Laboratory 2 Report
Model fitting and classification

17-06-2019

Diogo Correia s264324
diogo.valacorreia@studenti.polito.it

1 Theoretical Cheatsheet

1.0.1 Bayes Theorem

Describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

1.0.2 Gaussian Mixture Model

For **d dimensions**, the Gaussian distribution is a vector $x = (x^1, x^2, ..., x^d)^T$ is defined by the equation 2 [1].

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))} \tag{2}$$

Where μ is the mean and Σ is the covariance matrix of the Gaussian.

Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated.

2 Implementation & Results

2.1 Exercise 1

In this exercise emploid a dataset containing labelled data for two classes, i.e. males and females. Every row of the dataset contains three numbers: the gender (1=male, 2=female), the height (cm) and the weight (kg) of each person in the dataset.

Is asked to fit a **class-conditional Gaussian multivariate** distribution to these data, and **visualize the probability density function**.

2.1.1 Results

The plot the data of each class can be seen in figure 1.

The visualization of the histogram of weight and height can be seen in figure 2 and 3

The calculation of the maximum likelihood estimate of the mean and covariance matrix under a multivariate Gaussian model, independently for each class can be seen as a 2D representation of a pdf in figure 4 for the female class and in figure 5 for the male class.

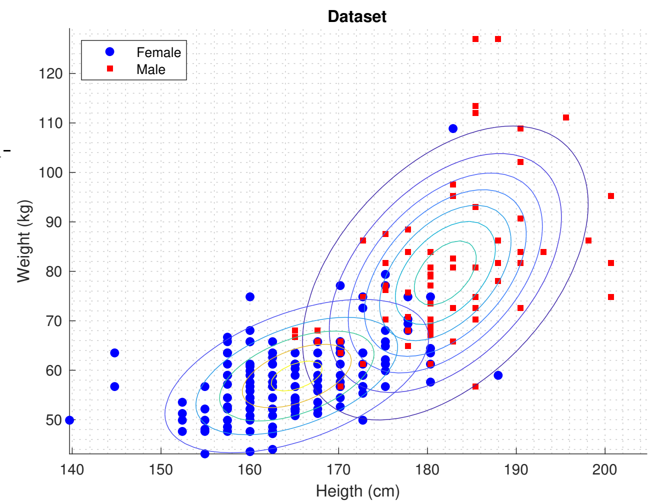

Figure 1: 2D scatter plot of each class

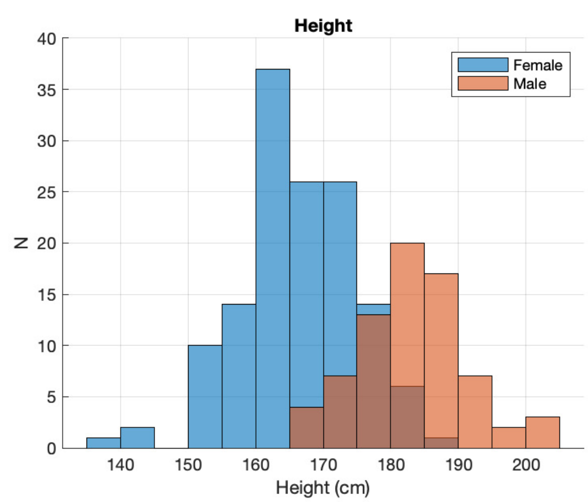

Figure 2: Histogram of the height of both classes

is the Gaussian model good for these data? TODO

The visualization the 2D joint pdf of weight and height can be seen in figure 4 for the female class and in figure 5 for the male class.

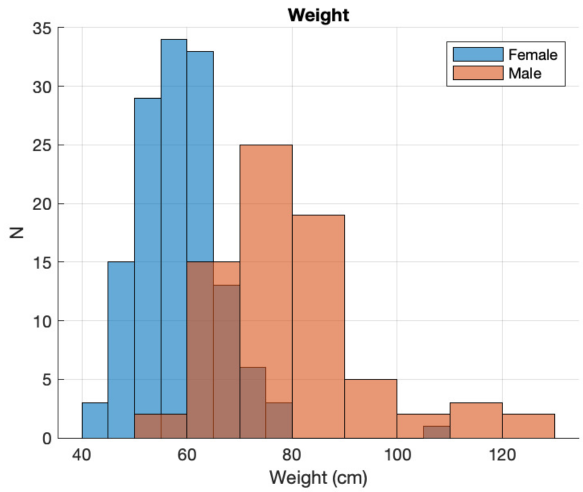

Figure 3:

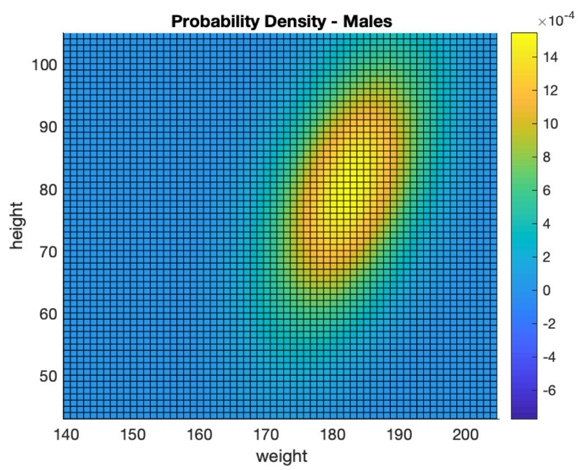

Figure 4: Multivariate Gaussian probability density function for the female class

2.2 Exercise 2

This exercise provides a dataset containing features for two classes, i.e. documents talking about Microsoft Windows, and documents talking about X Windows. The objective was to fit the parameters employed by a Naive Bayes Classifier, using a Bernoulli model.

2.2.1 Results

The plots of the class-conditional densities can be seen in figures 6 and 7.

The uninformative features are the ones that have high class-conditional probability but are common to both documents. This makes them not relevant since they

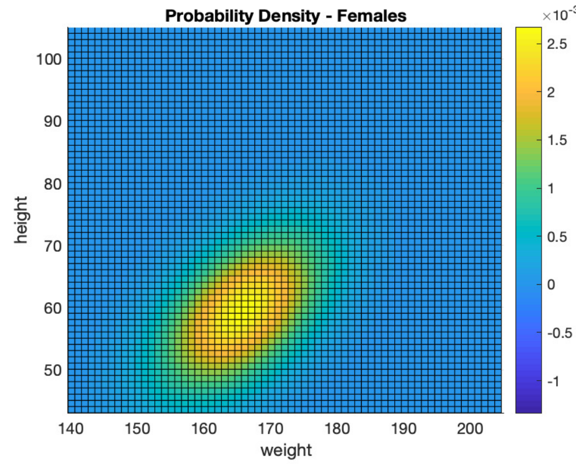

Figure 5: Multivariate Gaussian probability density function for the male class

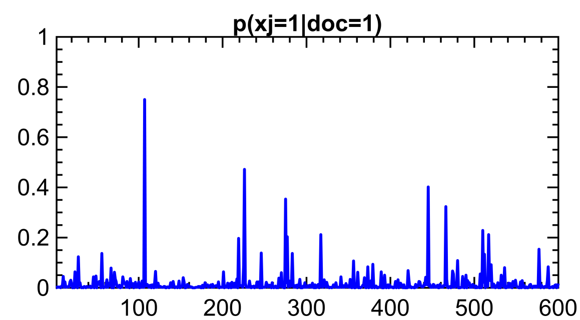

Figure 6: class-conditional densities of document 1

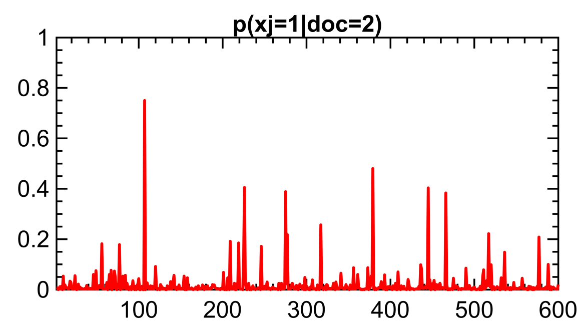

Figure 7: class-conditional densities of document 2

don't give any information.

Those features can be seen in the table 1.

subject	which	this	but	will	article	with	you	all

Table 1: Uninformative features

2.3 Exercise 3

In this exercise we have to design a Naive Bayes Classifier) for the Bag of Words features for document classification that have been prepared in section 2.2.

2.3.1 Results

The problem was completed but the result seem to not make a lot of sense. The accuracy of the classifier on the training and test data can be seen in table 2.

Error Test: 34.6667%	Error Train: 91.5556%

Table 2: Accuracy of the classifier on the training and test data

2.4 Exercise 4

This exercise employs the height/weight data already employed in section 2.1, and performs model fitting and classification using several versions of Gaussian discriminative analysis.

2.4.1 Results

The first one, *Two-class quadratic discriminant analysis*, the accuracy in classifying the test data was of 90.9091%.

The accuracy of the *quadratic discriminant analysis with diagonal covariance matrices* was the same has the first case 90.9091%.

Finally the accuracy of *quadratic discriminant analysis with shared covariance matrix* was 36.3636%.

References

[1] Enrico Magli. Slides: Statistical learning and neural, ict for smart societies.