Third practicals in Machine learning 1 – 2024 – Paper 1

1 Naive Bayes (September)

Naive Bayes (NB) is a particular form of classification that makes strong independence assumptions regarding the features of the data, conditional on the classes (see Bishop section 4.2.3). Specifically, NB assumes each feature is independent given the class label. In contrast, when we looked at probabilistic generative models for classification in the lecture, we used a full-covariance Gaussian to model data from each class, which incorporates correlation between all the input features (i.e. they are not conditionally independent).

If correlated features are treated independently, the evidence for a class will be overcounted. However, Naive Bayes is very simple to construct, because by ignoring correlations the *class-conditional likelihood*, $p(\mathbf{x}|\mathbf{C}_k)$, is a product of D univariate distributions, each of which is simple to learn:

$$p(\mathbf{x}|\mathbf{C}_k) = \prod_{d=1}^{D} p(x_d|\mathbf{C}_k) \tag{1}$$

Consider a document classification task, that classifies your documents into K classes $\mathbf{C}_k$. To do this you first make a bag-of-words (BoW) representation of your entire training set. A BoW is a vector $\mathbf{x}_n$ of dimension D for each document indicating whether each word in the vocabulary appears in the document (i.e. the words go into a bag and are shaken, losing their order so only their presence matters). This means that $x_{ni} = 1$ if word i is present in document n, $x_{ni} = 0$ otherwise. You can think of D as the vocabulary size of the training set, but it may also contain tokens or special features. Your training set therefore consists of an N by D matrix of word counts $\boldsymbol{X}$, and the target matrix $\boldsymbol{T}$, who's rows consist of the row vectors $\mathbf{t}_n^T = (t_{n1}, \ldots, t_{nk})$, one-hot-encoded such that $\mathbf{t}_n^T = (0, \ldots, 1, \ldots, 0)$ with the scalar 1 at position i if $n \in \mathbf{C}_i$. Assume we know $p(\mathbf{C}_i) = \pi_i$ (with the constraint $\sum_{i=1}^{K} \pi_i = 1$). We can model the word counts using different distributions (in the practice homework we modeled it with a Poisson distribution), for this question, we will use a Bernoulli distribution model and only account for the presence/absence of words, hence each word is distributed according to a Bernoulli distribution with parameter θ_{dk}, when conditioned on class $\mathbf{C}_k$:

$$p(\mathbf{x}|\mathbf{C}_k, \theta_{1k}, \ldots, \theta_{Dk}) = \prod_{d=1}^{D} \theta_{dk}^{x_d}(1 - \theta_{dk})^{1-x_d} \tag{2}$$

with distribution parameters $\theta_{dk} = P(x_d = 1|\mathbf{C}_k)$.

With this information answer the following questions:

(*a*) Write down the data likelihood, $p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\Theta})$ without NB independence assumption at the beginning. Then, derive the data likelihood for the *general* K classes naive Bayes classifier, stating where you make use of the product rule and the naive Bayes assumption.

You should write the likelihood in terms of $p(x_d|\mathbf{C}_k)$, meaning you should not assume the explicit Bernoulli distribution.

(*b*) Write down the data log-likelihood $\ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\Theta})$ for the Bernoulli model.

(*c*) Solve for the MLE estimators for θ_{dk}. Express in your own words how the result can be interpreted.

(*d*) Write $p(\mathbf{C}_1|\mathbf{x})$ for the *general* K classes naive Bayes classifier.

(*e*) Write $p(\mathbf{C}_1|\mathbf{x})$ for the Bernoulli model.