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.
    ⇨ 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?
    ⇨ 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
⇨

model	Deviance Chi-Square Test	degrees of freedom	significance
group_size	987.5766005262267	6	4.347870389027117e-210
group_size + homeowner	5867.781500353245	2	0.0
group_size + homeowner + married_couple	84.5780023841653	2	4.306457217534288e-19
group_size + homeowner + married_couple + group_size * homeowner	254.0781253632158	6	5.512105969198056e-52
group_size + homeowner + married_couple + group_size * homeowner + homeowner * married_couple	70.84227677015588	2	4.13804354648637e-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.
⇨

model	Feature Importance Index
group_size	209.36172341080683
group_size + homeowner	0
group_size + homeowner + married_couple	18.36587986292153
group_size + homeowner + married_couple + group_size * homeowner	51.25868244179064
group_size + homeowner + married_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.

⇨

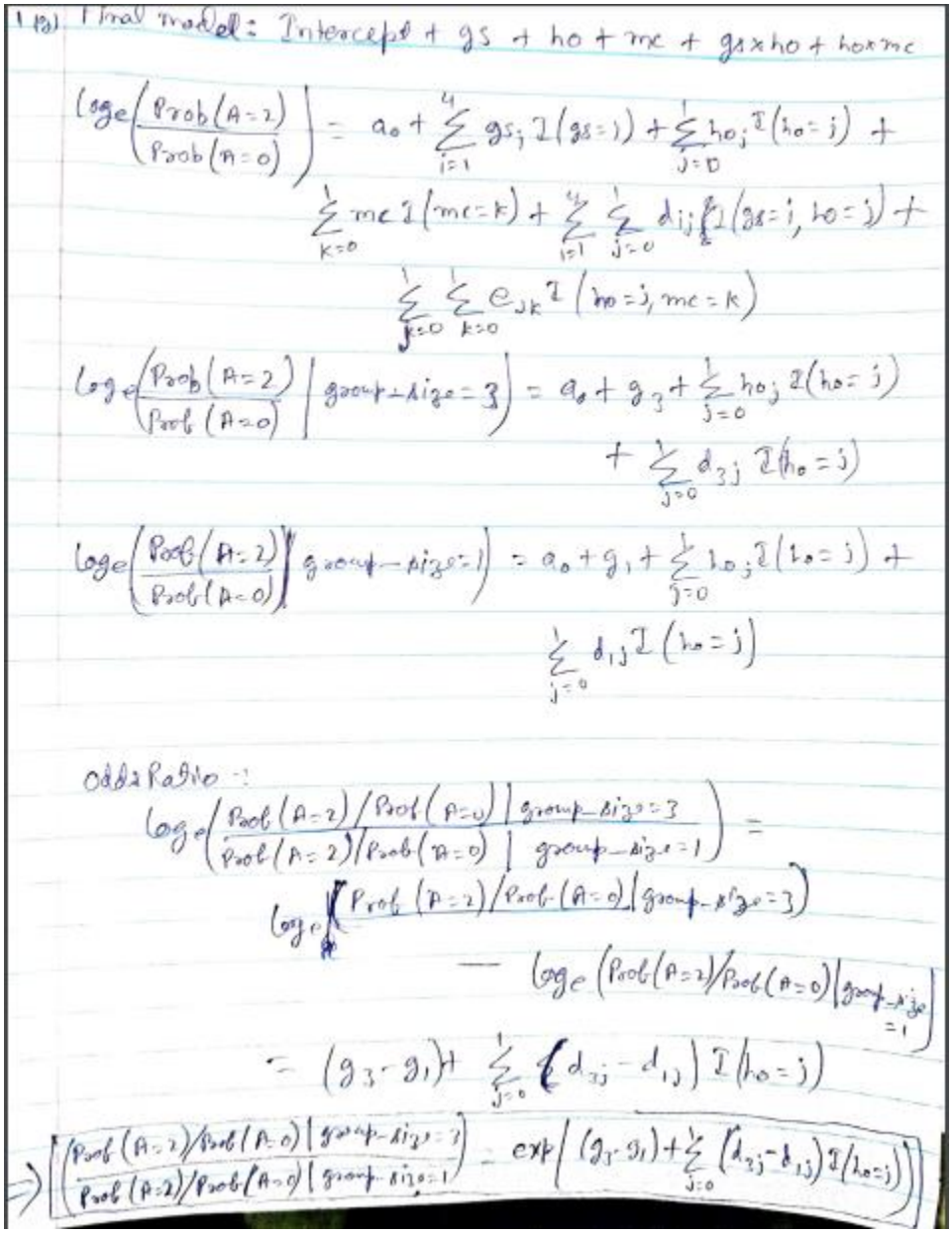
group_size	homeowner	married_couple	PA0	PA1	PA2
1	0	0	0.259651	0.589175	0.151174
1	0	1	0.260092	0.592106	0.147802
1	1	0	0.183602	0.68203	0.134368
1	1	1	0.154023	0.709918	0.136059
2	0	0	0.221936	0.621105	0.156959
2	0	1	0.222321	0.624216	0.153463
2	1	0	0.20251	0.659773	0.137718
2	1	1	0.170552	0.68945	0.139999
3	0	0	0.23957	0.604616	0.155814
3	0	1	0.239992	0.60766	0.152348
3	1	0	0.30114	0.531297	0.167563
3	1	1	0.259017	0.567017	0.173966
4	0	0	0.194485	0.669686	0.135829
4	0	1	0.194692	0.672592	0.132716
4	1	0	0.387719	0.484974	0.127306
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?

