Enron emails (text analytics)

Tool: Bag of words model, Classification And Regression Trees (CART), Random Forests, Naive Bayes Classifier.

The Analytics Edge: Predictive coding provides an innovative way for lawyers to review large sets of documents. Lawsuits often mean the review of letters, memoranda, and files to determine which ones are relevant. With more email communications, the amount of documents to review has exploded. It is no longer economical to review documents individually to find the typically smaller set of sensitive ones. Predictive coding uses sample sets reviewed by senior attorneys and analytics to predict the most important documents from the many available.

1 Overview

Enron is an American energy company established in 1985. It was originally involved in transmitting and distributing electricity and natural gas throughout the US. Before 2001, Enron employed nearly 20,000 staff, with claimed revenues of nearly $111 billion. By the end of 2001, it was revealed that the reported financial conditions were based on systematic accounting fraud (the *Enron scandal*).

The Enron corpus is a publicly-available dataset of the email messages sent or received by about 150 Enron senior managers. These data were obtained by the Federal Energy Regulatory Commission during its investigation. This corpus is one of the few publicly-available mass collection of real emails available for study (as such collections are often bond by privacy laws).

Predictive coding allows software to take information entered by people and generalize it to a larger group of documents. Traditionally, one searches a set of documents to identify documents relevant to a case. Rather, predictive coding considers the context. Instead of having legal experts going through all documents, one uses a training set, which has been categorized by legal experts. Then, the use of text analytics helps predict the categorization of future documents.

The California energy crisis (2000-2001) was a situation in which California had a shortage of electric supply caused by market manipulations and illegal shutdowns of pipelines. The state suffered from multiple blackouts, while Enron gamed the market, which was deregulated. The Federal Energy Regulatory Commission (FERC) investigated Enron involvement in the crisis and this ultimately led to a $ 1.52 billion settlement. As part of the investigations, FERC released the emails from some of the executives at Enron. One approach is to search for keywords like "electricity bid", "energy schedule" in emails, and then carefully review which ones are responsive. Predictive coding using text analytics as an alternate and newer way to do this. The data come from a 2010 Text Retrieval Conference (attorneys labeled emails responsive to energy schedule or bids).

Key Question: Is it possible to correctly classify an email (as responsive or non-responsive) based on the email text?

2 Summary

Data: large unstructured datasets containing emails and their corresponding classification (responsive or non-responsive).

Model: A classification model (e.g., logistic regression, CART, Random Forest, Naive Bayes classifier) that predicts the class of an email based on its text.

Value and Decision: the model replaces the expensive option of manually searching all emails.

3 Naive Bayes classifier

The Naive Bayes classifier is a simple classification rule based on Bayes' rule, which can be used with the bag of words model. Given a problem instance (say a document) to be classified, represented by a vector $\{x_1, \ldots, x_p\}$ of terms (predictors), a Naive Bayes classifier assigns to this instance probabilities $P(C_k|x_1, \ldots, x_p)$ for each of the K possible classes or outcomes C_k. Using Bayes' rule, let's rewrite this as:

$$P(C_k|x_1, \ldots, x_p) = \frac{P(C_k) \cdot P(x_1, \ldots, x_p|C_k)}{P(x_1, \ldots, x_p)}.$$

The denominator $P(x_1, \ldots, x_p)$ does not depend on C_k, so it can be treated as a constant Z (since the values of $x_1, \ldots, x_p$ are given). Let's thus focus on the numerator $P(C_k) \cdot P(x_1, \ldots, x_p|C_k)$, which is equal to the joint probability $P(C_k, x_1, \ldots, x_p) = P(x_1, \ldots, x_p, C_k)$. Using the chain rule, we can write:

$$\begin{aligned} P(x_1, \ldots, x_p, C_k) &= P(x_1|x_2, \ldots, x_p, C_k) \cdot P(x_2, \ldots, x_p, C_k) \\ &= P(x_1|x_2, \ldots, x_p, C_k) \cdot P(x_2|x_3, \ldots, x_p, C_k) \cdot \ldots \cdot P(x_{p-1}|x_p, C_k) \cdot P(x_p|C_k) \cdot P(C_k). \end{aligned}$$

We now make a (naive) hypothesis of conditional independence of the features, that is, x_i is conditionally independent of every other feature x_j (with $i \neq j$). With this hypothesis in place, we can write:

$$\begin{aligned} P(C_k|x_1, \ldots, x_p) \propto P(C_k, x_1, \ldots, x_p) &= P(x_1, \ldots, x_p, C_k) \\ &= P(x_1|C_k) \cdot P(x_2|C_k) \cdot \ldots \cdot P(x_{p-1}|C_k) \cdot P(x_p|C_k) \cdot P(C_k) \\ &= P(C_k) \prod_{i=1}^{p} P(x_i|C_k). \end{aligned}$$

Recalling that $P(x_1, \ldots, x_p)$ is a constant Z, we can finally write:

$$P(C_k|x_1, \ldots, x_p) = \frac{1}{Z} P(C_k) \prod_{i=1}^{p} P(x_i|C_k).$$

The prediction rule is to assign a class k such that:

