BYOUNGJIN OH 0030411229

1. NBC details (20 pts)

(a) Write down the mathematical expression for $P(Y \mid X)$ given by the NBC. (2 pts)

$$P(Y|X) \propto P(X|Y)P(Y)$$
$$\propto \sum_{i=1}^{n} P(x_i|Y)P(Y)$$

(b) Suppose your data has binary class labels , i.e $y \in \{0,1\}$, write the expression for predicting the class for a given input row. You will use this expression in your implementation. (1 pts)

$$y_{MAP} \propto \underset{y \in \{0,1\}}{\operatorname{argmax}} P(x_1, x_2, \cdots x_n | y) P(y)$$

(c) State the naive assumption that lets us simplify the expression $P(X \mid Y)P(Y)$. 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)

Naïve assumption assumes that attributes (X) are conditionally independent given the class (Y). Bayes rule was used to simplify the expression. This assumption can be true for the given Yelp data. Because, given the class, outdoorSeating is possibly dependent with each attributes but not more than one attribute. Therefore, no matter attributes are dependent or independent each other, they can be assumed conditionally independence given the class, outdoorSeating.

(d) What part of the expression corresponds to the class prior? Considering the entire Yelp data as the training dataset, calculate the maximum likelihood estimate for the class prior with and without smoothing. What is the effect of smoothing on the final probabilities? (4 pts)

P(Y) corresponds to the class prior. Without smoothing, finding the maximum likelihood will fail. Because, there will be an attribute which dose not appear given Y. In this case, the probabilities will be swiped out.

Ex) P(X|Y) = P(X1|Y) * P(X2|Y) * … * 0 * … * P(Xn|Y) = 0

With Laplace correction (smoothing), we can keep the probabilities and find the maximum likelihood.

(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)

The full set of parameters :

{ state, latitude, longitude, stars, open, alcohol, noiseLevel, attire, priceRange, delivery, waiterService, smoking, caters, goodForGroups, goodForKids, dessert, latenight, lunch, dinner, breakfast, brunch, romantic, initmate, touristy, hipster, divey, classy, trendy, upscale, casual, garage, validate, street, lot, valet, dairy-free, gluten-free, vegan, kosher, halal, soy-free, vegetarian, outdoorSeating }

43 parameters.

(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)

$$P(X_i = x_k \mid \underset{y \in \{0,1\}}{Y = y}) = \frac{\text{Count}(X_i = x_k, Y = y)}{\text{Count}(Y = y)}$$

$$\hat{\theta}_{MLE} = P(\underset{y=\{0,1\}}{X_i = x_k \mid Y = y}) \quad \text{smoothing}$$

$$= \begin{cases} \dfrac{\text{count}(X_i = x_k, Y = y) + 1}{\text{count}(Y = y) + k} & \text{if } \text{count}(X_i = x_k, Y = y) = 0 \\ \dfrac{\text{count}(X_i = x_k, Y = y)}{\text{count}(Y = y)} & \text{otherwise} \end{cases}$$

(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)

$$P(\underset{x \in X}{X = \text{priceRange}}, \underset{y \in Y}{Y = \text{outdoorSeating}})$$

$x \in \{1.0, 2.0, 3.0, 4.0\}$ $y \in \{\text{True}, \text{False}\}$

$$P(x=1.0, y=\text{true}) = \frac{\text{count}(x=1.0, y=\text{true}) + 1}{\text{count}(y=\text{true}) + 4}$$

$$P(x=2.0, y=\text{true}) = \frac{\text{count}(x=2.0, y=\text{true}) + 1}{\text{count}(y=\text{true}) + 4}$$

$$\vdots$$

$$P(x=4.0, y=\text{false}) = \frac{\text{count}(x=4.0, y=\text{false}) + 1}{\text{count}(y=\text{false}) + 4}$$

(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:

(i) Delivery
(ii) Alcohol
(iii) noiseLevel
(iv) attire

What is the effect of smoothing (e.g., any difference compared to Q1d)? Which attribute shows the most association with the class? (4 pts)

The effect of smoothing here is same as the answer of Q1d. The conditional probability distribution results zero.

In my observation, alcohol attribute shows the most associatie with the class outdoorSeating. The probabilities of 'none' and 'full_bar' in alcohol vary depends on class value while other attributes has similar probabilities each values with different class value. 'none' with false is bigger than 'none' with true and 'full_bar' with true is bigger than 'full_bar' with false.