⇨ From above mentioned table in the answer to question 1 e)

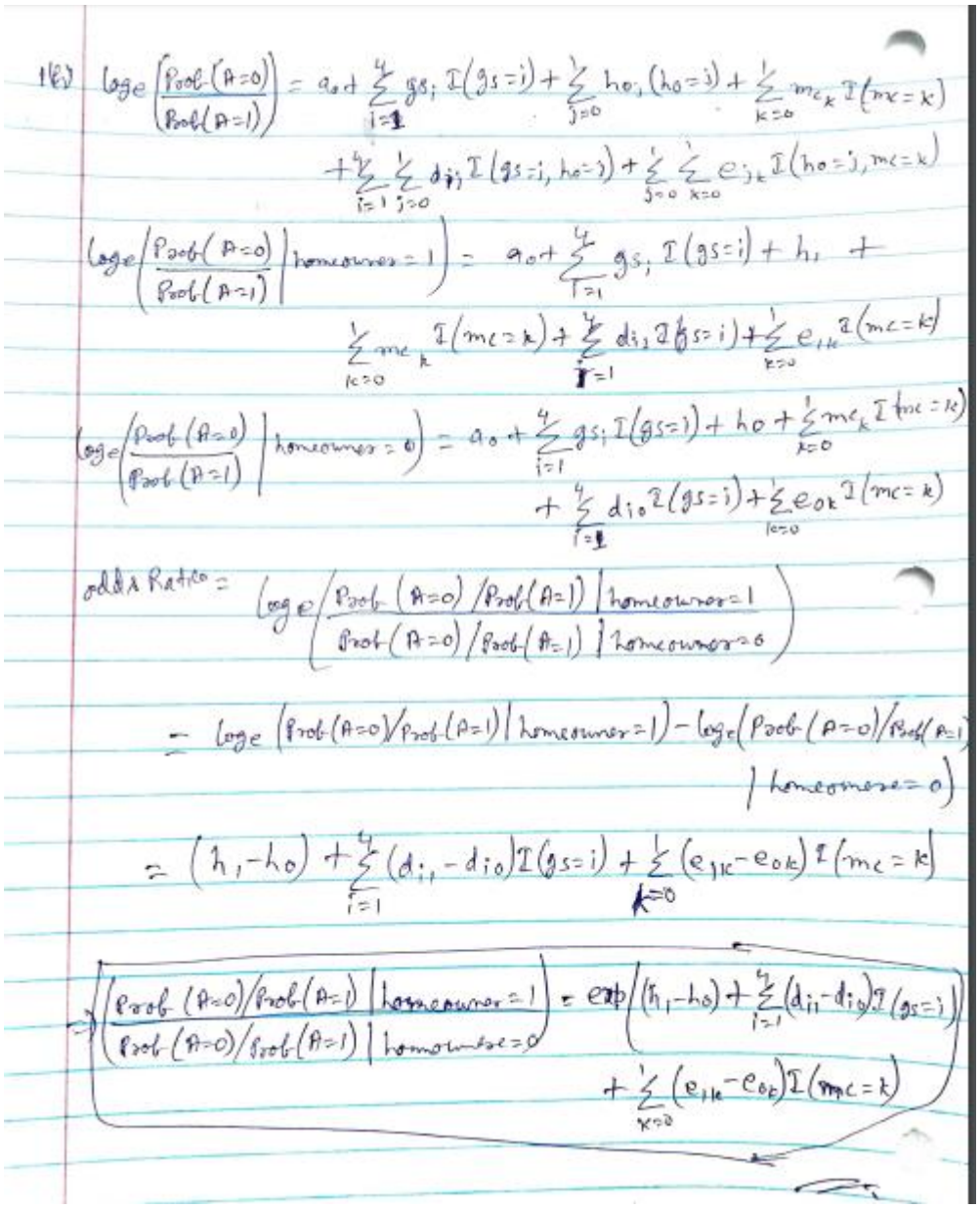
group_size = 1
homeowner = 1
married_couple= 1
PA0 = 0.154023
PA1 = 0.709918
PA2 = 0.136059
maximum odd value(PA1/PA0) = 4.609169

