Programming Assignment 4

Write a program 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[2] named *Example* and *Gauss* as tsv (tabular separated values) files from the last assignment. 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 should be a *single* tsv file 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 output file, 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 libraries for handling the CSV/TSV format and the input parameters.
Your program must accept the following parameters:

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

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

[2] http://wwwiti.cs.uni-magdeburg.de/iti_dke/Lehre/Materialien/WS2018_2019/ML/res/nb.zip

2. **output** - Where the output tsv should be written to.

Please prepare example statements on how to use your program. E.g. for a python program:

```
python3 nb.py --data Example.tsv --output Example_NB_Solution.tsv
```

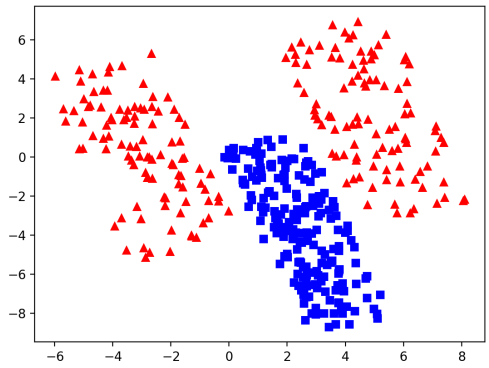
The final program code must be sent via email until Sunday, 6th of January 2019, 23:59 to marcus.thiel@ovgu.de. Please format your e-mail header as follows:

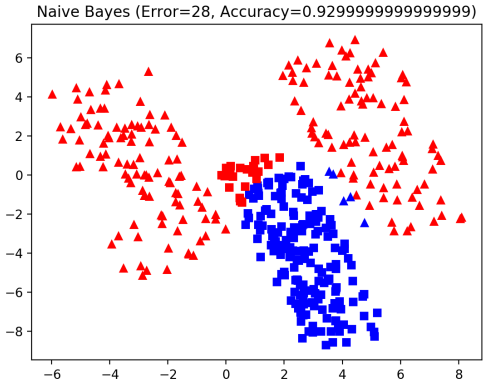
[Exercise Group] ML Programming Assignment 4

Replace *Exercise Group* with the day and time of your exercise group. E.g for Monday from 13:00 to 15:00 it would be:

[Monday 13-15] ML Programming Assignment 4

Please also be prepared to present your solution shortly in front of the class.
The figures below shows the data for the *Example* set and its Naive Bayes solution.





2 points