CS 584-04: Machine Learning

Autumn 2019 Assignment 4

Question 1 (50 points)

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge which is open. The data contain transaction history for customers that ended up purchasing a policy. For each Customer ID, you are given their quote history and the coverage options they purchased.

The data is available on the Blackboard as Purchase_Likelihood.csv. It contains 665,249 observations on 97,009 unique Customer ID. You will build a multinomial logistic model with the following specifications.

1. The nominal target variable is **A** which have these categories 0, 1, and 2
2. The nominal features are (categories are inside the parentheses):
   a. **group_size**. How many people will be 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)?
3. Include the Intercept term in the model
4. Enter the five model effects in this order: group_size, homeowner, married_couple, group_size * homeowner, and homeowner * married_couple (No forward or backward selection)
5. The optimization method is Newton
6. The maximum number of iterations is 100
7. The tolerance level is 1e-8.
8. Use the sympy.Matrix().rref() method to identify the non-aliased parameters

Please answer the following questions based on your model.

a) (5 points) List the aliased parameters that you found in your model.
Ans:
'group_size_4', 'homeowner_1', 'married_couple_1',
'group_size_1 * homeowner_1', 'group_size_2 * homeowner_1',
'group_size_3 * homeowner_1', 'group_size_4 * homeowner_0',
'group_size_4 * homeowner_1', 'homeowner_0 * married_couple_1',
'homeowner_1 * married_couple_0', 'homeowner_1 * married_couple_1'

b) (5 points) How many degrees of freedom do you have in your model?
Ans: 20

c) (10 points) After entering a model effect, calculate the Deviance test statistic, its degrees of freedom, and its significance value between the current model and the previous model. List your Deviance test results by the model effects in a table.
Ans:

Model Effects	**Difference in Deviance**	**Difference in Degrees of Freedom**	**Significance Value**
Intercept + group_size	987.576	6	4.34e-210
Intercept + group_size + homeowner	5867.781	2	0.0
Intercept + group_size + homeowner + married_couple	84.578	2	4.30e-19
Intercept + group_size + homeowner + married_couple + group_size * homeowner	254.078	6	5.51e-52
Intercept + group_size + homeowner + married_couple + group_size * homeowner + homeowner * married_couple	70.842	2	4.138e-16

d) (5 points) Calculate the Feature Importance Index as the negative base-10 logarithm of the significance value. List your indices by the model effects. Ans:

Model Effects	Feature Importance Index
Intercept + group_size	209.361
Intercept + group_size + homeowner	+inf
Intercept + group_size + homeowner + married_couple	18.365
Intercept + group_size + homeowner + married_couple + group_size * homeowner	51.258
Intercept + group_size + homeowner + married_couple + group_size * homeowner + homeowner * married_couple	15.383

e) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilties for A = 0, 1, 2 based on the multinomial logistic model. List your answers in a table with proper labelling. Ans: Below are cases for values of three features in order group_size, homeowner, and married_couple

Cases	A = 0	A = 1	A = 2
1, 0, 0	0.259	0.589	0.151
1, 0, 1	0.260	0.592	0.147
1, 1, 0	0.183	0.682	0.134
1, 1, 1	0.154	0.709	0.136
2, 0, 0	0.221	0.621	0.156
2, 0, 1	0.222	0.624	0.153
2, 1, 0	0.202	0.659	0.137
2, 1, 1	0.170	0.689	0.139

Cases	A = 0	A = 1	A = 2
3, 0, 0	0.239	0.604	0.153
3, 0, 1	0.239	0.607	0.152
3, 1, 0	0.301	0.531	0.167
3, 1, 1	0.259	0.567	0.173
4, 0, 0	0.194	0.669	0.133
4, 0, 1	0.194	0.672	0.132
4, 1, 0	0.387	0.484	0.127
4, 1, 1	0.339	0.526	0.134

f) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?
Ans: Max Odd value = 4.603
The values of group_size, homeowner, married_couple respectively are 1, 1, 1

g) (5 points) Based on your model, what is the odds ratio for group_size = 3 versus group_size = 1, and A = 2 versus A = 0? Mathematically, the odds ratio is (Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1).
Ans: Odds Ratio = 1.0249

h) (5 points) Based on your model, what is the odds ratio for homeowner = 1 versus homeowner = 0, and A = 0 versus A = 1? Mathematically, the odds ratio is (Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0).
Ans: Odds Ratio = 0.6232

Question 2 (50 points)

You are asked to build a Naïve Bayes model using the same Purchase_Likelihood.csv. The model specifications are:

1. No smoothing is needed. Therefore, the Laplace/Lidstone alpha is zero
2. The nominal target variable is **A** which have these categories 0, 1, and 2
3. The nominal features are (categories are inside the parentheses):
   a. **group_size**. How many people will be 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)?

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.
Ans:

A	Frequency Count	Class Probabilities
0	143691	0.2159
1	426067	0.6404
2	95491	0.1435

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

A	1	2	3	4
0	115460	25728	2282	221
1	329552	91065	5069	381
2	74293	19600	1505	93

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

A	0	1
0	78659	65032
1	183130	242937
2	46734	48757

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

A	0	1
0	11710	26581
1	333272	92795
2	75310	20181

e) (10 points) Calculate the Cramer's V statistics for the above three crosstabulations tables. Based on these Cramer's V statistics, which feature has the largest association with the target A?

Ans: Feature 'homeowner' has the largest association with the target A

f) (5 points) Based on the assumptions of the Naïve Bayes model, express the joint probability Prob(A = a, group_size = g, homeowner = h, married_couple = m) as a product of the appropriate probabilities.

Ans: P(A=a)*P(group_size=g|A=a)*P(homeowner=h|
A=a)*P(married_couple=m|A=a)

g) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for A = 0, 1, 2 based on the Naïve Bayes model. List your answers in a table with proper labelling.

Cases	A=0	A=1	A=2
1, 0, 0	0.227	0.627	0.145
1, 0, 1	0.214	0.637	0.148
1, 1, 0	0.205	0.654	0.140
1, 1, 1	0.193	0.663	0.142
2, 0, 0	0.238	0.614	0.147
2, 0, 1	0.225	0.624	0.150
2, 1, 0	0.216	0.641	0.142
2, 1, 1	0.204	0.651	0.144
3, 0, 0	0.250	0.601	0.148
3, 0, 1	0.236	0.611	0.151
3, 1, 0	0.227	0.628	0.144
3, 1, 1	0.214	0.638	0.146
4, 0, 0	0.262	0.587	0.150
4, 0, 1	0.248	0.598	0.153
4, 1, 0	0.238	0.615	0.145
4, 1, 1	0.225	0.625	0.148

h) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?
Ans: Maximum Odd Value = 3.42244,
Values of group_size, homeowner, married_couple are 1, 1, 1 respectively