Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

Using 3 late days

1. NBC details (20 pts)

(a) *Write down the mathematical expression for P(Y | X) given by the NBC.*

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(\frac{y}{x}) = \frac{\prod_{i=1}^{n} P(x_i|y)P(y)}{\sum_{j=1}^{c}\prod_{i=1}^{n} P\left(x_i\middle|y_j\right)P\left(y_j\right)}$$

i = features
j = classes

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(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.*

Given how input represents the feature vector of sample *s*

$$P_1 = \prod_{i=1}^{n} P(x_i|y_1)P(y_1)$$

$$y_1 = 1$$

$$P_2 = \prod_{i=1}^{n} P(x_i|y_2)P(y_2)$$

$$y_2 = 0$$

$$p_1{}' = \frac{p_1}{p_1+p_2} \qquad\qquad p_2{}' = \frac{p_2}{p_1+p_2}$$

If P_1' $>$ 0.5 then prediction = 1
Else, the prediction = 0

(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?*

We assume that all the features of the data are independent of each other.

→ We used the fact that $P(X \cap Y) = P(X)P(Y)$ given X and Y are independent.

→ This assumption is not true for the given Yelp data. The yelp dataset features are not independent. For example, if the restaurant is good for kids and groups, its noise level will probably be louder than a restaurant that is *not* good for kids and groups.

(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?*

In:

$$P\left(\frac{Y}{X}\right) = P\left(\frac{X}{Y}\right) * P(Y)$$

P(Y) corresponds to the class prior.

The Maximum Likelihood Estimate for the class prior:

(e) With smoothing

$$\frac{\text{\# samples with } (y = y_i) + 1}{\text{Total samples} + (\# \, classes)}$$

(ii) Without smoothing

$$\frac{\text{\# samples with } (y = y_i)}{\text{total samples}}$$

Since we add 1 sample for each class, it does affect the actual probabilities but helps in the case when training set does not contain any such sample which gives P(Y/X) = 0 and ensures that there could be a sample possible which is not in the train data but could exist in the test data.

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(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?*

The priors:

P($outdoorSeating = True$) and ($outdoorSeating = False$)

Other parameters:

$$\mathrm{P}\left(\frac{state}{y = True}\right) \& \mathrm{P}\left(\frac{state}{y = false}\right)$$

states = 14
Total = 28

$$\mathrm{P}\left(\frac{latitude}{Y = True}\right) \& \mathrm{P}\left(\frac{latitude}{Y = false}\right)$$

latitude = 5
Total = 10

Similarly:

	Total
longitude	6
stars	18
open	4
alcohol	4
noiseLevel	8
attire	6
priceRange	8
delivery	4
waiter	4
smoking	6
caters	4
goodForGroups	4
goodForKids	4

Binary Columns	27 attributes, 27 x 4 = 108

Total parameters = 2 + 28 + 10 + 6 + 18 + 4 + 4 + 8 + 6 + 8 + 4 + 4 + 6 + 4 + 4 + 4 + 108 = 228

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(f) *Write an expression for an arbitrary conditional probability distribution (CPD) of a discrete attribute Xi 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.*

Y \ X_i	a	b	c	d	… k
0	P(x=a | y = 0)	P(x = b| y = 0)			P(X = k | y = 0)
1	P(x = a | y = 1)	P(x = b | y = 1)			P(X = k | y = 1)

This is the discrete conditional distribution.

$$P(X_i = x \mid Y = y)$$

$$= \frac{P(X_i = x \cap Y = y)}{P(Y = y)}$$

Maximum Likelihood Function with Smoothing:

$$\frac{count(X_i = x, Y = y) + 1}{count(Y = y) + k}$$

Count(…) = Number of samples in dataset D which corresponds to the following properties *X=x Y=y*.

