Glossary

Bayes Optimal Classifier

The optimal classification rule if the true data distribution is known. In classification settings, the Bayes optimal classifier predicts the most likely label y given a feature vector $\mathbf{x}$ according to $P(y|\mathbf{x})$. In practice, the Bayes optimal classifier cannot be used directly, because the underlying data distribution is not typically known. However, it can be applied on distribution estimates or it can be used to derive theoretical optimal upper bounds on a classifier's performance.

Bayes Rule

Notation: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Bayes rule expresses the probability $P(A|B)$ as an expression of $P(B|A)$. This conversion is convenient in situations when we want to infer A from a given B but it is easier to estimate $P(B|A)$ from the data.

Categorical Naive Bayes (Binomial Naive Bayes)

Categorical Naive Bayes, also known as Binomial Naives Bayes, is a model used for data with catagorical features that take on one of a finite number of values, such as 0 or 1, "CA", "NJ", or "NY", or "red", "blue", "green" or "yellow" (in contrast to arbitrary natural or real numbers).

Class probability

The class probability (class prior) $P(y)$ represents the probability of y in the absence of any further information (i.e., features). For example, if spam email is 60% of the incoming email traffic, then $P(y = \text{spam}) = 0.06$ in the context of a spam classifier.

Conditional independence

Two random variables (A and B) are conditionally independent given a third random variable (C) if $P(A, B|C) = P(A|C)P(B|C)$.

Conditional probability with Naive Bayes

Naive Bayes classifier assumes that all features are conditionally independent given the class label:
$P(x_1, \ldots, x_d|y) = P(x_1|y) \ldots P(x_d|y)$

Feature extraction

The process of deriving quantitative values from a set of data to create a feature vector to represent each data point. Feature values are to be informative and not redundant. For example, if the data point is a name, useful features could be the number of letters in that name, the number of occurrences of any given letter, prefix, suffix, or the number of vowels.

Feature Hashing

A fast and effective way of creating a vector for a "bag of features," derived from an initial set of data.

Laplace Smoothing (plus one smoothing, additive smoothing)

A method for inserting additional "hallucinated" data into a small data set to avoid pathological examples (and divisions by zero) through unseen features. Laplacian smoothing is a convenient way to bias the probability estimates towards your prior believes in the presence of very little data. If a lot of data is observed, the few added examples become irrelevant. For example, adding one "heads" and one "tails" coin flip result to a set of coin flip data ensures each possible outcome is represented in the data.

Maximum likelihood estimation (MLE)

A method we use with the Bayes optimal classifier to estimate the distribution. The estimate is based on training data drawn from the distribution. MLE is a useful method for estimating the parameters of a distribution from observed data.

Naive Bayes assumption

The assumption that all features of a given data set are conditionally independent. This assumption makes high dimensional problems tractable. Conditional independence guarantees that given a label y, all features are independent of one other and can be estimated separately from each other. In other words, instead of estimating the (intractable) distribution of a d-dimensional feature vector $P(\mathbf{x}|y)$, we estimate the probabilities of d one-dimensional features $P(x_i|y)$.

Naive Bayes classifier

The Naive Bayes classifier uses the Naive Bayes assumption to estimate $P(x|y)$ efficiently - i.e., the likelihood of a feature value x given the label y. With Bayes Rule, we can then infer $P(y|\mathbf{x})$ - the probability of the label y given a data point $\mathbf{x}$.