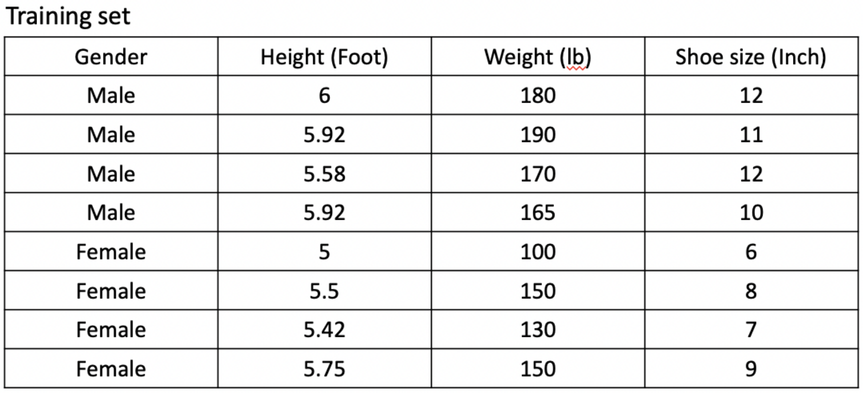
$$\underset{k=1,\cdots,K}{\arg\min}\ P(C_k) \prod_{i=1}^{p} P(x_i \mid C_k),$$

which is estimated from the training set and then applied to the test set.

To estimate the $P(x_i|C_k)$, one can, for example, use a Gaussian distribution. Denoting with μ_{ki} and σ^2_{ki} the mean and variance of x_i in class k (i.e., x_{ki}), we get:

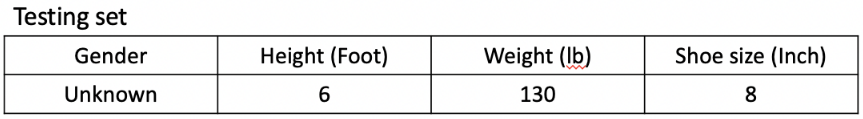
$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{ki}}}\, e^{-\frac{(x_{ki}-\mu_{ki})^2}{2\sigma^2_{ki}}}.$$

3.1 Example



Training set

Gender	Height (Foot)	Weight (lb)	Shoe size (Inch)
Male	6	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5	100	6
Female	5.5	150	8
Female	5.42	130	7
Female	5.75	150	9

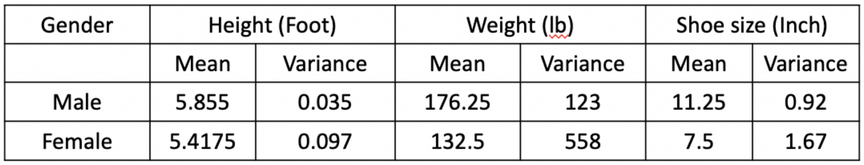


Testing set

Gender	Height (Foot)	Weight (lb)	Shoe size (Inch)
Unknown	6	130	8

$$\text{Posterior(Male)} = \frac{P(\text{M})P(\text{Height} \mid \text{M})P(\text{Weight} \mid \text{M})P(\text{ShoeSize} \mid \text{M})}{Z}$$
$$\text{Posterior(Female)} = \frac{P(\text{F})P(\text{Height} \mid \text{F})P(\text{Weight} \mid \text{F})P(\text{ShoeSize} \mid \text{F})}{Z}$$

- $P(\text{M}) = \frac{4}{8} = 0.5 \quad P(\text{F}) = \frac{4}{8} = 0.5$
- Assume Height, Weight, Shoe size follow Gaussian distributions.



Gender	Height (Foot)		Weight (lb)		Shoe size (Inch)	
	Mean	Variance	Mean	Variance	Mean	Variance
Male	5.855	0.035	176.25	123	11.25	0.92
Female	5.4175	0.097	132.5	558	7.5	1.67

$$P(\text{Height} \mid \text{M}) = \frac{1}{\sqrt{2\pi \times 0.035}} e^{-\frac{(6-5.855)^2}{2 \times 0.035}} \ldots$$

- For the testing data, Posterior(Male) < Posterior(Female) → Female!

3.2 Back to R!

In R, Naive Bayes Classifier can be carried out with the `e1071` package.

`model <- naiveBayes(formula,data)`
`model$apriori` To see $P(C_1), \cdots, P(C_K)$
`model$tables$`x_i To see mean and standard deviation of $P(x_i \mid C_k)$
`predict <- predict(model,newdata=test,type="class")` Prediction

3.3 Advantages and disadvantages of Naive Bayes classifier

Advantages

- Works with binary and multi-class problems
- Computationally efficient

Disadvantages

- Assumption of conditional independence of the predictors, which may (or may not) lead to poor performance
- Assumption of Gaussian distribution for $p(x_i \mid C_k)$, $i = 1, \cdots, p$