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(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 class label outdoorSeating.*

priceRange → { 1, 2, 3, 4 }
outdoorSeating → { True, False }

$$P(priceRange = 1 \mid outdoorSeating = \text{True}) = \frac{count(priceRange = 1, outdoorSeating = \text{True}) + 1}{count(outdoorSeating = \text{True}) + 4}$$

$$P(priceRange = 4 \mid outdoorSeating = \text{True}) = \frac{count(priceRange = 4, outdoorSeating = \text{True}) + 1}{count(outdoorSeating = \text{True}) + 4}$$

$$P(priceRange = 1 \mid outdoorSeating = \text{False}) = \frac{count(priceRange = 1, outdoorSeating = \text{False}) + 1}{count(outdoorSeating = \text{False}) + 4}$$

$$P(priceRange = 4 \mid outdoorSeating = \text{False}) = \frac{count(priceRange = 4, outdoorSeating = \text{False}) + 1}{count(outdoorSeating = \text{False}) + 4}$$

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(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?*

(i)

outdoorSeating\Delivery	True	False
True	0.157	0.842
False	0.209	0.790

(ii)

outdoorSeating\alcohol	Full bar	Beer & wine
True	0.797	0.202
False	0.702	0.297

(iii)

outdoorSeating\noiseLevel	Very loud	Loud	Quite	Avg
True	0.042	0.105	0.183	0.668
False	0.045	0.09	0.270	0.589

(iv)

outdoorSeating\attire	Casual	Dressy	Formal
True	0.959	0.0386	0.0014
False	0.965	0.0313	0.0032

No smoothing this probability $P(\mathit{attire} = \mathit{formal} \mid \mathit{outdoorSeating} = \mathit{True}) = 0$, stating that there would be no restaurant with outdoorSeating where formal attire is used, which is practically false. So, suing this without smoothing we would get:

$P(outdoorSeating = True \mid X_i = x, attire = formal) = 0$ independent of other features.

Log(P) → -∞ which is impractical, so we smooth out the probability by adding one sample such that its *attire* = *Formal* for *outdoorSeating* = *True*.

Attire shows most association with the label class since its label class since its CPD contains either very high or very low values which clearly indicates that attire plays a major role in predicting the output on its own without being affected much by the other 3 attributes.

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

3. Evaluate the NBC using cross validation and learning curves.

(a)

```
For Train Data Fraction =  0.01
Iteration Number:  0
ZERO-ONE LOSS =  8.491126772522713e-05
SQUARED LOSS =  0.00012778270844816946
Iteration Number:  1
ZERO-ONE LOSS =  0.00012731285011033781
SQUARED LOSS =  0.00011431138340221865
Iteration Number:  2
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  5.1805966076304304e-05
Iteration Number:  3
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  7.725233261723279e-05
Iteration Number:  4
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  7.064658607159695e-05
Iteration Number:  5
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  8.844823253968705e-05
Iteration Number:  6
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  3.135449870057197e-05
Iteration Number:  7
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  3.385010944312899e-05
Iteration Number:  8
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  7.733444317280218e-05
Iteration Number:  9
ZERO-ONE LOSS =  8.474576271186441e-05
SQUARED LOSS =  0.0001704195559079858

ZERO-ONE ERROR MEAN =  2.9696988054742937e-05
SQUARED ERROR MEAN =  8.432058163796982e-05

For Train Data Fraction =  0.1
Iteration Number:  0
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.2307833779974837e-05
Iteration Number:  1
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.163175063106656e-05
Iteration Number:  2
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.211586464341888e-05
Iteration Number:  3
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.0840120126873216e-05
Iteration Number:  4
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.7812470461900057e-05
Iteration Number:  5
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  9.29902995518841e-06
Iteration Number:  6
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.9326962356990725e-05
Iteration Number:  7
ZERO-ONE LOSS =  0.0
```

