Lab 4: Classification

Prof. Lorenzo Pellis
(Original lab material from Prof. T. House)

20 Feb 2024 – 13:00-14:30 or 14:30-16:00

Supporting materials

The following materials are available on Blackboard:

- The dataset `pokemon.csv` used in lecture.
- Two data sets, `RogersGirolami_GaussianClassConditional.csv` and another derived from the `20 Newsgroups` collection of texts, that are used to illustrate probabilistic classification.
- A Python notebook, `poke.ipynb`, that illustrates the use of Scikit-learn to classify the Pokémon data and another notebook, `ProbabilisticClassification.ipynb`, that illustrates two approaches to probabilistic classification.
- Two datasets, `vertebral_column_data.txt` and `vertebral_column_metadata.txt`, that list biomechanical attributes together with whether a patient was classified as normal or abnormal. You will need these for the coursework.

COURSEWORK: Classification

Using the vertebral column data, apply at least one unsupervised and one supervised classification procedure. Produce a short report, which should contain the following sections:

1. A description of the unsupervised clustering method, using your own words, including equations and citations as appropriate. **[3]**
2. A description of the supervised classification method, using your own words, including equations and citations as appropriate. **[3]**
3. Exploratory analysis of the data and any processing / transformations performed on the basis of this. **[3]**
4. Results of the analyses, including appropriate figures and tables to support the conclusions, and a discussion of how the supervised and unsupervised analyses inform each other. **[8]**
5. R or Python code used to produce the analysis. Note that it is expected that you will use packages such as *scikit-learn* rather than code from scratch. **[3]**

This gives a total of **20** marks.

Due Date: 3:00pm on Mon 11 March 2024, uploaded to BlackBoard as a PDF. Length: 1000 words.