

CS 584-04: Machine Learning

Fall 2018 Assignment 3

Suhas Sreenivas - A20423132

Question 1

- a) (5 points). What is the Gini metric for the root node?

The Gini metric for the root node is 0.7670692.

- b) (5 points). How many possible binary-splits that you can generate from the CarOwnership predictor?

$$2^{k-1} - 1$$

$$= 2^{3-1} - 1 = 3$$

- c) (10 points). Calculate the Gini metric for each possibly binary split that you can generate from the CarOwnership predictor. List your answers in a table. The table should have three columns: the sequence index of the split, the contents of the two branches, the split Gini metric.

| Contents of the 2 branches | Sequence index of the split | Split GINI metric |
|----------------------------|-----------------------------|--------------------|
| (Lease, none) and own | 0,1 and 2 | 0.7660055295576353 |
| (None, own) and Lease | 1,2 and 0 | 0.7657091011467262 |
| (Own, Lease) and None | 0,2 and 1 | 0.7667762687160046 |

- d) (5 points). What is the optimal split for the CarOwnership predictor?

(Own, None) and Lease with a GINI index of 0.7657091011467262

- e) (5 points). How many possible binary-splits that you can generate from the JobCategory predictor?

$$2^{7-1} - 1 = 64 - 1 = 63$$

- f) (10 points). Calculate the Gini metric for each possibly binary split that you can generate from the JobCategory predictor. List your answers in a table. The table should have three columns: the sequence index of the split, the contents of the two branches, the split Gini metric.

| | Contents of the 2 branches | Sequence index | Split GINI metric |
|----|--|---------------------------|--------------------|
| 1 | Agriculture and (craft,labor,missing,professional,sales,service) | [(0), (1,2,3,4,5,6)] | 0.7670516826657839 |
| 2 | craft and (Agriculture,labor,missing,professional,sales,service) | [(1), (0, 2, 3, 4, 5, 6)] | 0.7669942084821879 |
| 3 | labor and (craft, Agriculture,missing,professional,sales,service) | [(2), (0, 1, 3, 4, 5, 6)] | 0.7670315160755341 |
| 4 | missing and (craft,labor, Agriculture,professional,sales,service) | [(3), (0, 1, 2, 4, 5, 6)] | 0.7668979204279504 |
| 5 | professional and (craft,labor,missing, Agriculture,sales,service) | [(4), (0, 1, 2, 3, 5, 6)] | 0.7667678180799103 |
| 6 | sales and (craft,labor,missing,professional, Agriculture,service) | [(5), (0, 1, 2, 3, 4, 6)] | 0.7666854311563086 |
| 7 | service and (craft,labor,missing,professional,sales, Agriculture) | [(6), (0, 1, 2, 3, 4, 5)] | 0.7670150066283693 |
| 8 | Agriculture, Crafts and (labor,missing,professional,sales,service) | [(0, 1), (2, 3, 4, 5, 6)] | 0.7670091884141732 |
| 9 | Agriculture Labor and (craft,missing,professional,sales,service) | [(0, 2), (1, 3, 4, 5, 6)] | 0.7670390199142361 |
| 10 | Agriculture, Missing and (craft,labor,professional,sales,service) | [(0, 3), (1, 2, 4, 5, 6)] | 0.7670229036125564 |
| 11 | Agriculture, Professional and (craft,labor,missing,sales,service) | [(0, 4), (1, 2, 3, 5, 6)] | 0.7667752040720033 |
| 12 | Agriculture, Sales and (craft,labor,missing,professional,service) | [(0, 5), (1, 2, 3, 4, 6)] | 0.7667187854366833 |
| 13 | Agriculture, Service and (craft,labor,missing,professional,sales) | [(0, 6), (1, 2, 3, 4, 5)] | 0.7670277019046212 |
| 14 | craft Labor and (Agriculture,missing,professional,sales,service) | [(1, 2), (0, 3, 4, 5, 6)] | 0.7669835211091989 |
| 15 | craft, Missing and (Agriculture,labor,professional,sales,service) | [(1, 3), (0, 2, 4, 5, 6)] | 0.7669933691118804 |
| 16 | craft, Professional and (Agriculture,labor,missing,sales,service) | [(1, 4), (0, 2, 3, 5, 6)] | 0.7667013453310396 |

