3.1 Classification: introduction

Classification for campaign optimisation:

Campaign to promote new vehicle: large list of prospects, invitations for test-drive, gifts and free phone line for enquiries/reservations.

- Goals: reduce costs and maximise returns;
- Stategy: analyse response to previous campaigns, build customer relating characteristics and response, apply model to prospects and invite prospects selected by the model.

Data for classification:

- Prospects (customers that didn't buy a car in the last 4 years);
- Results from previous campaigns (customers who were contacted and their response).

Model: classification tree (or decision tree):

- Root contains all cases;
- Each node is called a test, which divides into decisions (leaves) or other tests using an attribute;
- The distribution of classes is shown on the leaves.
- Steps:
    1. Set of labelled examples;
    2. If all the examples are of the same class, stop;
    3. Otherwise, divide (split) in the root, each test being of [variable type = value] or [variable > value];
    4. Create two descendant nodes according to the test selected;
    5. Repeat process in each of these descendant nodes.
- Split:
    - Entropy of a variable can be used to decide a good split;
    - The bigger the decrease in diversity, the better.
- However, decreasing the training error does not mean the model is better on other cases: an overfitted model generalizes poorly.

Classifier Evaluation

Confusion Matrix

- prediction vs. reality

- number of right answers on the main diagonal
- sum of the array is the total number of examples
- **Error rate**: percentage/proportion of cases where the model misses

Measures Computed

- False Positive/Negative Rate (FPR/FNR);
- True Positive Rate (TPR), also known as Recall or Sensitivity;
- True Negative Rate (TNR), also known as Specificity;
- Positive predictive value (PPV) also known as precision;
- Negative predictive value (NPV);
- Accuracy (proportion of correct predictions);
- F1-measure.
- **Baseline**: simplest model that can be obtained from the data (e.g. most common value).
- **Split validation**: using different operations for training data and for test data; model obtained on the train side is passed to the test side.

Classification for Scoring

Binary Classification

- Direct application of the model is not suitable for all problems (e.g. list of 1000 prospects but send only to the 200 with the highest probability);
- Scoring can use estimated probability;

Where to cut?

- Arbitrary threshold (by default, Prob(X) > 0.5);
- If it is important to find all cases of X, maybe reduce the threshold (Prob(X) > 0.3).

Evaluate Scoring Models

- Sort prediction by increasing order of belonging to positive class
- ROC (Receiver Operating Characteristic) analysis:
    - visualize proportion TP vs. FP to find best compromise

[< Go back]