**Probabilistic Models**

With Naive Bayes, we can classify data instance x to belong to the class of k, for which probability P( Y=k | X = x) is maximal, with Y being the label.

Knowing P(Y, X) = P(Y) P(X | Y)

**P(Y = k | X = x) = P(Y=k, X=x) / P(X=x) = P(X = x | Y=k) [P(Y=k) / P(X = x)]**

The main criterion for learning probabilistic models is maximizing the likelihood, which has two versions:

**Generative:** MUL P(xn, yn) with n 🡪 [1..N]

**Discriminative:** MUL P(yn | xn) with n 🡪 [1..N]

We require a parametric model for the conditional distributions, called a parameter vector w, s.t: P( Y | w), P(X | Y,w), P(Y, X,w).

Multinoulli gives the probability for each label Y.

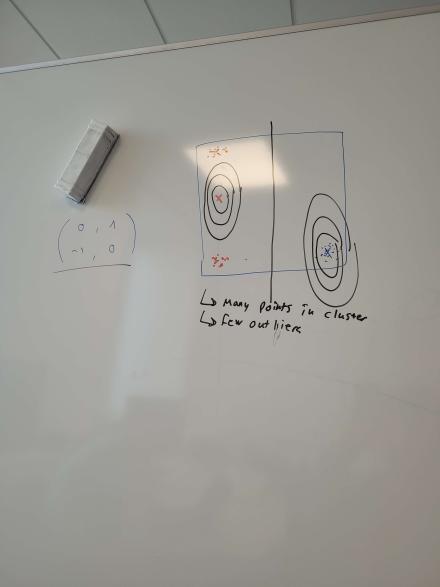
The model depends on the nature of the input. If the input is categorical then the table is OK because all values can be taken into account. This is different when the input is continuous, as there exist infinite values. For that reason, we need to describe the distribution of the input (e.g mean, median…)

**Gaussian Mixtures:**

review or something

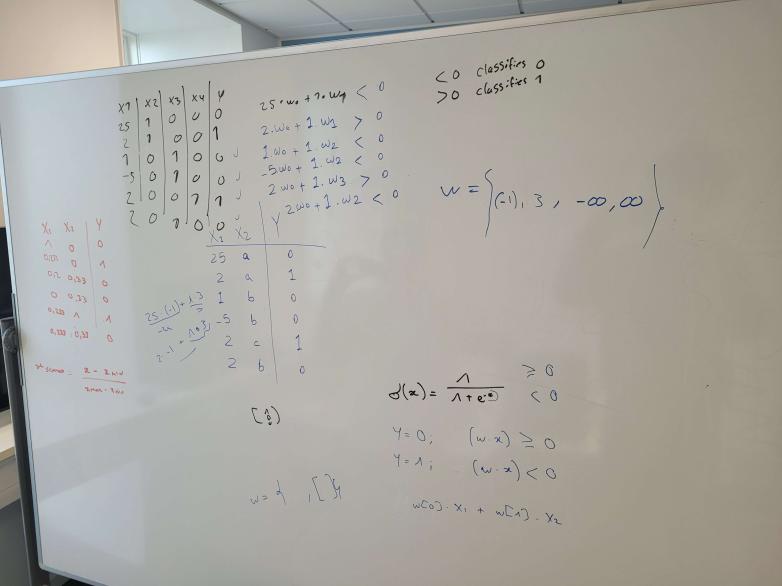
**Logistic Regression:**

**EXERCISE 1:**



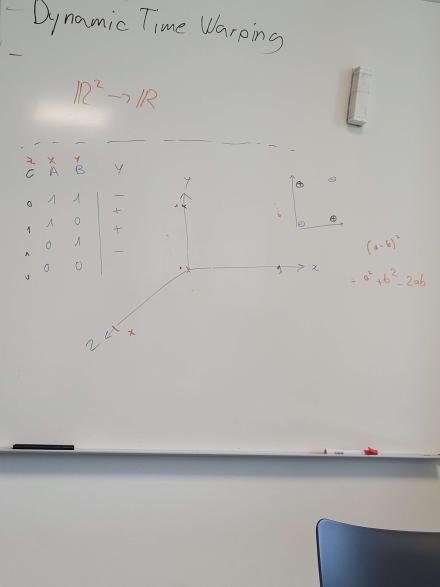
LDA is not great with outliers, since covariances have to be the same. In this example the blue points are all clustered together with a few outliers. These outliers are close to the other cluster but can still be linearly separated.

**EXERCISE 2:**

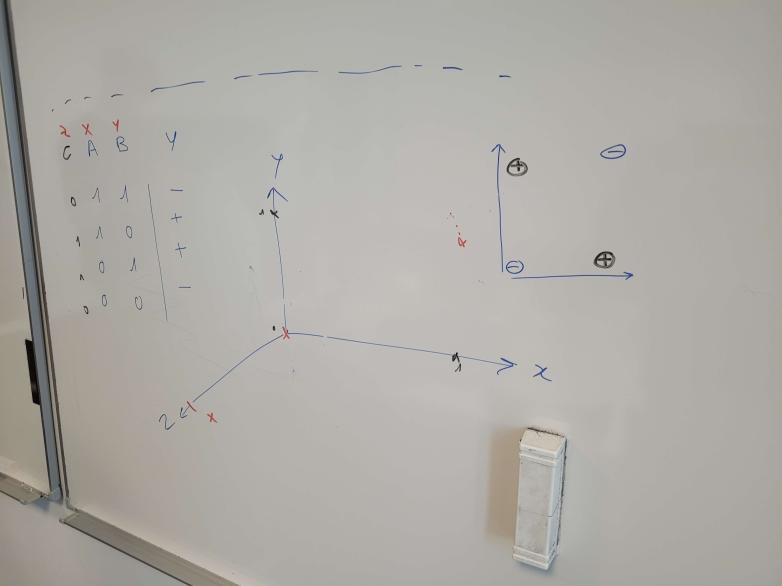


One hot encoding to go from 2 dimensions to 4. We have 4 weights, w3 and w4 can be infinity to segregate the classes. Then w1 and w2 are decided accordingly for the two first rows

**Exercise 3:**



We go from R2 to R by finding a function, such that: f(x,y) = (a-b)^2



We simply add a new dimension to separate the points through the new dimension C.