2. Implement a naive Bayes classification algorithm in python. (20 pts)
We will run several tests on your code to assess the accuracy for different samples.

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

Cross-validation is a powerful tool to assess how well your learned model generalizes: a subset of the data is kept aside (called the *validation set*) before training begins. After a model is learned, the validation set can be used to test the performance of the learned model.

In k-fold cross-validation, data is divided into k subsets. In each iteration, one such subset is used as the test set and the remaining $k - 1$ subsets are collectively used as the training set. This process is repeated k times. We then take the average error across

all *k* trials. This gives a more accurate measure of model quality than if you were to hold out some fixed subset of the training data as validation set.

A **learning curve** is a way to plot how the scoring function (e.g., 0/1-loss, squared loss) evolves as size of the training set changes.

(a) For each k in [1,10,50]:

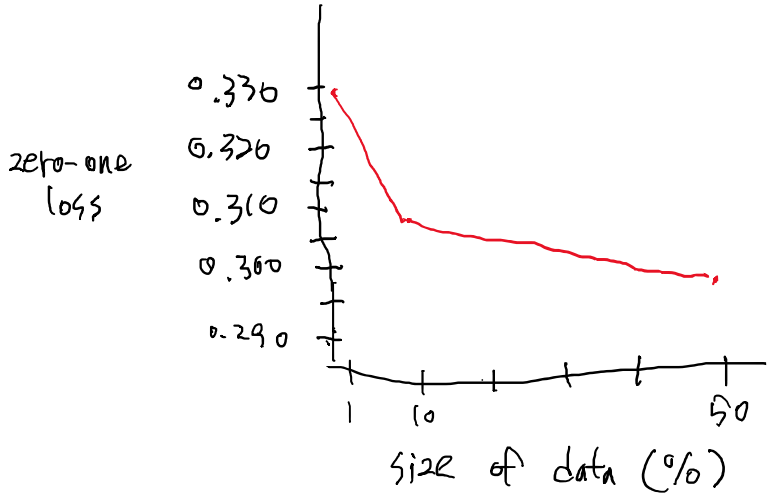
For *i* in [0...9]:

- Randomly sample k % of the data to use as the training dataset.
- Use the remaining (100-k)% of the data as the test dataset.
- Learn a model from the training data and apply it to the test data.
- Measure the performance on the test data using zero-one loss.

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

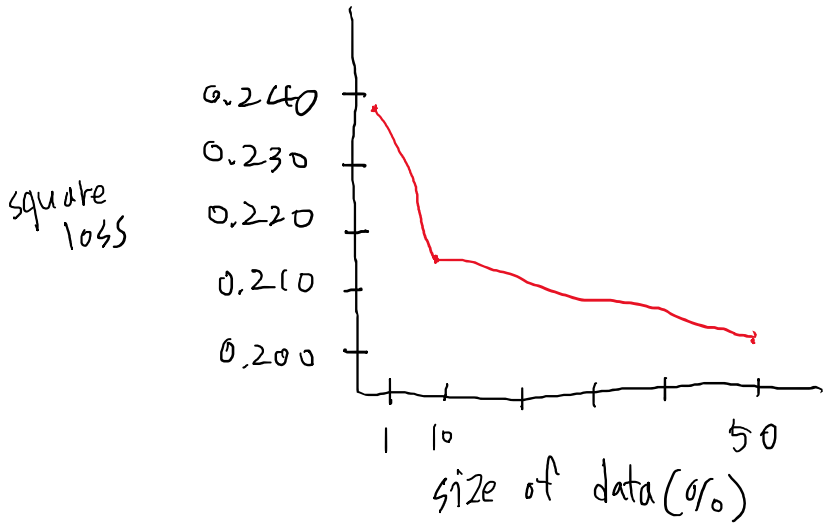
	K=1	K=10	K=50
Zero-one	0.329577	0.308152	0.297770
squared	0.238737	0.215561	0.203946

(b) Plot a learning curve of training set size vs. zero-one-loss (report the mean performance measured above). Compare 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). (6 pts)



The baseline error here is 0.35054 (1 – 16115/(8698 + 16115)). The zero-one loss with small training size is close to the baseline error but as the the training set size increases, the zero-one-loss decreases and the zero-one-loss is getting far from the baseline error.

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



As the training set size increases, the square-loss decreases. The trends of the graphs between zero-one loss and square-loss are same but the value of the square-loss is smaller than zero-one loss.