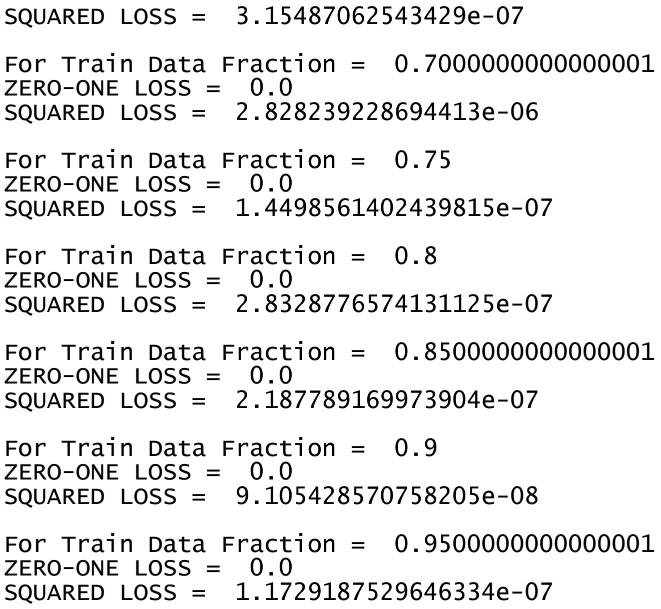
```
SQUARED LOSS =  1.1782306333000258e-05
Iteration Number:  8
ZERO-ONE LOSS =  0.00022450720668133447
SQUARED LOSS =  0.00014823371801332892
Iteration Number:  9
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.3482880934602215e-05

ZERO-ONE ERROR MEAN =  2.2450720668133446e-05
SQUARED ERROR MEAN =  2.668329372363441e-05

For Train Data Fraction =  0.5
Iteration Number:  0
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  6.771174570856554e-07
Iteration Number:  1
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  5.270462218750309e-07
Iteration Number:  2
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.0401174261330438e-06
Iteration Number:  3
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  7.67063274624224e-07
Iteration Number:  4
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.6366984725729306e-06
Iteration Number:  5
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  6.04523104209967e-07
Iteration Number:  6
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  4.423082847467262e-07
Iteration Number:  7
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  3.919335970682046e-07
Iteration Number:  8
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.1909856532138716e-06
Iteration Number:  9
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  3.821775454612783e-07

ZERO-ONE ERROR MEAN =  0.0
SQUARED ERROR MEAN =  8.659971036990933e-07
```

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019

(b & c)
Here since number of 'no' is greater than 'yes', the baseline will be that outdoorSeating = False, this will give us a baseline zero-one error of yes/(yes+no) = 0.326, which is very high as compared to the bayes classifier

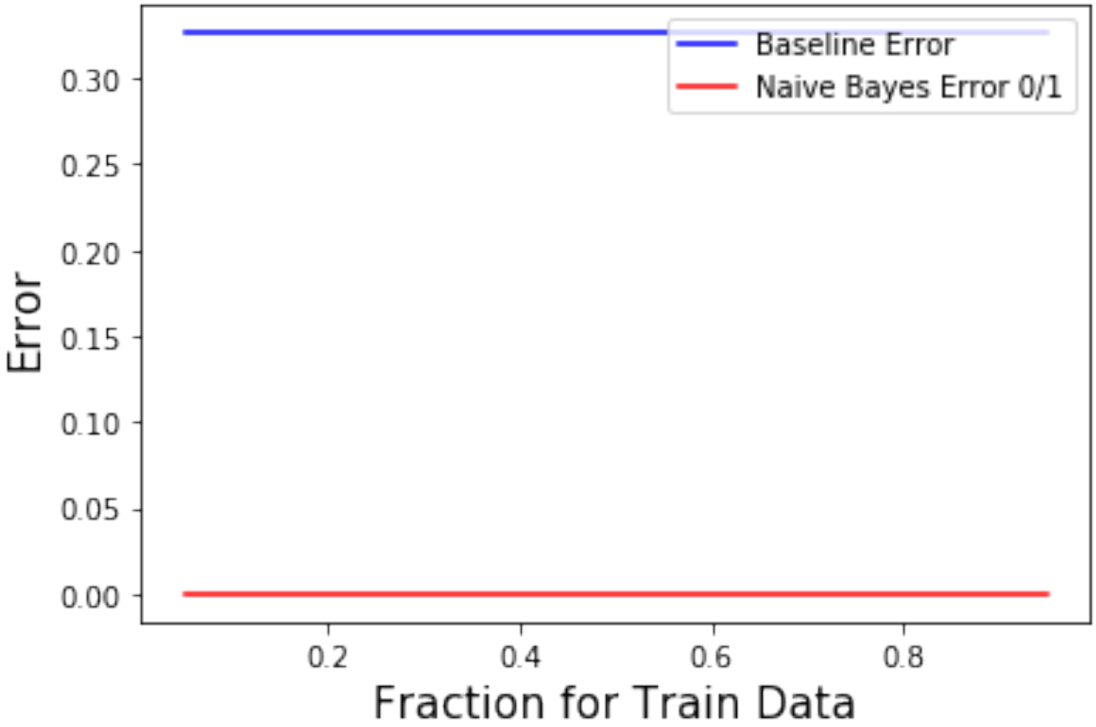
Zero-One Loss tells just, if the test sample is correctly classified or not, Squared Loss indicates the closeness of the test sample with the correct label

