# Project-I by Group Sydney

**Diego Antognin and Jason Racine**
EPFL
diego.antognini@epfl.ch, jason.racine@epfl.ch

## Abstract

## 1 Regression

## 2 Data Description

The train-data for regression consists of $N = 2800$ input ($\mathbf{X}$) and output ($\mathbf{y}$) data samples. Each input sample is a vector $\mathbf{x}_n$ with dimension $D = 76$. Out of these 76 variables, 63 are real, 3 binary, 4 categorical with 3 categories, 6 are categorical with 4 categories.

We also have test-data of size $N = 1200$ without their corresponding output. Our goal is to produce predictions for those data, as well as an approximation of the test-error.

## 3 Data visualization and cleaning

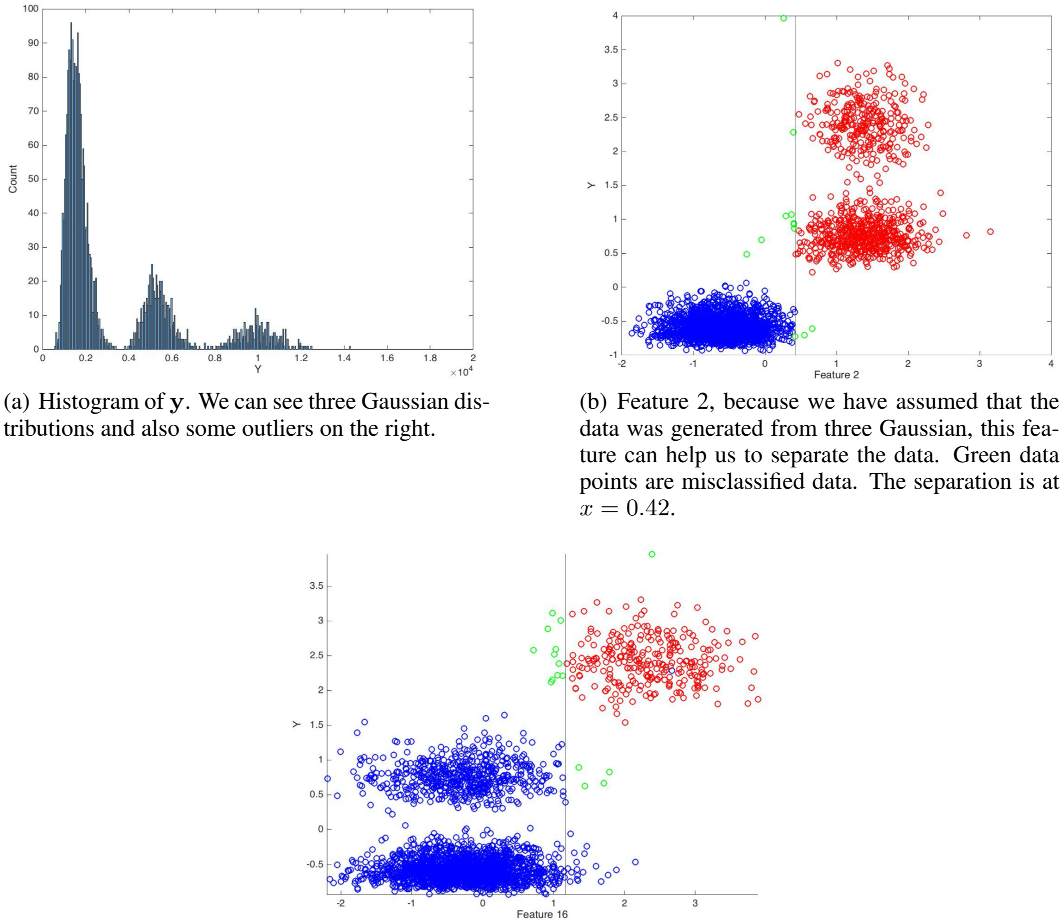
We first have plot the distribution of our features (plot not shown because too big for the 76 features). As expected, they are not center and we should normalized them (each cluster will be normalized independently). The Figure 1(a) shows an histogram of the output ($\mathbf{y}$) and we can conclude our data seem to be a combination of three Gaussian distributions. It will be used later in order to separate the data in three sets and apply different regression models on them. We can also observe that each cluster have different sizes : 1946, 576 and 278. Moreover, on the right, we can see some data points (there are 2) which have a higher values than the others. We consider them as outliers and we will remove.

To separate the data, we have observed that the feature 2 and 16 could help us. Figure 1(b). We can observe 11 misclassified data (green points), which we will remove them in order to not corrupt our model. So, this feature can allow us to find the first cluster. For the two others clusters, we need to oberve the feature 16. Figure 1(c). We can observe 14 misclassified data (green points and one blue). They also will be consider as outliers and remove. We set the threshold in order to minimize the number of misclassified data samples. Those thresholds are $0.42$ for the feature 2 and $1.17$ for the feature 16.

We are also interested about the correlation between the input and output variables. We have observed the correlation for each cluster and conclude that for the first cluster, they are mainly in $[-0.1, 0.1]$, except two features which are highly correlated. For the second and third cluster, also mainly in $[-0.1, 0.1]$ but this time, there are more correlated features ( 15). Moreover, the features seem not have correlation between them.

We use dummy encoding for the categorical variables (for a categorical variable of size $k$, we need $k - 1$ features), which gives us a total of 93 input variable.

We can note that the rank of our input matrix $\mathbf{X}$ is rank-deficient with a rank of 66 instead of 76, and is rank-deficient of 65 for each cluster.



(a) Histogram of $\mathbf{y}$. We can see three Gaussian distributions and also some outliers on the right.

(b) Feature 2, because we have assumed that the data was generated from three Gaussian, this feature can help us to separate the data. Green data points are misclassified data. The separation is at $x = 0.42$.

(c) Feature 16, because we have assumed that the data was generated from three Gaussian, this feature can help us to separate the data. Green data points are misclassified data. We can see see a blue data point which is misclassified, it will be an outlier. The separation is at $x = 1.17$.

Figure 1: