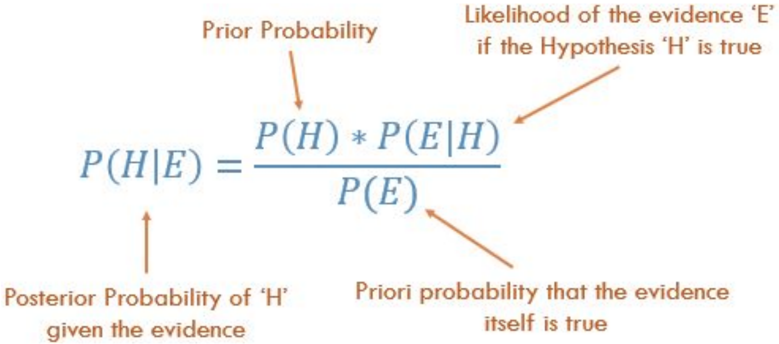
spam-musubayes

An exploration of Naive Bayes algorithms on the Spambase data set

Hannah Galbraith

What is Naive Bayes?

- A family of classification algorithms based on Bayes' Theorem



- All of the algorithms share a common principle: every feature that is being classified is independent of the value of any other feature. Since features are not always independent, this is why we label the algorithm "naive".

Spambase Data Set

- Created by researchers at Hewlett Packard Labs in the ‘90s
- Contains 4,601 spam and non-spam email samples (~60% non-spam, ~40% spam) with 57 features.
- Features 1-48: percentage of words in the email that match a given word
- Features 49-54: percentage of characters in the email that match a given character
- Feature 55: average length of uninterrupted sequences of capital letters
- Feature 56: length of longest uninterrupted sequence of capital letters
- Feature 57: total number of capital letters in the email
- Each sample has a label: ‘0’ for non-spam and ‘1’ for spam

Gaussian Naive Bayes

- Bayes by definition finds probabilities for either a binary or a discrete list of attribute values.
- Gaussian Bayes provides a method of normalizing continuous data into values we can apply Bayes' Theorem to.
- The Gaussian distribution tends to be the most fitting model of natural occurrences.
- Instead of just looking at frequency, we find the mean and standard deviation of the values to represent the distributions.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / (2\sigma^2)}$$

Results from Gaussian Naive Bayes Program

```
[1] "Confusion Matrices for each fold:"
[1] "TN   |   FP"
[1] "-----|-----"
[1] "FN   |   TP"
     [,1] [,2]
[1,]  181   75
[2,]    7  184
[1] "***************"
     [,1] [,2]
[1,]  196   64
[2,]    8  151
[1] "***************"
     [,1] [,2]
[1,]  187   89
[2,]    9  174
[1] "***************"
     [,1] [,2]
[1,]  218   78
[2,]    6  148
[1] "***************"
     [,1] [,2]
[1,]  210   78
[2,]    3  187
[1] "***************"
```

```
[1] "Accuracy of each fold (%):"
[[1]]
[1] 81.65548

[[2]]
[1] 82.81623

[[3]]
[1] 78.64924

[[4]]
[1] 81.33333

[[5]]
[1] 83.05439

[[6]]
[1] 82.12766

[[7]]
[1] 80.46218

[[8]]
[1] 80.7947

[[9]]
[1] 83.57588

[[10]]
[1] 81.41026
```

```
[1] "Average Accuracy (%):"
[1] 81.58794
[1] "Max Accuracy (%):"
[1] 83.57588
[1] "Min Accuracy (%):"
[1] 78.64924
[1] "Standard Deviation:"
[1] 1.43511
```

Multivariate Bernoulli Naive Bayes

- Simply looks at the frequency of a given feature for each class
- Using the training set, you calculate the following:
    - $P(x_i = 0 \mid \text{non-spam})$
    - $P(x_i = 1 \mid \text{non-spam})$
    - $P(x_i = 0 \mid \text{spam})$
    - $P(x_i = 1 \mid \text{spam})$

On the test set, you use the above probabilities to predict each sample. E.g.:

Sample : $<x_1 = 1, x_2 = 0, \ldots, x_n = 1>$

$P(\text{spam} \mid x_1 = 1, x_2 = 0, \ldots, x_n = 1) = P(\text{spam}) * P(x_1 = 1 \mid \text{spam}) * P(x_2 = 0 \mid \text{spam}) * \ldots * P(x_n = 1 \mid \text{spam})$

Results from Bernoulli Naive Bayes Program

```
[1] "Confusion Matrices for each fold:"
[1] "TN    |    FP"
[1] "-----|-----"
[1] "FN    |    TP"
     [,1] [,2]
[1,]   20  236
[2,]   40  151
[1] "***************"
     [,1] [,2]
[1,]   21  239
[2,]   41  118
[1] "***************"
     [,1] [,2]
[1,]   19  257
[2,]   32  151
[1] "***************"
     [,1] [,2]
[1,]    8  288
[2,]   13  141
[1] "***************"
     [,1] [,2]
[1,]   23  265
[2,]   35  155
[1] "***************"
```

```
[1] "Accuracy of each fold (%):"
[[1]]
[1] 38.25503

[[2]]
[1] 33.17422

[[3]]
[1] 37.03704

[[4]]
[1] 33.11111

[[5]]
[1] 37.23849

[[6]]
[1] 35.74468

[[7]]
[1] 35.92437

[[8]]
[1] 34.21634

[[9]]
[1] 38.46154

[[10]]
[1] 37.39316
```

```
[1] "Average Accuracy (%):"
[1] 36.0556
[1] "Max Accuracy (%):"
[1] 38.46154
[1] "Min Accuracy (%):"
[1] 33.11111
[1] "Standard Deviation:"
[1] 1.978964
```

References

http://blog.aylien.com/naive-bayes-for-dummies-a-simple-explanation/

https://www.quora.com/Why-do-I-need-to-use-the-Gaussian-Naive-Bayes-for-continuous-data-and-not-the-classical-Naive-Bayes

https://archive.ics.uci.edu/ml/datasets/spambase

Leila Hawana, Ph.D. candidate in Machine Learning, Portland State University