Rahul Chamarthi
CS373
9/17/19

HW2

1. **NBC details (20 pts)**
    a. Write down the mathematical expression for P(Y|X) given by the NBC. (2pts)
        i. **P(y | x) =** $\frac{P(x|y)P(y)}{P(x)}$
            - **[prior] → P(x) = probability this particular sample of data is observed**
            - **[likelihood] → P(x | y) = probability of observing the sample x given the class label (y)**
            - **[evidence] → P(y) = prior probability of class label y. This reflects background knowledge before data is observed.**
            - **[posterior probability] → P(y | x) = posterior probability of y is the probability that y is the class label, given that some feature set (x) is observed.**
        ii. **BACKGROUND:**
            1. **class label: outdoorSeating**
            2. **feature set: all attributes up to, but not including, outdoorSeating. Ex. state, latitude, longitude, stars, attire, etc.**
    b. Suppose your data has binary class labels , i.e y∈ {0,1}, write the expression for predicting the class for a given input row. You will use this expression in your implementation. (1 pts)
        i. **P(Y | ALL_ATTRS) =** $\frac{P(ALL\_ATTRS \mid Y)\,P(Y)}{P(ALL\_ATTRS)}$ **=** $\frac{P(ATR_1 \mid Y)\,P(ATR_2 \mid Y).....P(ATR_{N-1} \mid Y)P(Y)}{P(ALL\_ATTRS) = P(ATR_1, ATR_2, \ldots ATR_{N-1})}$ **=**
        $P(ATR_1 \mid Y)P(ATR_2 \mid Y)...P(ATR_{N-1} \mid Y)P(Y)$
        ii. **BACKGROUND:**
            1. **Substitute Y for both 0 and 1 and select the probability that maximizes. The value (0,1) belonging to the selected probability will be our prediction.**
            2. **Most other explanations can be derived from the breakdown of the mathematical expression in a**

3. **ALL_ATTRS = (ATR_1, ATR_2 …. ATR_{N-1}) is simply saying from the first feature for a data vector to the last feature in this data vector. Excluding the last (outdoorSeating) which is our class label.**
    a. **Example: For our data set this could be considered ATR_1 = state and ATR_{N-1} = goodForKids.**

c. State the naive assumption that lets us simplify the expression P(X|Y)P(Y).What rule(s) of probability are used to simplify the expression? Is this assumption true for the given Yelp data? Explain why or why not? (3 pts)
    i. **The naive assumption that allows to simplify the expression is assuming the attributes are conditionally independent provide the class.**
    ii. **In terms of which rule we use to simplify; we use Bayes Rule to simplify the expression.**
    iii. **I would say that this feature set could be assumed to be conditionally independent.**
        1. **Example for assumption of conditional independence:**
            a. **Suppose we are given the value of one class label(outdoorSeating) and we are given the fact that the ambience is romantic. It is not possible based on this information to tell what someone's personal dietary restrictions are with absolute certainty. The aforementioned example when scaled to all other columns is what allows me to derive my assumption.**

d. What part of the expression corresponds to the class prior? Considering the entire Yelp data as the training dataset, calculate the maximum likelihood es-timate for the class prior with and without smoothing. What is the effect of smoothing on the final probabilities? (4 pts)
    i. **The class prior corresponds to P(y) in the equation in part a.**
    ii. **MLE(class prior) → smooth vs unsmoothed:**
        1. **Smoothed**
            a. **P(YES) = 0.650421**
            b. **P(NO) = 0.349579**
        2. **Unsmoothed**
            a. **P(YES) = 0.650433**
            b. **P(NO) = 0.349567**

iii. **The effect of smoothing on the data is a small increase in P(NO) and a small decrease in P(YES). This makes sense considering that we knew almost all of the attribute feature values. The whole point of smoothing the data is to try and minimize the influence of unforseen feature attribute values.**

e. Specify the full set of parameters that need to be estimated for the NBC model of the Yelp data. How many parameters are there? (2 pts)

i. **REASONING: We were given the basic NBC Formula of: P(BC | A, I, S, C, R) ~ P(A|BC)P(I|BC)P(S|BC)P(R|BC)P(BC) in the slides. This equates to multiplying all of the CPDs and prior values. There is one CPD for every attribute. These CPDs model their individual relationships with the class label. These probabilities can be generated by applying the MLE method.**

