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. =

I have 20 Degree of Freedom.

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. =

```
Deviance Chi=Square Test
==>for (Intercept + group_size) model
Chi-Square Statistic =  987.5766005262267
  Degrees of Freedom =  6
        Significance =  4.347870389027117e-210
==>for (Intercept + group_size + homeowner) model
Chi-Square Statistic =  5867.781500353245
  Degrees of Freedom =  2
        Significance =  0.0
==>for (Intercept + group_size + homeowner + married_couple) model
Chi-Square Statistic =  84.5780023841653
  Degrees of Freedom =  2
        Significance =  4.306457217534288e-19
==>for (Intercept + group_size + homeowner + married_couple + group_
size * homeowner) model
Chi-Square Statistic =  254.0781253632158
  Degrees of Freedom =  6
        Significance =  5.512105969198056e-52
==>for (Intercept + group_size + homeowner + married_couple + group_
size * homeowner + homeowner * married_couple) model
Chi-Square Statistic =  70.84227677015588
  Degrees of Freedom =  2
        Significance =  4.13804354648637e-16
```

	Test Statistic	DF	Significance	Association Measure	
group_size	Deviance	329.43	2	2.91862e-72	McFaddens R^2	0.000276643
homeowner	Deviance	6264.24	2	0	McFaddens R^2	0.00526047
married_couple	Deviance	714.265	2	7.93e-156	McFaddens R^2	0.000599813

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. =

```
Feature Importance Index for (Intercept + group_size)
= 209.36172341080683
Feature Importance Index for (Intercept + group_size + homeowner)
= inf
Feature Importance Index for (Intercept + group_size + homeowner + m
arried_couple)
= 18.36587986292153
Feature Importance Index for (Intercept + group_size + homeowner + m
arried_couple + group_size * homeowner)
= 51.25868244179064
Feature Importance Index for (Intercept + group_size + homeowner + m
arried_couple + group_size * homeowner + homeowner * married_couple)
= 15.38320494337081
```

e) (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 multinomial logistic model. List your answers in a table with proper labelling.

Ans. =

	group_size	homeowner	married_couple	p_a_0	p_a_1	p_a_2
0	1	0	0	0.259651	0.589175	0.151174
1	1	0	1	0.260092	0.592106	0.147802
2	1	1	0	0.183602	0.682030	0.134368
3	1	1	1	0.154023	0.709918	0.136059
4	2	0	0	0.221936	0.621105	0.156959
5	2	0	1	0.222321	0.624216	0.153463
6	2	1	0	0.202510	0.659773	0.137718
7	2	1	1	0.170552	0.689450	0.139999
8	3	0	0	0.239570	0.604616	0.155814
9	3	0	1	0.239992	0.607660	0.152348
10	3	1	0	0.301140	0.531297	0.167563
11	3	1	1	0.259017	0.567017	0.173966
12	4	0	0	0.194485	0.669686	0.135829
13	4	0	1	0.194692	0.672592	0.132716
14	4	1	0	0.387719	0.484974	0.127306
15	4	1	1	0.339172	0.526404	0.134424

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. =

```
group_size=1
homeowner=1
married_couple=1

Maximum value = 4.609169
```

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. =

= loge((prob(A=2) / prob(A=0) | group_size=3)) - loge((prob(A=2) / prob(A=0) | group_size=1))
= parameter of (group_size = 3 | A=2) – parameter of (group_size = 1 | A=2)
= 0.527471 - 0.80149
= -0.274022
= exp (-0.274022)
= 0.76031534813

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. =

= log(prob(A=0) / prob(A=1) | homeowner = 1) – log(prob(A=0) / prob(A=1) | homeowner = 0)

= (0.800157 – 1.505554 * g1 – 1.164638 * g2 – 0.654639 * g3 + 0.212483 *(1-m)

= (Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0)

= exp ((0.800157 – 1.505554 * g1 – 1.164638 * g2 – 0.654639 * g3 + 0.212483 *(1-m))

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. =

```
    count  class probability
A
0  143691           0.215996
1  426067           0.640462
2   95491           0.143542
```

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

Ans. =

```
Frequency Table:
 group_size       1      2     3    4
A
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. =

```
Frequency Table:
 homeowner       0       1
A
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. =

```
Frequency Table:
 married_couple       0      1
A
0               117110  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. =

```
homeowner       CramerV  0.0970864
married_couple  CramerV  0.0324216
group_size      CramerV   0.027102
```

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. =

Prob(A = a, group_size = g, homeowner = h, married_couple = m)
=
Prob(A = a | group_size = g, homeowner = h, married_couple = m)
*
Prob(group_size = g, homeowner = h, married_couple = m)
=
Prob(A =a)
*
Prob(group_size = g, homeowner = h, married_couple = m | A = a)
=
Prob(A =a)
*
Prob(group_size = g | A = a) * Prob(homeowner = h | A = a) * Prob(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.

Ans. =

A=0	A=1	A=2
0.26972190083648967	0.5801333993691891	0.15014469979432118
0.23278921851630957	0.6142185578024016	0.15299222368128876
0.19403790475559898	0.6696590048821739	0.1363030903622272
0.164935004743777,	0.6982780459509148	0.13678694930530805
0.2311433273249531	0.6165184597447714	0.15233821293027552
0.198015591405003	0.6479067807659843	0.15407762782901277
0.16362752552123652	0.7002878088359464	0.13608466564281702
0.13827417044457968	0.7259549630220522	0.13577086653336812
0.30821939378427693	0.5159241677311622	0.17585643848456095
0.26831105711605896	0.5509508971155715	0.18073804576836952
0.22697183146374494	0.6096117811433283	0.16341638739292683
0.19436951362831584	0.6404097735081213	0.16522071286356266
0.3754903907259939	0.4878101005336526	0.13669950874035344
0.3307434441365481	0.527098304946624	0.14215825091682782
0.2821726796029393	0.5881964548622688	0.1296308655347919
0.24393033920041854	0.6237659642682374	0.13230369653134402

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:
group_size=2,
homeowner=1,
married_couple=1
The maximum value is: 5.25011258