| | | | |
|----|--|---------------------------|------------------------|
| 17 | craft, Sales and (Agriculture,labor,missing,professional,serv ice) | [(1, 5), (0, 2, 3, 4, 6)] | 0.8059158490275 |
| 18 | craft, Service and (Agriculture,labor,missing,professional,sale s) | [(1, 6), (0, 2, 3, 4, 5)] | 0.7670440336504 847 |
| 19 | Labor, Missing and (Agriculture, Crafts,professional,sales,service) | [(2, 3), (0, 1, 4, 5, 6)] | 0.7670487631566 11 |
| 20 | Labor, Professional and (Agriculture, Crafts,missing,sales,service) | [(2, 4), (0, 1, 3, 5, 6)] | 0.7668199919361 81 |
| 21 | Labor, Sales and (Agriculture, Crafts,missing,professional,service) | [(2, 5), (0, 1, 3, 4, 6)] | 0.7667483725125 881 |
| 22 | Labor, Service and (Agriculture, Crafts,missing,professional,sales) | [(2, 6), (0, 1, 3, 4, 5)] | 0.7669986871637 38 |
| 23 | Professional, Missing and (Agriculture, Crafts,Labor,sales,service) | [(3, 4), (0, 1, 2, 5, 6)] | 0.7667445819476 135 |
| 24 | Missing, Sales and (Agriculture, Crafts,Professional,sales,service) | [(3, 5), (0, 1, 2, 4, 6)] | 0.7666929425599 277 |
| 25 | Missing, Service and (Agriculture, Crafts,Labor,professional,service) | [(3, 6), (0, 1, 2, 4, 5)] | 0.7670279564538 968 |
| 26 | Professional, Sales and (Agriculture, Crafts, Labor, Missing, Service) | [(4, 5), (0, 1, 2, 3, 6)] | 0.7670179293290 057 |
| 27 | Professional, Service and (Agriculture, Crafts, Labor, Missing, Sales) | [(4, 6), (0, 1, 2, 3, 5)] | 0.7668363333333 333 |
| 28 | Sales , Services and (Agriculture, Crafts, Labor, Missing, Professional) | [(5, 6), (0, 1, 2, 3, 4)] | 0.7666814182609 785 |
| 29 | [(Agriculture, Craft, Labor), (Missing, Professional, Sales, Services)] | [(0, 1, 2), (3, 4, 5, 6)] | 0.7669926818873 669 |
| 30 | [(Agriculture, Craft, Missing), (Labor, Professional, Sales, Services)] | [(0, 1, 3), (2, 4, 5, 6)] | 0.7670014692093 515 |
| 31 | [(Agriculture, Craft, Professional), (Labor, Missing, Sales, Services)] | [(0, 1, 4), (2, 3, 5, 6)] | 0.7667009008842 012 |
| 32 | [(Agriculture, Craft, Sales), (Labor, Missing, Professional, Services)] | [(0, 1, 5), (2, 3, 4, 6)] | 0.7668272791206 003 |
| 33 | [(Agriculture, Craft, Services), (Labor, Missing, Professional, Sales)] | [(0, 1, 6), (2, 3, 4, 5)] | 0.7670474978957 71 |
| 34 | [(Agriculture, Labor, Missing), (Craft, Professional, Sales, Services)] | [(0, 2, 3), (1, 4, 5, 6)] | 0.7670475434223 493 |
| 35 | [(Agriculture, Labor, Professional), (Craft, Missing, Sales, Services)] | [(0, 2, 4), (1, 3, 5, 6)] | 0.7668140712782 201 |

