

Information Theory and Data Science

Assignment no. 4

Mauro De Sanctis

General Recommendations

The project should be carried out using Python. Good programming discipline should be followed when writing the Python code. This means that the variable names should be logical, the code must be commented and it should be written in such a way that it is easy to follow and understand.

Project

- Build a Bayes classifier function which takes a training dataset, a class label vector for the training dataset and a test dataset and returns a class label vector for the test dataset. Assume that features are continuous random variables and estimate the multivariate probability density function by using a multivariate kernel density estimator.
- Build a naïve Bayes classifier function by assuming that features are continuous and independent random variables. Use a univariate kernel density estimator to estimate the probability density function of each feature.
- Build a naïve Bayes classifier function by assuming that features are continuous, independent and Gaussian distributed random variables with mean μ_j and variance σ_j^2 .
- Compute and compare the average accuracy of the previous classifiers by applying it to the Iris dataset taking for each class label 50% of the rows as training dataset and the remaining 50% of the rows as test dataset.

Hint: Start from the example in `test_pdf_multivariate_iris.py`