EECS 553: Machine Learning (ECE) **University of Michigan**

The Naïve Bayes Classifier

Winter 2023 **Clayton Scott**

Naïve Bayes is a family of plug-in methods. It could be generative or discriminative, parametric or nonparametric, and linear or nonlinear, depending on design choices.

1 The Naïve Bayes Assumption

Let $\boldsymbol{X} = [X_1 \cdots X_d]^T \in \mathbb{R}^d$ denote the random feature vector in a classification problem, and Y the corresponding label. The Naïve Bayes method assumes that, given Y, the features $X_1, \ldots, X_d$ are *independent*. This means that the class-conditional pmfs/pdfs $g_k(\boldsymbol{x})$ can be factored

$$g_k(\boldsymbol{x}) = \prod_{j=1}^{d} g_{kj}(x_j)$$

where g_{kj} is the pmf/pdf of the feature X_j. A naïve Bayes classifier thus has the form

$$\widehat{f}(\boldsymbol{x}) = \arg\max \ \widehat{\pi}_k \prod_{j=1}^{d} \widehat{g}_{kj}(x_j),$$

where $\widehat{g}_{kj}$ is an estimate of g_{kj} from labeled training data. Different NB methods differ in how they estimate these marginal class-conditional pmfs/pdfs.

2 Example: Document Classification

Suppose we wish to classify documents into categories like "business," "politics," "sports," etc. A simple feature representation of a document is a so-called *bag-of-words* model, in which a document is represented as a vector

$$\boldsymbol{X} = [X_1 \cdots X_d]^T$$

where d is the number of words in the vocabulary, and X_j is the number of times word j occurs in the document.

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data, and let

$$\widehat{\pi}_k = \frac{|\{i : y_i = k\}|}{n}.$$

This is the same estimate of π_k used for LDA.

To estimate g_{kj}, introduce

$$n_k = |\{i : y_i = k\}|$$
$$n_{kj\ell} = |\{i : y_i = k \wedge x_{ij} = \ell\}|$$

where x_{ij} is the jth feature of the ith training instance, and $\wedge$ indicates the logical "and" operator.

Then a natural estimate of $g_{kj}(\ell)$ is

$$\widehat{g}_{kj}(\ell) = \frac{n_{kj\ell}}{n_k}.$$

In other words, the estimated probability that (in a random document) word j occurs ℓ times in class k is the empirical frequency with which that event occurs in the training data. This estimate may be derived as the maximum likelihood estimate for a multinomial data model, although we may also accept it as an intuitive estimate.

One issue that occurs in practice is that we may have a class of documents where a certain word never occurs. For example, suppose all sports documents in our training data contain the word "ball". Then a new document that does not contain the word ball would never be classified as sports, because the estimated probability $\widehat{g}_{kj}(0)$ is zero, and all the probabilities for all words get multiplied together. To overcome this, a simple fix is to modify the estimate so that all counts for all words occur with some nonzero probability. For example, we may redefine

$$\widehat{g}_{kj}(\ell) = \frac{n_{kj\ell} + 1}{n_k + L + 1} \tag{1}$$

where L is the maximum number of times a word can occur in a document. These probabilities still sum to one (when summed over ℓ) and are never zero. This technique is known as additive smoothing or Laplace smoothing and can be derived as a Bayesian estimate based on the multinomial likelihood and uniform prior.

3 Other Models

Naïve Bayes can be used when the X_j are continuous random variables. Then we could model X_j as a univariate Gaussian and estimate the parameters via maximum likelihood. Alternatively, we could estimate the marginal densities g_{kj} with a nonparametric density estimator, such as the kernel density estimator.

Exercises

1. (★) As a simplification of the bag-of-words model, consider the numerical representation of a document given by

$$X_j = \begin{cases} 1 & \text{if } j^{th} \text{ word occurs at least once in document,} \\ 0 & \text{otherwise.} \end{cases}$$

   State the NB classifier in this case and include the version based on additive smoothing.

2. (★★) Assume the features are continuous and Gaussian distributed, and that for any feature, the two classes have a common variance parameter. Using maximum likelihood estimates for the Gaussian model parameters, state the NB classifier and argue that it is linear.

3. (☆☆) Show that the formula in (1) is the minimum mean squared error estimate based on the multinomial likelihood and uniform prior. How does the formula change if the uniform prior is generalized to a Dirichlet prior?