TDA231 - Algorithms for Machine Learning & Inference

Chalmers University of Technology

LP3– 2017-02-06

By:
Karabo Ikaneng, 891024-8239, kara.ikaneng@gmail.com
Elias Svensson, 920406-3052, elias.svensson.1992@gmail.com

Goal

Classification

1 Theoretical problems

1.1 Bayes Classifier, 6 points

Information:

Features:

$$x = (rich, married, healthy)$$

Classes:

1 = Content

0 = Not Content

The response of each feature for the respective classes:

$$1 : (1,1,1), (0,0,1), (1,1,0), (1,0,1)$$

$$0 : (0,0,0), (1,0,0), (0,0,1), (0,1,0)$$

Bayes classifier:

$$P(t_{new} = k|X, t, x_{new}) = \frac{P(x_{new}|t_{new} = k, X, t)P(t_{new} = k)}{\sum_j P(x_{new}|t_{new} = j, X, t)P(t_{new} = j)} \tag{1}$$

Naive-Bayes

$$P(x_{new}|t_{new} = k, X, t) = \prod_{d=1}^{D} P(x_d^{new}|t_{new}, X, t) \tag{2}$$

Naive-Bayes for binomial distribution:

$$P(x_{new}|t_{new} = k, X, t) = \prod_{i=1}^{n} P_{ki}^{x_i}(1 - P_{ki})^{1-x_i} \tag{3}$$

The prior distribution for the respective classes 1 and 0:

$$P(t_{new} = 1) = \frac{1}{2}$$

$$P(t_{new} = 0) = \frac{1}{2}$$

The matrice given below is for descriptive purposes, and should be considered as two independent row vectors for calculations in equation 3.

$$\begin{array}{c} \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{ccc} R & M & H \end{array} \\ \begin{pmatrix} \frac{3}{4} & \frac{1}{2} & \frac{3}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \end{array}$$

Where:

R = Rich ; M = Married and H = Healthy

Therefore, the probability of being content(1) and married (M), using index this notation is $P_{1,12} = \frac{1}{2}$.

Likewise, the probability of not being content(0) and married (M), using this index notation $P_{0,12} = \frac{1}{2}$

From equation 3:

$$P(x_{new} = [0,1,1]|t_{new} = 1, X, t) = P_{1,11}^{0}\, P_{1,12}^{1}\, P_{1,13}^{1}\, (1-P_{1,11})^{1}(1-P_{1,12})^{0}(1-P_{1,13})^{0} \quad (4)$$

$$P(x_{new} = [0,1,1]|t_{new} = 0, X, t) = P_{1,11}^{0}\, P_{0,12}^{1}\, P_{0,13}^{1}\, (1-P_{0,11})^{1}(1-P_{0,12})^{0}(1-P_{0,13})^{0} \quad (5)$$

Substituting the appropriate indexes into equation 4 and 5 respectively:

$$P(x_{new} = [0,1,1]|t_{new} = 1, X, t) = \frac{1}{2} * \frac{3}{4}(1 - \frac{3}{4}) = \frac{3}{32}$$

$$P(x_{new} = [0,1,1]|t_{new} = 0, X, t) = \frac{1}{4} * \frac{1}{4}(1 - \frac{1}{4}) = \frac{3}{64}$$

Therefore, the Bayes Classifier is given as:

$$P(t_{new} = 1|X, t, x_{new}) = \frac{\frac{3}{32} * \frac{1}{2}}{\frac{3}{32} * \frac{1}{2} + \frac{3}{64} * \frac{1}{2}} = \frac{2}{3}$$

b)

Since the features are independent, we can simply leave out the data pertaining to "healthy". Therefore, when training and classifying using the Naive-Bayes classifier, we simply use the available independent data. We can then calculate the Bayes Classifier as illustrated below.

Using the same method as before, however adjusting for the missing data, we ge the two independent row vectors given in matrice notation for descriptive purposes. Note that the priors remain the same, since they are not dependent on the number of features, but rather the number of classes which remain unchanged at 2.

$$\begin{array}{c} \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{cc} R & M \end{array} \\ \begin{pmatrix} \frac{3}{4} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} \end{array}$$

$$P(x_{new} = [0, 1]|t_{new} = 1, X, t) = \frac{1}{2}(1 - \frac{3}{4}) = \frac{1}{8}$$

$$P(x_{new} = [0, 1]|t_{new} = 0, X, t) = \frac{1}{4}(1 - \frac{1}{4}) = \frac{3}{6}$$

Therefore the Bayes-Classifier is calculated as:

$$P(t_{new} = 1|X, t, x_{new}) = \frac{\frac{1}{8} * \frac{1}{2}}{\frac{1}{8} * \frac{1}{2} + \frac{3}{16} * \frac{1}{2}} = \frac{2}{5}$$

1.2 Extending Naive-Bayes, 4 points

The Naive-Bayes assumption is only valid when the features are independent. However, in the given problem the data is not independent since only one age related feature can be 1 for a individual "data sample". In order to use Naive-Bayes we can collapse x_1, x_2, x_3 into a single variable. Therefore, $x_1^{'} \in (0, 1, 2)$ where , $x_1^{'} = 0$ corresponds to $x_1 = 1$, $x_1^{'} = 1$ corresponds to $x_2 = 1$ and $x_1^{'} = 2$ corresponds to $x_3 = 1$.

