Week-6, Practice, Solution

Question-1

A Bernoulli NB model for a binary classification problem with m features has $2m$ parameters that are used to model the class conditional distribution, $P(x \mid y)$. There is one parameter for each feature belonging to each class. That gives us $2m$ parameters. In addition to this, the priors give another 2 parameters. Therefore, the total number of parameters is $2m + 2$. However, only $2m + 1$ parameters are independent. This is because, the prior probabilities sum to 1. So, knowing one prior probability is enough to estimate the other.

Question-2

$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{evidence}}$$

Question-3

F_1, ROC curve, precision and recall are used to evaluate the performance of a classification model.

Question-4

This is just restating the Naive Bayes assumption.

Question-5

NB models the conditional probability of the feature vector given the label. This is typically what happens in generative classifiers.

Question-6

The advantages of operating in the log-space is both mathematical and computational. Underflow is typically caused when many small fractions are multiplied together.

Common Data for questions 7, 8 and 9

Consider a balanced training dataset $D = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{100}$ for a binary classification problem, where the feature vector $x = (x_1, x_2)$ is a two-dimensional binary vector, i.e., each feature is binary. The class label y is indexed using 1 and 2. A sample feature matrix and label vector is given below:

$$X = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{bmatrix}$$

Assume that the features are conditionally independent given the class labels. A Bernoulli Naive-Bayes classifier is trained for this data (refer to PPA-1 to see how this is done computationally). Specifically, the following parameter matrix is estimated:

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

This matrix is to be understood as follows. For features x_1 and x_2:

$$p_{ij} = P(x_i = 1 \mid y = j)$$

In p_{ij}, the first index stands for the feature and the second stands for the class-label.

Question-7

The various feature vectors are as follows:

$$
\begin{gathered}
(0,0)\\
(0,1)\\
(1,0)\\
(1,1)
\end{gathered}
$$

Question-8

We need to compute the value of the following probability:

$$P(x = (1, 1) \mid y = 1)$$

Using the class conditional independence assumption:

$$P(x = (1, 1) \mid y = 1) = P(x_1 = 1 \mid y = 1) \cdot P(x_2 = 1 \mid y = 1)$$

We can now read off these values from the parameter matrix:

$$p_{11} \cdot p_{21}$$

Question-9

Consider the following expression:

$$\frac{(1-p_{11}) \cdot p_{21}}{(1-p_{11}) \cdot p_{21} + (1-p_{12}) \cdot p_{22}}$$

We need to find what this is equal to. We shall look at all distinct terms in the expression one by one:

(1) $1 - p_{11}$

This has something to do with the conditional probability of x_1 with respect to class 1:

$1 - p_{11} = P(x_1 = 0 \mid y = 1)$

(2) p_{21}

This has something to do with the conditional probability of x_2 with respect to class 1:

$p_{21} = P(x_2 = 1 \mid y = 1)$

(3) $1 - p_{12}$

This has something to do with the conditional probability of x_1 with respect to class 2:

$1 - p_{12} = P(x_1 = 0 \mid y = 2)$

(4) p_{22}

This has something to do with the conditional probability of x_2 with respect to class 2:

$p_{22} = P(x_2 = 1 \mid y = 2)$

Now, plugging all of this together, the expression becomes:

$$\frac{P(x_1 = 0 \mid y = 1) \cdot P(x_2 = 1 \mid y = 1)}{P(x_1 = 0 \mid y = 1) \cdot P(x_2 = 1 \mid y = 1) + P(x_1 = 0 \mid y = 2) \cdot P(x_2 = 1 \mid y = 2)}$$

Using the Naive Bayes' assumption, we can convert this into:

$$\frac{P(x = (0,1) \mid y = 1)}{P(x = (0,1) \mid y = 1) + P(x = (0,1) \mid y = 2)}$$

Since the dataset is balanced, the priors are equal to 0.5. So, using the Bayes' theorem, this is nothing but:

$$P(y = 1 \mid x = (0,1))$$