| | | | |
|----|---|---------------------------|--------------------|
| 36 | [(Agriculture, Labor, Sales), (Craft, Missing, Professional, Services)] | [(0, 2, 5), (1, 3, 4, 6)] | 0.7667641833913141 |
| 37 | [(Agriculture, Labor, Services), (Craft, Missing, Professional, Sales)] | [(0, 2, 6), (1, 3, 4, 5)] | 0.7670088886265447 |
| 38 | [(Agriculture, Missing, Professional), (Craft, Labor, Sales, Services)] | [(0, 3, 4), (1, 2, 5, 6)] | 0.7667497953159906 |
| 39 | [(Agriculture, Missing, Sales), (Craft, Labor, Professional, Services)] | [(0, 3, 5), (1, 2, 4, 6)] | 0.7667220133097015 |
| 40 | [(Agriculture, Missing, Services), (Craft, Labor, Professional, Sales)] | [(0, 3, 6), (1, 2, 4, 5)] | 0.7670320641598185 |
| 41 | [(Agriculture, Professional, Sales), (Craft, Labor, Missing, Services)] | [(0, 4, 5), (1, 2, 3, 6)] | 0.7670153981729524 |
| 42 | [(Agriculture, Professional, Services), (Craft, Labor, Missing, Sales)] | [(0, 4, 6), (1, 2, 3, 5)] | 0.766831540495173 |
| 43 | [(Agriculture, Sales, Services), (Craft, Labor, Missing, Professional)] | [(0, 5, 6), (1, 2, 3, 4)] | 0.76669976314273 |
| 44 | [(Craft, Labor, Missing), (Agriculture, Professional, Sales, Services)] | [(1, 2, 3), (0, 4, 5, 6)] | 0.7669970153695415 |
| 45 | [(Craft, Labor, Professional), (Agriculture, Missing, Sales, Services)] | [(1, 2, 4), (0, 3, 5, 6)] | 0.7667090459089398 |
| 46 | [(Craft, Labor, Sales), (Agriculture, Missing, Professional, Services)] | [(1, 2, 5), (0, 3, 4, 6)] | 0.7668155372116556 |
| 47 | [(Craft, Labor, Services), (Agriculture, Missing, Professional, Sales)] | [(1, 2, 6), (0, 3, 4, 5)] | 0.7669972882930872 |
| 48 | [(Craft, Missing, Professional), (Agriculture, Labor, Sales, Services)] | [(1, 3, 4), (0, 2, 5, 6)] | 0.7666819099761641 |
| 49 | [(Craft, Missing, Sales), (Agriculture, Labor, Professional, Services)] | [(1, 3, 5), (0, 2, 4, 6)] | 0.7668173702917922 |
| 50 | [(Craft, Missing, Services), (Agriculture, Labor, Professional, Sales)] | [(1, 3, 6), (0, 2, 4, 5)] | 0.7670535229559383 |
| 51 | [(Craft, Professional, Sales), (Agriculture, Labor, Missing, Services)] | [(1, 4, 5), (0, 2, 3, 6)] | 0.767023523216099 |
| 52 | [(Craft, Professional, Services), (Agriculture, Labor, Missing, Sales)] | [(1, 4, 6), (0, 2, 3, 5)] | 0.7667760989651224 |
| 53 | [(Craft, Sales, Services), (Agriculture, Labor, Missing, Professional)] | [(1, 5, 6), (0, 2, 3, 4)] | 0.7668008442520571 |
| 54 | [(Labor, Missing, Professional), (Agriculture, Craft, Sales, Services)] | [(2, 3, 4), (0, 1, 5, 6)] | 0.7668099526557102 |