2 Practical Problems

2.1 Bayes Classifier, 5 points

a)

$$P(x_{new}|y_{new} = \pm 1, X, y) = P(x_{new}|\hat{\mu}_{\pm}, \hat{\sigma}^2_{\pm}) \tag{6}$$

where $\mu_{\pm}$ and $\sigma^2_{\pm}$ are the MLE parameters. The Multivariate Spherical Gaussian Distribution is therefore given as:

$$P(x_{new}|\hat{\mu}_{\pm}, \hat{\sigma}^2_{\pm}) = \frac{1}{2\pi^{\frac{d}{2}} \times \hat{\sigma}^2_{\pm}} \exp(\frac{1}{2\hat{\sigma}^2_{\pm}}||x - \hat{\mu}_{\pm}||^2_2) \tag{7}$$

The Bayes-Classifier for the two respective classes 1 and -1 is given below. Note that the prior probabilities $P(t_{new} = \pm 1)$ are equal and have therefore been omitted since they would simply factor out of the respective equations.

$$P(y_{new} = 1|x_{new}, X, y) = \frac{\frac{1}{\hat{\sigma}^2_1} \exp(\frac{1}{2\hat{\sigma}^2_1}||x - \hat{\mu}_1||^2_2)}{\frac{1}{\hat{\sigma}^2_1} \exp(\frac{1}{2\hat{\sigma}^2_1}||x - \hat{\mu}_1||^2_2) + \frac{1}{\hat{\sigma}^2_{-1}} \exp(\frac{1}{2\hat{\sigma}^2_{-1}}||x - \hat{\mu}_{-1}||^2_2)}$$

$$P(y_{new} = -1|x_{new}, X, y) = \frac{\frac{1}{\hat{\sigma}^2_{-1}} \exp(\frac{1}{2\hat{\sigma}^2_{-1}}||x - \hat{\mu}_{-1}||^2_2)}{\frac{1}{\hat{\sigma}^2_1} \exp(\frac{1}{2\hat{\sigma}^2_1}||x - \hat{\mu}_1||^2_2) + \frac{1}{\hat{\sigma}^2_{-1}} \exp(\frac{1}{2\hat{\sigma}^2_{-1}}||x - \hat{\mu}_{-1}||^2_2)}$$

b-c) See code.

d)

Both classifiers model the given data equally well. The training data also has sufficient spread (variance) between samples for the two classes to be distinguished from one another.

Table 1: 5-fold-cross validation error using different classifier models

Classifier	5-fold-cross validation error
Bayes	0 %
New	0 %

The new classifier function is more efficient, as it utilises less computation power than the spherical Bayes.However, the spherical Bayes has more functionality than the new classifier, as it can calculate the probability for which some data belongs to a certain class.

2.2 Handwritten digit recognition,5 points

a-b) See code.

c)

Comparison from 2.2a:

Table 2: 5-fold-cross validation error for the unscaled dataset

Pixels	5-fold-cross validation error
Errors for class 8	85
Errors for class 5	37
Total Errors	122
Total Errors Ratio	5.55%

Comparison from 2.2b:

Table 3: 5-fold-cross validation error for the scaled feature vector

Variance	5-fold-cross validation error
Errors for class 8	247
Errors for class 5	258
Total Errors	505
Total Errors Ratio	22.95%

Table 2 illustrates the comparison of the unscaled feature vector. The total error ratio is lower for the unscaled than the scaled feature vector. However, the difference between the errors for the respective classes is larger in the unscaled feature vector than the scaled feature vector.

In the first instance the variance in the unscaled data set is large and therefore there is a significant difference in the errors of the two classes, but once the data set is scaled between to the values [0,1] the variance between the two classes becomes similar and the errors begin to approximate each other. As a result of this, the errors are equally bad since the are both receiving the same scaled data.

d)

To improve the results we could obtain more data, in addition to this we could reduce noise in the data, using appropriate filters. We could also explore other Classification models to see if they predict data better, especially in the case of the scaled feature vector.