— *Solution notes* —

Third practicals in Machine learning 1 – 2023 – Paper 1

1 Naive Bayes (practice) (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 $\mathbf{t}_n^T = (t_{n1}, \dots, t_{nk})$, one-hot-encoded such that $\mathbf{t}_n^T = (0, \dots, 1, \dots, 0)$ 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 $\mathbf{C}_k$:

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

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.

Answer:

Develop using the IID and NB independence assumptions and product rule of probabilities:

$$p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\Theta}) \overset{\text{IID}}{=} \prod_{n=1}^{N} p(\mathbf{t}_n, \mathbf{x}_n|\boldsymbol{\Theta}) \overset{\text{prod rule}}{=} \prod_{n=1}^{N} p(\mathbf{t}_n|\boldsymbol{\Theta})p(\mathbf{x}_n|\mathbf{t}_n, \boldsymbol{\Theta})$$

$$= \prod_{n=1}^{N}\prod_{k=1}^{K} \Big(\underbrace{p(t_{nk} = 1|\boldsymbol{\theta}_k)}_{\text{prior}} \quad \underbrace{p(\mathbf{x}_n|t_{nk} = 1, \boldsymbol{\theta}_k)}_{\text{class conditional likelihood}} \Big)^{t_{nk}}.$$

Note that since $\mathbf{t}_n = \{0, .., 1, .., 0\} \in \mathbb{R}^K$, then, it holds that $\sum_{k=1}^{K} t_{nk} = 1$. For that reason, we can develop the joint likelihood as a product over all classes K, where, t_{nk} can work as a selection mechanism that filters out all the cases that $t_{nk} \neq 1$. Moreover, we can write:

$$p(C_k|\boldsymbol{\theta}_k) = p(t_{nk} = 1|\boldsymbol{\theta}_k)$$

.

Then, in terms of $p(\mathbf{C}_i) = \pi_i$ and Naive Bayes assumption that $p(\mathbf{x}_n|\mathbf{C}_k, \boldsymbol{\theta}_k) = \prod_{d=1}^{D} p(x_{nd}|\mathbf{C}_k, \theta_{dk})$ we get:

$$p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{k=1}^{K}(\pi_k \prod_{d=1}^{D} p(x_{nd}|\mathbf{C}_k, \theta_{dk}))^{t_{nk}}.$$

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

Answer: With the Bernoulli model $p(x_{nd}|\mathbf{C}_k, \theta_{dk}) = \theta_{dk}^{x_{nd}}(1 - \theta_{dk})^{1-x_{nd}})^{t_{nk}}$ we get

$$p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{k=1}^{K}(\pi_k \prod_{d=1}^{D} \theta_{dk}^{x_{nd}}(1 - \theta_{dk})^{1-x_{nd}})^{t_{nk}}.$$

Then, we can calculate the log-likelihood as

$$\ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left(\ln \pi_k + \sum_{d=1}^{D}(x_{nd} \ln \theta_{dk} + (1 - x_{nd}) \ln(1 - \theta_{dk}))\right).$$

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

Answer: The MLE estimator is obtained by solving $\frac{\partial \ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_{dk}} = 0$ for θ_{dk}. The derivation is as follows:

$$
\begin{aligned}
& \frac{\partial \ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\theta})}{\partial \theta_{dk}} = 0 \\
\Rightarrow \quad & \frac{\partial}{\partial \theta_{dk}} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\ln \pi_k + \sum_{d=1}^{D} (x_{nd} \ln \theta_{dk} + (1 - x_{nd}) \ln(1 - \theta_{dk})) \right) = 0 \\
\Rightarrow \quad & \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\sum_{d=1}^{D} \frac{\partial}{\partial \theta_{dk}} (x_{nd} \ln \theta_{dk} + (1 - x_{nd}) \ln(1 - \theta_{dk})) \right) = 0 \\
\Rightarrow \quad & \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\sum_{d=1}^{D} \delta_{dk} (\frac{x_{nd}}{\theta_{dk}} - \frac{1 - x_{nd}}{1 - \theta_{dk}}) \right) = 0 \\
\Rightarrow \quad & \sum_{n=1}^{N} t_{nk} \left(\frac{x_{nd}}{\theta_{dk}} - \frac{1 - x_{nd}}{1 - \theta_{dk}} \right) = 0 \\
\Leftrightarrow \quad & \sum_{n=1}^{N} t_{nk} \frac{x_{nd} - x_{nd}\theta_{dk} - \theta_{dk} + x_{nd}\theta_{dk}}{\theta_{dk}(1 - \theta_{dk})} = 0 \\
\Leftrightarrow \quad & \frac{1}{\theta_{dk}(1 - \theta_{dk})} \sum_{n=1}^{N} t_{nk} \left(x_{nd} - \theta_{dk} \right) = 0 \\
\overset{*}{\Rightarrow} \quad & \sum_{n=1}^{N} t_{nk} \, \theta_{dk} = \sum_{n=1} t_{nk} \, x_{nd} \\
\overset{**}{\Leftrightarrow} \quad & N_k \, \theta_{dk} = N_{dk} \\
\Leftrightarrow \quad & \theta_{dk} = \frac{N_{dk}}{N_k},
\end{aligned}
$$

where at $\overset{*}{=}$ we used that $\theta_{dk}(1 - \theta_d k) \neq 0$ and at $\overset{**}{=}$ we defined N_k as the count of mails in class k (sum of cases for which $t_{nk} = 1$) and N_{dk} the count of word d occuring in a class k mail (sum of cases for which both $t_{nk} = 1$ and $x_{nd} = 1$). The MLE estimator θ_{dk} is thus frequency of emails containing word d for class $\mathbf{C}_k$.

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

Answer: The class posterior is obtained using bayes rule as

$$
p(\mathbf{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{C}_1)p(\mathbf{C}_1)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathbf{C}_1)p(\mathbf{C}_1)}{\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{C}_k)p(\mathbf{C}_k)} = \frac{\pi_1 \prod_{d=1}^{D} p(x_d|\theta_{d1})}{\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} p(x_d|\theta_{dk})}.
$$

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

Answer: In the above we substitute $p(x_d|\theta_{d1}) = \theta_{d1}^{x_d}(1 - \theta_{d1})^{1 - x_d}$ and obtain

$$
p(\mathbf{C}_1|\mathbf{x}) = \frac{\pi_1 \prod_{d=1}^{D} \theta_{d1}^{x_d}(1 - \theta_{d1})^{1 - x_d}}{\sum_{k=1}^{K} \pi_k \prod_{d=1}^{D} \theta_{dk}^{x_d}(1 - \theta_{dk})^{1 - x_d}}.
$$