cpsc340 A2

zach2940

January 2020

1 Training and Testing Error Curves

The training errror(red dotted line) is lower than the testing error for each iteration but both curves decrease at roughly the same rate and plateus out at a depth of 4

1.2 Depth of 4. Due to the random nature of the samples we get a slightly different graph. Use more training data instead of validation data and use cross validation

2.1 p(spam) = 6/10 p(not spam) = 4/10

2.2 1/6 5/6 2/6 1 1/4 3/4

2.3 naive bayes
$p(x_i|y_i = "spam")p(y_i = "spam") > p(x_i|y_i = "notspam")p(y_i = "notspam")$
Assuming features are independent,
$\frac{1}{6} * \frac{5}{6} * \frac{2}{6} * \frac{6}{10} > 1 * \frac{1}{4} * \frac{3}{4} * \frac{4}{10}$ $0.0278 is not > 0.075$ so it is not spam.

2.4 $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$
$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Any training set such that $p(x_i|y_i = "spam")p(y_i = "spam") or p(x_i|y_i = "notspam")p(y_i = "notspam")$ does not give 0

2.5 Lunar ['car' 'fact' 'gun' 'video'] talk.*

2.6 Naive Bayes (ours) validation error: 0.188 Naive Bayes (sklearn) validation error: 0.201

2.7 $O(kd)$

3. For k=1, Training Error is: 0.0 Testing Error is: 0.0645 For k=3, Training Error is: 0.0275 Testing Error is: 0.066 For k=10, Training Error is: 0.0725 Testing Error is: 0.097

The training error increases with k whereas it decreases with depth of decision tree. The approximation error also decreases with k.

Insert 2 images

When k=1, training error is 0 because the model is complicated and just memorises where all the training points are.

I would iterate over a range of k values and select the one with the lowest approximation error.