COMP47750
Machine Learning with Python

2021 Assignment 1

Gaussian Naive Bayes

Objective

The objective of this assignment is to implement a Gaussian Naive Bayes classifier in the scikit-learn framework. A notebook (**`MajorityClassClf`**) is provided with a simple example of a classifier that works with scikit-learn.

Note: The code developed in this assignment will be extended in the second assignment to allow for missing values.

Requirements

The notebook **`MajorityClassClf`** contains some basic code to help you get started.

1. Provide a python class MyGaussianNB that implements Gaussian Naive Bayes. The conditional probabilities should be calculated as follows:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

   where μ_y is the mean for variable *i* for class *y and* σ_y is the corresponding standard deviation. Thereafter the classification should use the NB formulae presented in the lectures. Alternatives that use addition of conditional probabilities or logs should **not** be used.

   The API specification for sklearn classifiers is here:
   https://scikit-learn.org/stable/developers/develop.html
   You should implement the **`'fit'`** and **`'predict'`** methods, there is no need to implement '**`predict_proba`**'.

   Prior probabilities should be calculated from the training data. With this, there will be no need to pass parameters when instances are created.

2. Test the performance of your implementation against the **`GaussianNB`** implementation in scikit-learn. You should use a range of datasets for this testing. Possible test sets used in lectures are penguins, diabetes and glassV2.

Submission: This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise your notebook and the second dataset

that you use. Clear all outputs in the notebook before saving for submission. You can use markdown cells in the notebook to report your findings and conclusions.