*radius – 4.934

g) (5 points) Please plot the theta-coordinate against the radius-coordinate in a scatterplot. Please color-code the points using the new Group target variable (0 = Red, 1 = Blue, 2 = Green, 3 = Black). Please add the hyperplanes to the graph. To obtain the full credits, you should properly label the axes, the legend, and the chart title. Also, grid lines should be added to the axes.



h) (5 points) Convert the observations along with the hyperplanes from the polar coordinates back to the Cartesian coordinates. Please plot the y-coordinate against the x-coordinate in a scatterplot. Please color-code the points using the SpectralCluster (0 = Red and 1 = Blue). Besides, plot the hyper-curves as dotted lines to the graph. To obtain the full credits, you should properly label the axes, the legend, and the chart title. Also, grid lines should be added to the axes.

Based on your graph, which hypercurve do you think is not needed?



The curve that corresponds to the SVM0 is not needed. Besides, it gets too close to the 0^{th} spectral cluster in the center and too far away from the 0^{th} spectral cluster in the edge.