Project III: Statistical Classification (Ric Crossman)

Topic

There are many real-world situations in which we might encounter some variation on the following problem. We want to know what category an object belongs to, but we cannot determine this directly. Instead, we need to consider observed characteristics the object possesses, and use those characteristics to decide what category the object is most likely to belong to. An (extremely) incomplete list of such situations is given below.

- Medicine, in which disease types might be the categories, and symptoms might be the observed characteristics.
- Economics, in which general situations for next year's economy ("boom", "recession", "depression", etc.) might be the categories, and economic properties/measures of this year's situation might be the observed characteristics.
- Finance, in which credit ratings might be the categories, and age, familial and residential status, previous credit history etc. might be the observed characteristics.
- Ornithology, in which bird species might be the categories, and properties of the bird (colour, size, song, location found, etc.) might be the observed characteristics.

Having taken Statistical Modelling II (a pre-requisite for this project), you are aware of how linear regression can be used to provide an expected value of a continuous response variable, given values of either continuous or categorial variables. **Classification** focuses instead on providing what is, in some defined sense, the "most likely" value of a categorial variable. This one apparently small tweak requires a total change of approach.

In this project we will begin by looking at the simplest form of predictor, an extension of linear regression to the case of a binary response variable. This approach is referred to as **logistic regression**, and it allows us to estimate the probability of an object belonging to each category.

From this initial starting point (which will also be covered in Advanced Statistical Modelling III, and therefore make up only a small part of the overall project), there are three general routes for project progression, each of which have a great deal of flexibility within them.

1. **Generalised logit model regression** – an extension of logistic regression which allows probability distributions for the category of an object to be generated, given the object's observed characteristics. This is done by comparing odds ratios of each category to a single reference category. One possible extension to this topic would be considering the issue of **overdispersion**.

2. **Classification trees** – an alternative to the regression approach, in which directed acyclic graphs are mathematically constructed to give us what one might think of as flow charts. Each non-leaf node represents an observed characteristic, each edge from such a node is labelled with one or more values that observed characteristic can take, and the leaf nodes represent categories. Such a graph can therefore be used to label an object with a category. There are many different ways to build such trees, depending on the choice of **splitting rule**, which determines which observed category is the most valuable to assign to each node. One possible extension to this topic would be **random forests**.

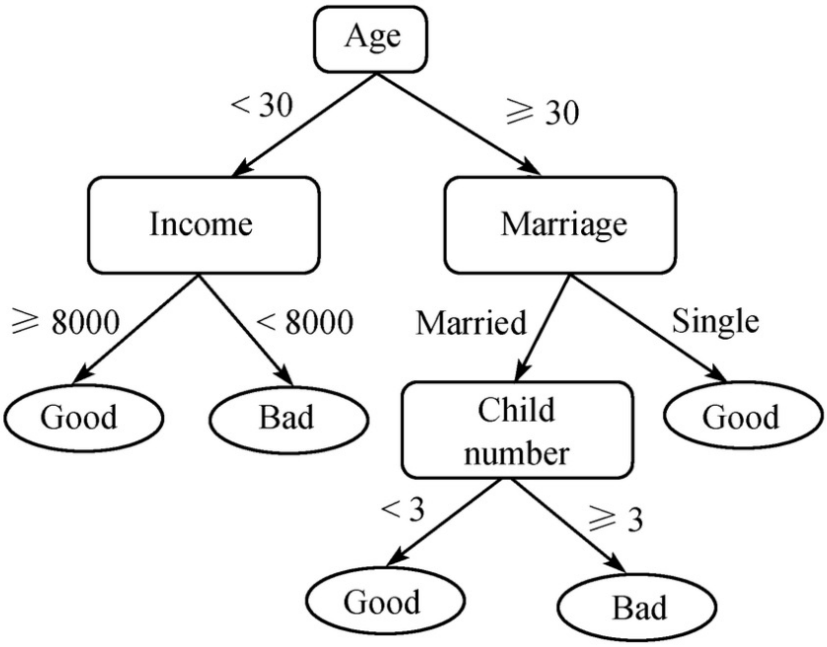

Figure 1: Example of a decision tree for credit scoring (taken from this paper)

3. **Naïve Bayes Classifiers** – A set of approaches to classification which commonly use Maximum Likelihood Estimation to determine the most likely category of an object, under the assumption that, given the actual value of the category, the probability distributions of each observable characteristic are independent of each other. This assumption simplifies the process of creating the **likelihood function**, which – as one would expect from a Bayesian process – is then combined with a **prior distribution** for the categories in order to produce a **posterior distribution** for the categories given the observed characteristics. One possible extension to this topic would be **Non-naïve Bayes Classifiers**.

Other forms of classifier can also be touched upon where appropriate. Toy data sets will be available to allow small-scale applications of these approaches in practice. Students wishing to apply classification approaches to larger data sets would be encouraged to seek out such a data set themselves early in the project.

An alternative extension could be to investigate the many ways in which the performance of a classifier can be evaluated. Do you just consider the proportion of correct classifications, for instance? Or are some errors more costly than others?

Prerequisites

Statistical Inference II, Statistical Modelling II

Web Resources

- The Elements of Statistical Learning, Hastie, T; Tibshirani, R; Friedman, J; 2nd edition 2009, Springer.
- An Introduction to Statistical Learning, James, G; Witten, D; Hastie, T; Tibshirani, R; 2nd edition 2021, Springer.

Further Information

- Statistical Regression and Classification - From Linear Models to Machine Learning Matloff, N; 2017, CRC Press.
- Naive Bayes classifiers, Murphy, K; 2006, Instituto de Computação, Universidade Estadual de Campinas.