Bear Test Analyzer:
An experiment in the Bayesian classification of text

Rob Menke

February 8, 2004

1 Problem Definition

The Bear Test provides a means to determine a subject's hidden personality traits. The test consists of a combination of multiple-choice and short answer questions. The multiple-choice questions have direct mappings to their underlying answers and will not be considered here.

The text answers are more difficult to process mechanically. Currently, a human reader reviews the answers and provides a short paragraph of analysis that more often than not follows a standard boilerplate. Since the answers are structured and repetitive, it was decided to classify them into categories that could easily be assigned to future tests.

The goal of the Bear Test Analyzer (Figure 1) is to produce an automated system using Bayesian decision theory for analyzing and classifying Bear Test submissions so that the need for a human reader is reduced (not eliminated) and the results can be delivered to the subject much more quickly than the current system allows.

2 A Short Introduction to Bayesian Decision Theory

A classifier is defined over a finite set of classes (or *states of nature*) Ω, each denoted by ω_i, where $1 \leq i \leq |\Omega|$. A general classifier consists of a set of discriminant functions, one per class, $g_i(\mathbf{x} \in \mathbf{X}), 1 \leq i \leq |\Omega|$. The *feature set* $\mathbf{x}$ is considered to be of class ω_i if and only if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \text{for all } i \neq j. \tag{1}$$

A Bayes classifier is represented this way. Bayesian decision theory tells us that to minimize the probability of error in making a decision, the rule

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x}) \qquad \omega_i, \omega_j \in \Omega \tag{2}$$

should be used. In terms of a general classifier, $g_i(\mathbf{x}) \equiv P(\omega_i|\mathbf{x})$.

Problems arise when the probabilities $P(\omega_i|\mathbf{x})$ are not well known. Fortunately, the celebrated Bayes formula can rewrite the probabilities into a much more manageable form. Bayes formula states that

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}. \tag{3}$$

Informally, this is stating that

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{4}$$

where the *a posteriori* (posterior) probability is the probability of the feature set $\mathbf{x}$ belonging to class ω_i and the *likelihood* represents the probability of feature set $\mathbf{x}$ being seen in class ω_i. The belief in the natural occurrence of ω_i without additional evidence is defined by the *a priori* (prior) probability. The *evidence* is used as a normalization factor so that the probabilities sum to one. Because of its importance in connecting classes to feature sets, we define the likelihood of ω_i given $\mathbf{x}$ by the special notation $L(\omega_i|\mathbf{x}) \equiv P(\mathbf{x}|\omega_i)$. Since $P(\omega_i|\mathbf{x}) \propto P(\omega_i)L(\omega_i|\mathbf{x})$ with respect to ω_i, (2) can be rewritten as:

$$\text{Decide } \omega_i \text{ if } P(\omega_i)L(\omega_i|\mathbf{x}) \geq P(\omega_j)L(\omega_j|\mathbf{x}) \qquad \omega_i, \omega_j \in \Omega. \tag{5}$$

This is the *maximum a posteriori* (MAP) decision rule.

Most of the probabilities encountered are vanishingly small. Since the logarithmic function is monotonic and therefore invariant under maximum:

$$\omega_{MAP} = \arg\max_{\omega_i \in \Omega} \{P(\omega_i)L(\omega_i|\mathbf{x})\} = \arg\max_{\omega_i \in \Omega} \{\ln P(\omega_i) + l(\omega_i|\mathbf{x})\} \tag{6}$$

where $l(\omega_i|\mathbf{x}) \equiv \ln L(\omega_i|\mathbf{x})$. This reduces the problem to the determination of $P(\omega_i)$ and $l(\omega_j|\mathbf{x})$.