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).



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).

⇨



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.
⇨

Target Varable	Count	Class_probability
0	143691	0.215995815
1	426067	0.640462443
2	95491	0.143541742

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

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

Row Table:

A	1	2	3	4
0	0.8035297965774	0.1790508800134	0.0158813008470	0.0015380225623
1	0.7734745943713	0.2137339901940	0.0118971898786	0.0008942255561
2	0.7780104931355	0.2052549454922	0.0157606476003	0.0009739137720

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

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

Row Table:

A	0	1
0	0.5474177227523	0.4525822772477

1	0.4298150290917	0.5701849709083
2	0.4894073787058	0.5105926212942

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

A	0	1
0	117110	26581
1	333272	92795
2	75310	20181

Row Table:

A	0	1
0	0.8150127704588	0.1849872295412
1	0.7822056155487	0.2177943844513
2	0.7886607114807	0.2113392885193

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?

	Test	Statistic	DF	Significance	Association	Measure
group_size	Chi-square	977.276	6	7.34301e-208	CramerV	0.027102
married_couple	Chi-square	699.285	2	1.41953e-152	CramerV	0.0324216
homeowner	Chi-square	6270.49	2	0	CramerV	0.0970864

From the above table we can say that 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.

$$\text{Prob}(A=a, \text{group\_size}=g, \text{homeowner}=h, \text{married\_couple}=m)$$

$$= \text{Prob}(A=a \mid \text{group\_size}=g, \text{homeowner}=h, \text{married\_couple}=m) \times \text{Prob}(\text{group\_size}=g, \text{homeowner}=h, \text{married\_couple}=m)$$

$$= \text{Prob}(A=a) \times \text{Prob}(\text{group\_size}=g, \text{homeowner}=h, \text{married\_couple}=m \mid A=a)$$

$$= \text{Prob}(A=a) \times \text{Prob}(\text{group\_size}=g \mid A=a) \times \text{Prob}(\text{homeowner}=h \mid A=a) \times \text{Prob}(\text{married\_couple}=m \mid A=a)$$

Therefore,

$$\text{Prob}(A=a, \text{group\_size}=g, \text{homeowner}=h, \text{married\_couple}=m) = \text{Prob}(A=a) \times \text{Prob}(\text{group\_size}=g \mid A=a) \times \text{Prob}(\text{homeowner}=h \mid A=a) \times \text{Prob}(\text{married\_couple}=m \mid A=a)$$

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.

⇨

group_size	homeowner	married_couple	PA0	PA1	PA2
1	0	0	0.227037	0.627593	0.14537
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	0.14572
4	1	1	0.225656	0.62572	0.148624

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?

⇨ From the above mentioned table in the answer to question 2 g)

group_size = 1
homeowner = 1
married_couple= 1
PA0 = 0.193842
PA1 = 0.663414
PA2 = 0.142744
maximum odd value(PA1/PA0) = 3.422447148