ii. **ANSWER: There are a grand total of 42 parameters for our method. It will be all of the below values multiplied by the class prior. In our case the class prior(P(Y)) will be reflective of the feature outdoorSeating which is what we are trying to predict.**

iii. **TOTAL LIST OF PARAMETERS:**

1. **P(outdoorSeating) → class prior**
2. **P(state | outdoorSeating)**
3. **P(latitude | outdoorSeating)**
4. **P(longitude | outdoorSeating)**
5. **P(stars | outdoorSeating)**
6. **P(open | outdoorSeating)**
7. **P(alcohol | outdoorSeating)**
8. **P(noiseLevel | outdoorSeating)**
9. **P(attire | outdoorSeating)**
10. **P(priceRange | outdoorSeating)**
11. **P(delivery | outdoorSeating)**
12. **P(waiterService | outdoorSeating)**
13. **P(smoking | outdoorSeating)**
14. **P(caters, outdoorSeating)**
15. **P(goodForGroups | outdoorSeating)**
16. **P(goodForKids | outdoorSeating)**
17. **P(amb_romantic | outdoorSeating)**
18. **P(amb_intimate | outdoorSeating)**
19. **P(amb_touristy | outdoorSeating)**
20. **P(amb_trendy | outdoorSeating)**

21. **P(amb_classy | outdoorSeating)**
22. **P(amb_casual | outdoorSeating)**
23. **P(amb_divey | outdoorSeating)**
24. **P(amb_hipster | outdoorSeating)**
25. **P(amb_upscale | outdoorSeating)**
26. **P(p_garage | outdoorSeating)**
27. **P(p_valet | outdoorSeating)**
28. **P(p_validate | outdoorSeating)**
29. **P(p_street | outdoorSeating)**
30. **P(p_lot | outdoorSeating)**
31. **P(diet_halal | outdoorSeating)**
32. **P(diet_gluten | outdoorSeating)**
33. **P(diet_dairy | outdoorSeating)**
34. **P(diet_kosher | outdoorSeating)**
35. **P(diet_soy | outdoorSeating)**
36. **P(diet_vegetarian | outdoorSeating)**
37. **P(diet_vegan | outdoorSeating)**
38. **P(rec_breakfast | outdoorSeating)**
39. **P(rec_brunch, outdoorSeating)**
40. **P(rec_dinner | outdoorSeating)**
41. **P(rec_lunch | outdoorSeating)**
42. **P(rec_latenight | outdoorSeating)**
43. **P(rec_dessert | outdoorSeating)**

f. Write an expression for an arbitrary conditional probability distribution (CPD) of a discrete attribute X_i with k distinct values (conditioned on a binary class Y). Include a mathematical expression for the maximum likelihood estimates of the parameters of this distribution (with smoothing), which correspond to counts of attribute value combinations in a data set D. (2 pts)

i. **BACKGROUND:**
   1. **Y = y means for some class label value y**

ii. **ANSWER:**
   1. **$P(X_i = k \mid Y) = \frac{P(X_i=k) \cap P(Y=y)}{P(Y=y)}$**

g. For the Yelp data, explicitly state the mathematical expression for the maximum likelihood estimates (with smoothing) of the CPD parameters for the attribute `priceRange` conditioned on the the class label `outdoorSeating`. (2 pts)

i. **BACKGROUND:**
   1. **pR = priceRange**

2. **Y = y means for some class label value y**

ii. **ANSWER:**

1. **P(pR = k | Y=y) =** $$\frac{(\sum_{i=1}^{n}(pR=k) * (Y=y)) + 1}{(\sum_{i=1}^{n}(Y=y)) + \Sigma k \in k}$$

h. Consider the entire Yelp data as the training dataset and `outdoorSeating` as the class label. Estimate the conditional probability distributions of the following attributes with and without smoothing: What is the effect of smoothing (e.g., any difference compared to Q1d)? Which attribute shows the most association with the class? (4 pts)

i. **The effect of smoothing stayed the same as in Q1d. when we compared the smoothed CPDs to the unsmoothed CPDs there was a very small adjustment in the probabilities given outdoorSeating was true or given outdoorSeating was false. This could follow our explanation in Q1d. We must consider that we knew almost all of the attribute feature values for the entire set from the beginning. Meaning there was almost no feature attribute value left unaccounted. Seeing as the whole point of smoothing the data is to try and minimize the influence of unforseen feature attribute values, we can see why the difference would be negligible.**

ii. **The attribute that showed the most association with the class (outdoorSeating) was the feature Smoking with the attribute outdoor. We can infer this quite logically. If there is outdoorSeating people are more likely to smoke outdoors. Similarly, it would make sense that the probability of smoking outdoors would reduce if there is no outdoorSeating. Why would you want to leave your meal to go smoke outdoors. The probability for exact comparison went from 0.232 for outdoorSeating being false, to 0.671 if outdoorSeating was true.**

2. **Implement a naive Bayes classification algorithm in python. (20 pts)**

a. **This is done in our code file that we submitted via turnin so I am not writing any information here.**

3. **Evaluate the NBC using cross validation and learning curves. (10 pts)**

a. Record the mean zero-one loss observed across the ten trials for each training setsize (i.e., sample %). Record the mean squared loss across the ten trials for each training set size. (8 pts)

Zero = Zero Square Loss

Square = MSE

TRIAL #1 (1%)

1. **ZERO-ONE LOSS=0.344376**
   SQUARED LOSS=0.000307
2. **ZERO-ONE LOSS=0.322291**
   SQUARED LOSS=0.000209
3. **ZERO-ONE LOSS=0.332366**
   SQUARED LOSS=0.000167
4. **ZERO-ONE LOSS=0.334623**
   SQUARED LOSS=0.000296
5. **ZERO-ONE LOSS=0.311409**
   SQUARED LOSS=0.000138
6. **ZERO-ONE LOSS=0.309354**
   SQUARED LOSS=0.000227
7. **ZERO-ONE LOSS=0.332084**
   SQUARED LOSS=0.000247
8. **ZERO-ONE LOSS=0.325112**
   SQUARED LOSS=0.000307
9. **ZERO-ONE LOSS=0.319228**
   SQUARED LOSS=0.000155
10. **ZERO-ONE LOSS=0.319378**
    SQUARED LOSS=0.000175

TRIAL #2 (10%)

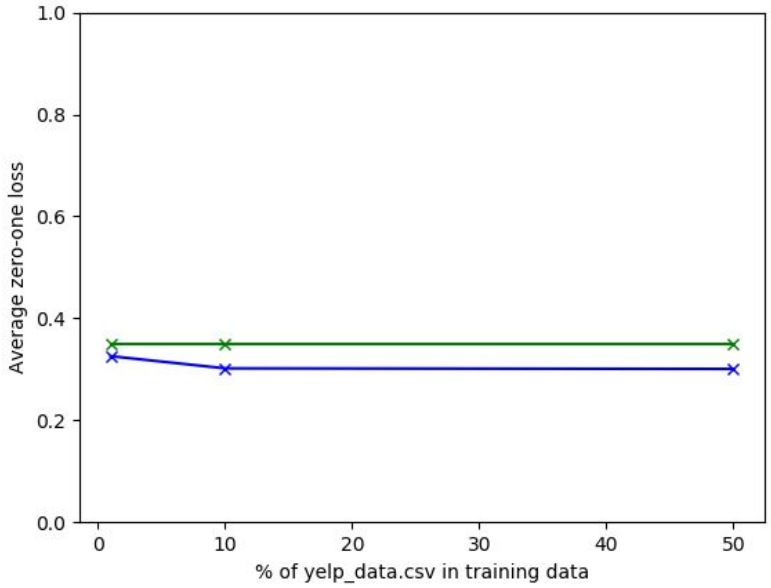
1. **ZERO-ONE LOSS=0.309354**
   SQUARED LOSS=0.000205
2. **ZERO-ONE LOSS=0.299440**
   SQUARED LOSS=0.000168
3. **ZERO-ONE LOSS=0.300488**
   SQUARED LOSS=0.000151
4. **ZERO-ONE LOSS=0.298352**
   SQUARED LOSS=0.000153
5. **ZERO-ONE LOSS=0.307258**
   SQUARED LOSS=0.000177
6. **ZERO-ONE LOSS=0.305566**
   SQUARED LOSS=0.000161
7. **ZERO-ONE LOSS=0.295934**
   SQUARED LOSS=0.000159

8. ZERO-ONE LOSS=0.300165
   SQUARED LOSS=0.000169
9. ZERO-ONE LOSS=0.300004
   SQUARED LOSS=0.000157
10. ZERO-ONE LOSS=0.300304
    SQUARED LOSS=0.000158

TRIAL #3 (50%)

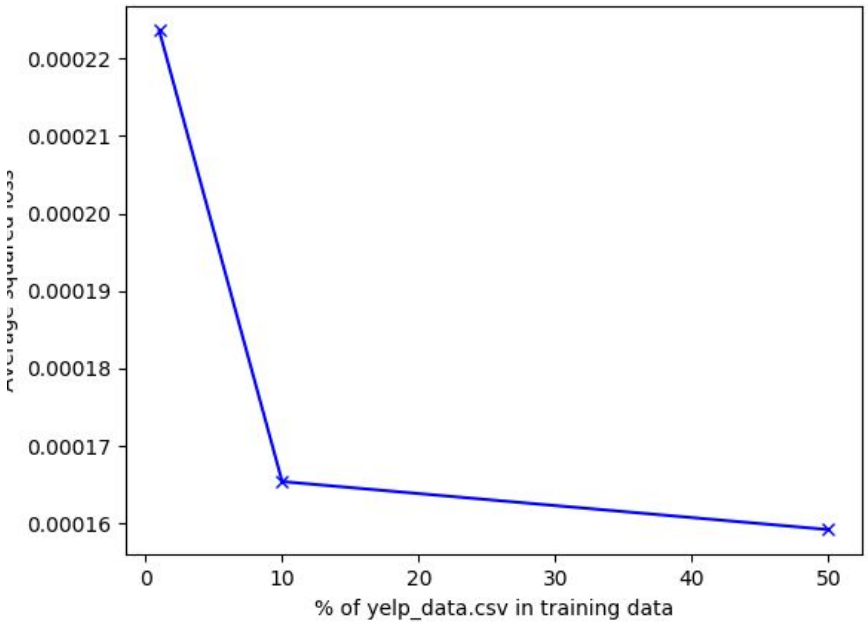
1. ZERO-ONE LOSS=0.300085
   SQUARED LOSS=0.000167
2. ZERO-ONE LOSS=0.303067
   SQUARED LOSS=0.000175
3. ZERO-ONE LOSS=0.299803
   SQUARED LOSS=0.000170
4. ZERO-ONE LOSS=0.301898
   SQUARED LOSS=0.000170
5. ZERO-ONE LOSS=0.298231
   SQUARED LOSS=0.000174
6. ZERO-ONE LOSS=0.298835
   SQUARED LOSS=0.000170
7. ZERO-ONE LOSS=0.299279
   SQUARED LOSS=0.000168
8. ZERO-ONE LOSS=0.300729
   SQUARED LOSS=0.000169
9. ZERO-ONE LOSS=0.300649
   SQUARED LOSS=0.000168
10. ZERO-ONE LOSS=0.288835
    SQUARED LOSS=0.000165

b. Plot a learning curve of training set size vs. zero-one-loss (report the mean performance measured above). Compared to the baseline default error that would be achieved if you just predicted the most frequent class label in the overall data. Discuss the results (e.g., how is zero-one loss impacted by training set size). (6pts)



- **We can see that as we increase our dataset size we diverge further and further away from the accuracy of just guessing the most frequent class label value. This makes sense if we think about it because it means that as we go along we acquire a more accurate sense of what the value might be given a certain data vector. More specifically what the value might be given specific feature attribute values from the inputted test vector.**

c. Plot a learning curve of training set size vs. square-loss. Discuss how zero-one loss performance compares to square-loss.(6 pts)



- **We want the MSE to be minimized and we want the Zero-One loss to be minimized. Because they are on the same dataset both are showing roughly the same behavior. The aforementioned statement is probably why our graphs for both metrics look so similar.**