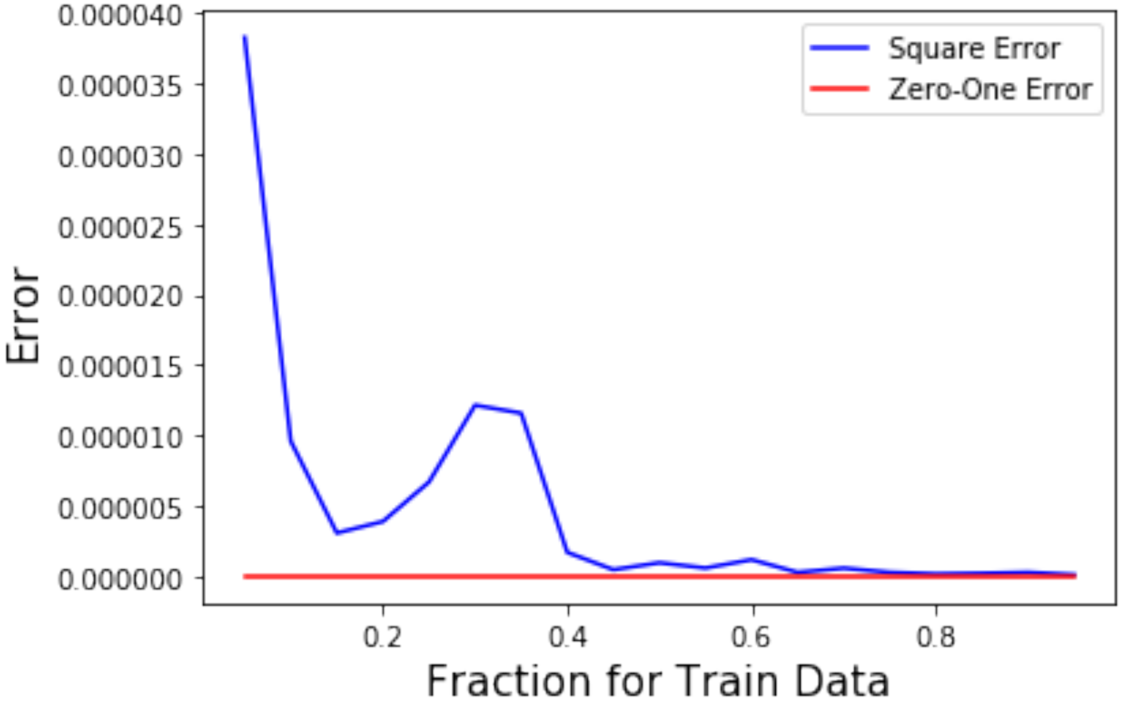
```
For Train Data Fraction =  0.05
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  4.296356997262736e-05

For Train Data Fraction =  0.1
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.5984650574543373e-05

For Train Data Fraction =  0.15000000000000002
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  4.363734240743267e-06

For Train Data Fraction =  0.2
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.4817167423898446e-06

For Train Data Fraction =  0.25
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.4792851451948946e-06

For Train Data Fraction =  0.30000000000000004
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.3382212829244567e-06

For Train Data Fraction =  0.35000000000000003
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.0470230875053162e-06

For Train Data Fraction =  0.4
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.569220499039436e-05

For Train Data Fraction =  0.45
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  5.433506951893648e-07

For Train Data Fraction =  0.5
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  3.9094724200776167e-07

For Train Data Fraction =  0.55
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  5.171363572479467e-07

For Train Data Fraction =  0.6000000000000001
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.199275664978143e-06

For Train Data Fraction =  0.65
ZERO-ONE LOSS =  0.0
```



```
SQUARED LOSS =  3.15487062543429e-07

For Train Data Fraction =  0.7000000000000001
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.828239228694413e-06

For Train Data Fraction =  0.75
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.4498561402439815e-07

For Train Data Fraction =  0.8
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.8328776574131125e-07

For Train Data Fraction =  0.8500000000000001
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  2.187789169973904e-07

For Train Data Fraction =  0.9
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  9.105428570758205e-08

For Train Data Fraction =  0.9500000000000001
ZERO-ONE LOSS =  0.0
SQUARED LOSS =  1.1729187529646334e-07
```

Akash Lankala
PUID: 0027710383
CS 373
Homework 2
October 14th, 2019


Baseline Error
Naive Bayes Error 0/1
Error
Fraction for Train Data


Square Error
Zero-One Error
Error
Fraction for Train Data