Programming Assignment 4

Write a program (for Python 3.6.8) that implements a 2-class Naive Bayes algorithm with an apriori decision rule using a *multinomial* estimation for the classes and a gaussian estimation for the attributes. The formulas to be used are therefore[1]:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} = \frac{P(X|C) \cdot P(C)}{\sum_{C'} P(X, C')} = \frac{P(X|C) \cdot P(C)}{\sum_{C'} P(X|C') \cdot P(C')}$$

$$P(c_i) = p_i$$

$$P(x_a|c_i) = \mathcal{N}(x_a|\mu_{a,c_i}, \sigma_{a,c_i}) := \frac{1}{\sqrt{2\pi\sigma^2_{a,c_i}}} exp\left\{ - \frac{(x_a - \mu_{a,c_i})^2}{2\sigma^2_{a,c_i}} \right\}$$

where x_a is an instance x with an attribute a and μ and σ being the parameters of the Gaussian. The parameter estimates are given as follows:

$$p_i = \frac{n_{c_i}}{\sum_i^c n_{c_i}}$$

$$\hat{\mu}_{a,c_i} = \frac{1}{n_{c_i}} \cdot \sum_{k=1}^{n_{c_i}} x_{k,a}$$

$$\hat{\sigma}^2_{a,c_i} = \frac{1}{n_{c_i} - 1} \cdot \sum_{k=1}^{n_{c_i}} (x_{k,a} - \hat{\mu}_{a,c_i})^2$$

where n_{c_i} is the amount of instances for class c_i.
Given are the two data sets on Moodle named *Example* and *Gauss2* as csv files. Your program should be able to read both data sets and treat the *first* value of each line as the class (A or B). The output of your algorithm *on the console* should be comma separated values per data set, which contains a row for each class:

$$\hat{\mu}_{1,c} \quad \hat{\sigma}^2_{1,c} \quad \hat{\mu}_{2,c} \quad \hat{\sigma}^2_{2,c} \quad \hat{p}_c$$

The last (third) row contains the absolute number of misclassifications for the data. Any other information should **not** be inside the console output, only the requested values. You can check the solution for the *Example* data set in order to compare it to your output file. For each data set, you can acquire one point, if the solution of your program returns correct results. If the program fails, the data format is incorrect or I have to change source code, in order to make it work, you will get zero points. Machine learning libraries are not allowed. You can use numpy 1.12.1 again.
Your program must accept the following parameters:

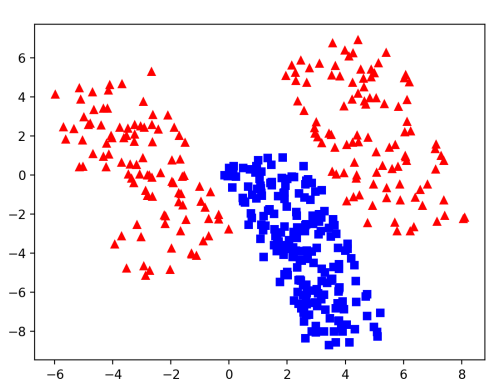
1. **data** - The location of the data file (e.g. /media/data/Example.tsv).

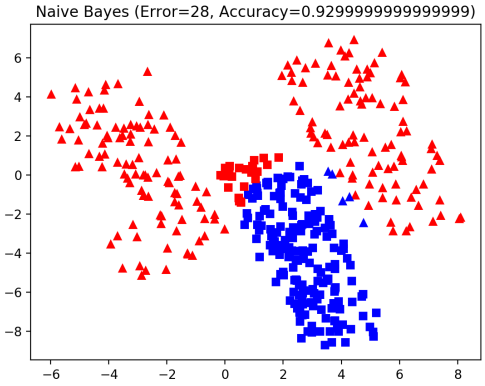
I.e. your program will be executed as follows:

[1] $P(c_i)$ is a simplification of $P(c) = \mathcal{M}(n_1, n_2|p_1, p_2) := \binom{n_1+n_2}{n_1,n_2} \cdot p_1^{n_1} \cdot p_2^{n_2}$

```
python3 student.py --data Example.csv
```

The final program code must be uploaded to Moodle until Wednesday, the 13th of January 2021, 1am. The figures below shows the data for the *Example* set and its Naive Bayes solution.





2 points