| | | | |
|----|---|---------------------------|--------------------|
| 55 | [(Labor, Missing, Sales), (Agriculture, Craft, Professional, Services)] | [(2, 3, 5), (0, 1, 4, 6)] | 0.7667638050721957 |
| 56 | [(Labor, Missing, Services), (Agriculture, Craft, Professional, Sales)] | [(2, 3, 6), (0, 1, 4, 5)] | 0.7670188726283493 |
| 57 | [(Labor, Professional, Sales), (Agriculture, Craft, Missing, Services)] | [(2, 4, 5), (0, 1, 3, 6)] | 0.7670517628446133 |
| 58 | [(Labor, Professional, Services), (Agriculture, Craft, Missing, Sales)] | [(2, 4, 6), (0, 1, 3, 5)] | 0.7668304838752236 |
| 59 | [(Labor, Sales, Services), (Agriculture, Craft, Missing, Professional)] | [(2, 5, 6), (0, 1, 3, 4)] | 0.7666785924047138 |
| 60 | [(Missing, Professional, Sales), (Agriculture, Craft, Labor, Services)] | [(3, 4, 5), (0, 1, 2, 6)] | 0.7670039804995727 |
| 61 | [(Missing, Professional, Services), (Agriculture, Craft, Labor, Sales)] | [(3, 4, 6), (0, 1, 2, 5)] | 0.7668233960538839 |
| 62 | [(Missing, Sales, Services), (Agriculture, Craft, Labor, Professional)] | [(3, 5, 6), (0, 1, 2, 4)] | 0.7666943383195367 |
| 63 | [(Professional, Sales, Services), (Agriculture, Craft, Labor, Missing)] | [(4, 5, 6), (0, 1, 2, 3)] | 0.7670006273964458 |

g) (5 points). What is the optimal split for the JobCategory predictor?

| Optimal Split | | GINI for split |
|---|---------------------------|--------------------|
| [(Labor, Sales, Services), (Agriculture, Craft, Missing, Professional)] | [(2, 5, 6), (0, 1, 3, 4)] | 0.7666785924047138 |

h) (5 points). Between the CarOwnership and the JobCategory predictors, which predictor will you choose for the second layer (i.e., depth 1) of your decision tree?

I will choose CarOwnership as the predictor for the second layer as the GINI for CarOwnership is lesser than the GINI for JobCategory.

Question 2

- a) (10 points) Suppose you start with a model with only the Intercept term (i.e., without any predictors). You are asked to mathematically calculate the maximum likelihood estimates of the predicted probabilities $\pi_{ij}, j = 1, 2, 3$ without calling the MNLogit function. Show all the necessary steps and the estimates for the $\pi_{ij}, j = 1, 2, 3$. (Hint: equate the first derivatives of the log-likelihood function to zeros for this Intercept-only model)

Equating the first derivative of the log-likelihood function to zero.

$$\sum_{i=1}^m x_{i,j} (n_{ij} - n_i \pi_{ij}) = 0$$

Since there are no subpopulations, $m=1$

$$x_{i,j} (n_{ij} - n_i \pi_{ij}) = 0$$

$$n_{ij} - n_i \pi_{ij} = 0$$

$$\pi_{ij} = \frac{n_{ij}}{n_i}$$

| No. of observations | A | j | total no. of observations = 665249 |
|---------------------|---|---|------------------------------------|
| 143691 (n_{i1}) | 0 | 1 | |
| 426067 (n_{i2}) | 1 | 2 | |
| 95491 (n_{i3}) | 2 | 3 | |

$$\pi_{ij} = \frac{n_{ij}}{n_{\text{total}}}$$

$$\pi_{i1} = \frac{143691}{665249} = 0.215996$$

$$\pi_{i2} = \frac{426067}{665249} = 0.640462$$

$$\pi_{i3} = \frac{95491}{665249} = 0.143542$$

$$\begin{aligned}
 \log \text{ likelihood} &= \sum_{i=1}^n \sum_{j=1}^K n_{ij} \log_e(x_{ij}) \\
 \Rightarrow l &= 143891 \times \log_e(0.215996) + \\
 &\quad 426067 \times \log_e(0.640462) + \\
 &\quad 95491 \times \log_e(0.143542) \\
 &= -595406.7619
 \end{aligned}$$

$$\log L = l$$

$$\log_e L = l$$

$$\log_e L = -595406.7619$$

$$L = e^{-595406.7619} = \text{Maximum Likelihood.}$$

- b) (10 points) Next, you are asked to mathematically calculate the maximum likelihood estimates of the Intercept terms $\beta_{j0}, j = 1, \dots, K$. The convention is to set the Intercept term to zero for the target category $A = 0$, i.e., $\beta_{10} = 0$. (Hint: use the mathematical formula of the logit of π_{ij} (i.e., $\log_e(\pi_{ij}/\pi_{i1})$) for this Intercept only model, then solve for the betas)?

$$b) \quad \log_e\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = x_i^T \beta_j$$

for $j=1$ $j=J$ as target category $A=0$ is the reference. $\times \beta_J = 0$

~~From~~ From question (a),
 $\pi_{ij} = \pi_{i1} = 0.215996$

For $j=2$

$$\beta_{j=2} = \log_e\left(\frac{\pi_{i2}}{\pi_{i1}}\right) = \frac{\pi_{i2}}{\pi_{i1}}$$

$$\beta_{j=2} = \log_e\left(\frac{0.640462}{0.215996}\right)$$

$$= \log_e(2.965156762)$$

$$= 1.08692$$

For $j=3$

$$\beta_{j=3} = \log_e\left(\frac{\pi_{i3}}{\pi_{i1}}\right) = \log_e\left(\frac{0.143542}{0.215996}\right)$$

$$= \log_e(0.6645586)$$

$$= -0.40863221$$

- c) (4 points) Now, you will use the MNLogit function to build the multinomial logistic model. What value of the target variable A is used by the MNLogit function as the reference category? The statsmodels.api.MNLogit function conventionally takes the lexically first target category as the reference. In this case, A = 0 is the reference category.

- d) (2 points) How many iterations are performed before convergence is achieved?
57 Iterations were performed before convergence.
With default settings, the algorithm did not converge. Hence the maximum number of iterations was set to 100.
`thisFit = logit.fit(method='newton', full_output = True, maxiter = 100, tol = 1e-8)`

- e) (4 points) How many parameters (including the redundant ones) are in the model?
There are 7 parameters.

- f) (5 points) When group_size = 2, homeowner = 1, and married_couple = 1, what are the predicted probabilities: Prob(A = 0), Prob(A = 1), and Prob(A = 2)?

`[[0.16697404 0.69415127 0.13887469]]`

Prob(A = 0) = 0.16697404

Prob(A = 1) = 0.69415127

Prob(A = 2) = 0.13887469

- g) (10 points) What are the values of the predictors group_size, homeowner, and married_couple such that Prob(A = 0) will attain its maximum? What is the maximum Prob(A = 0) value?

`[1,0,0,0,1,0,0]`

The values of the predictors are :

group_size = 4 homeowner = 0 married_couple = 0

Maximum Prob(A = 0) value = 0.42117104

- h) (5 points) According to the logistic model, what is the odds ratio for group_size = 4 versus group_size = 1, and A = 1 versus A = 0? Mathematically, the odds ratio is $(\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 4) / ((\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 1))$.

For group_size , 4 = $0.408236 - 0.359197 = 0.049039$

For group_size, 1 = $0.408236 + 0.424367 = 0.832603$

Group_size 4 – group_size 1 = -0.783564

$\log_e (\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 4) - \log_e (\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 1) = -0.783564$

$\log_e \{ (\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 4) / (\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 1) \} =$

-0.783564

$(\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 4) / (\text{Prob}(A=1)/\text{Prob}(A=0) \mid \text{group_size} = 1) =$
 $e^{-0.783564} = 0.45678$