3 Likelihood and the Naïve Bayes Assumption

In analysis, the text is split into words and common connective words are removed. The resultant arbitrarily-sized vector is the feature set. The typical answer is twenty words long, so even after this conversion the likelihood of an exact match is practically zero. We can make the (not completely correct) assumption that a conditional independence exists between the words ($x \in \mathbf{x}$) relative to

the class of the full text, and simply compute the product of the likelihoods relative to each word of the set:

$$\hat{L}(\omega_i|\mathbf{x}) = \prod_{j=1}^{|\mathbf{x}|} P(x_j|\omega_i). \tag{7}$$

This is known as the *Naïve Bayes assumption* and—while the assumption of independence is normally incorrect—it has been shown both theoretically and empirically to be very accurate under certain conditions [1]. Using Naïve Bayes in this manner is common in statistical natural language parsing and is known as the *bag of words* approach [3].

Under logarithms, (7) becomes

$$\hat{l}(\omega_i|\mathbf{x}) = \sum_{j=1}^{|\mathbf{x}|} \ln P(x_j|\omega_i). \tag{8}$$

Combining (6) and (8) gives the feature-based form of the classifier:

$$\omega_{MAP} = \arg\max_{\omega_i \in \Omega} \left\{ \ln \hat{P}(\omega_i) + \sum_{j=1}^{|\mathbf{x}|} \ln \hat{P}(x_j|\omega_i) \right\}. \tag{9}$$

where the (unknown) true probabilities $P(\omega_i)$ and $P(x_j|\omega_i)$ have been replaced by the training set estimators discussed below.

4 Estimators and the Problem of Infinite Surprise

Estimating the prior probability is easy. If c is the total number of samples and c_i is the number of samples of class ω_i, then $\hat{P}(\omega_i) = c_i/c$.

The estimation of the likelihood probabilities is more difficult. Note that the notation $P(x_j|\omega_i)$ is really shorthand for $P(X_j = w_k|\omega_i)$; that is, the probability that the jth word in the text is $w_k \in V$ when the class of the text is ω_i. From this the definition is straightforward:

$$P(X_j = w_k|\omega_i) = \frac{c_{ijk}}{c_{ij}}, \tag{10}$$

where c_{ijk} is the number of times word w_k has appeared in position j in text labeled ω_i and c_{ij} is the total number of words in position j in all the samples from class ω_i.

We can make another simplification here by assuming that the probabilities are positionally independent; that is, $P(X_m = w_k|\omega_i) = P(X_n = w_k|\omega_i)$ for all $1 \leq m < n$. If c_i is the total number of words in all examples of class ω_i and c_{ik} is the total number of times word w_k has appeared *in any position* in examples of class ω_i, then our estimator is

$$\hat{P}(x_j|\omega_i) = \hat{P}(X_j = w_k|\omega_i) = \frac{c_{ik}}{c_i}. \tag{11}$$

The problem here is that w_k is defined over some vocabulary V, which is expected to be very large. There is a strong possibility that many of the estimates $\hat{P}(x_j|\omega_i)$ as defined above would be zero. Examining (9), we see that any probability of zero automatically eliminates that class from consideration: $\ln \hat{P}(x_j|\omega_i)$ goes to negative infinity. Information theorists define *surprise*$(\mathbf{x}) \equiv -\ln P(\mathbf{x})$, so this condition is known as the *problem of infinite surprise.*

What we need to do is come up with an estimate that insures $\hat{P}(x_j|\omega_i) > 0$ for all x_j and ω_i. One such function is the *Laplace estimator*, which is deceptively simple:

$$\hat{P}(x_j|\omega_i) = \frac{c_{ik}+1}{c_i+|V|}. \tag{12}$$

The Laplace estimator simply creates a virtual sample for each class containing every word in the vocabulary. Its strength is its simplicity; its weakness is that it places far too much weight on the unseen words, diluting the probability estimates. For a classifier with a large corpus this is tolerable; however, the estimated probabilities returned are closer to the *a priori* estimate than the true probabilities.

References

[1] DOMINGOS, P., AND PAZZANI, M. J. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning 29*, 2-3 (1997), 103–130.

[2] DUDA, R. O., HART, P. E., AND STORK, D. G. *Pattern Classification*, second ed. John Wiley & Sons, Inc., New York, 2001.

[3] MANNING, C. New fronteers in statistical natural language processing. Workshop, Seventeeth National Conference on Artificial Intelligence, July 2000.