Complement on Classification

This short segment addresses the assessment of the quality of classification methods.

Given the large number of methods available, one is often faced with the following situation in practice:

- ✓ the data set was split into a training and a validation part;
- ✓ M different methods were applied to the training data set;
- ✓ the validation data are to be used to select the best method.

Tools are then needed to measure the performance of different methods.

A Simple Measure

As was seen in a previous segment, the first step consists of comparing the success rate of each method on the validation data.

In some cases, it is also possible to assign a profit to each type of proper classification and a cost to each type of incorrect classification.

One can then choose the method that maximizes the estimate difference

$$\text{profits} - \text{costs}$$

on the validation data.

One can also resort to the "F_1 score" if desired, viz.

$$F_1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} .$$

Example (1–2)

Method A			Method B		
	C1	C2		C1	C2
C1	**25**	10	C1	**14**	21
C2	5	**20**	C2	2	**23**

Correct classification rates for Method A and Method B:

$$\frac{25+20}{25+20+5+15} = 45/60, \quad \frac{14+23}{14+23+2+21} = 37/60$$

F_1 score for Method A and Method B:

$$\frac{2\times 25}{2\times 25+5+10} = 0.760, \quad \frac{2\times 14}{2\times 14+2+21} = 0.683$$

Method A is then preferable to Method B.

Example (2–2)

Method A			Method B		
	C1	C2		C1	C2
C1	**25**	10	C1	**14**	21
C2	5	**20**	C2	2	**23**

Suppose that a profit of \$1 is assigned for a properly classified C1 or C2, a cost of \$1 for a poorly classified C1 but \$10 for a poorly classified C2.

Cost for Method A:

$$25 \times \$1 - 10 \times \$1 - 5 \times \$10 + 20 \times \$1 = -\$15$$

Cost for Method B:

$$14 \times \$1 - 21 \times \$1 - 2 \times \$10 + 23 \times \$1 = -\$4$$

Method B is then preferable to Method A.

Relying on Probability Estimates

More informative criteria exploit the fact that classification methods often give a probability of belonging to a given class, not just a prediction.

A better measure of the performance of such methods can be devised by checking whether these probabilities are realistic.

Two popular tools used in finance are based on this notion, namely

- ✓ lifting,
- ✓ gain.

Both of them are described below in the special case of binary classification.

Lift Curve (1–3)

Suppose that one is interested in predicting items belonging to Group 2.

Let $n_{\text{valid}}^{(2)}$ be the number of items in Group 2 in the validation dataset.

For any fixed $p \in (0, 1)$,

- ✓ select at random $p \times 100\%$ of the items in the validation dataset and let N_p^r be the number of items from Group 2 therein;
- ✓ find the $p \times 100\%$ of the items in the validation dataset to which the proposed classification method assigns the highest probability of belonging to Group 2;
- ✓ denote by N_p^c the latter number of of items.

Lift Curve (2–3)

The lift ratio for this choice of $p \in (0, 1)$ is then given by

$$\ell_p = \frac{N_p^c}{\mathsf{E}(N_p^r)} = \frac{N_p^c}{p \times n_{\text{valid}}^{(2)}}.$$

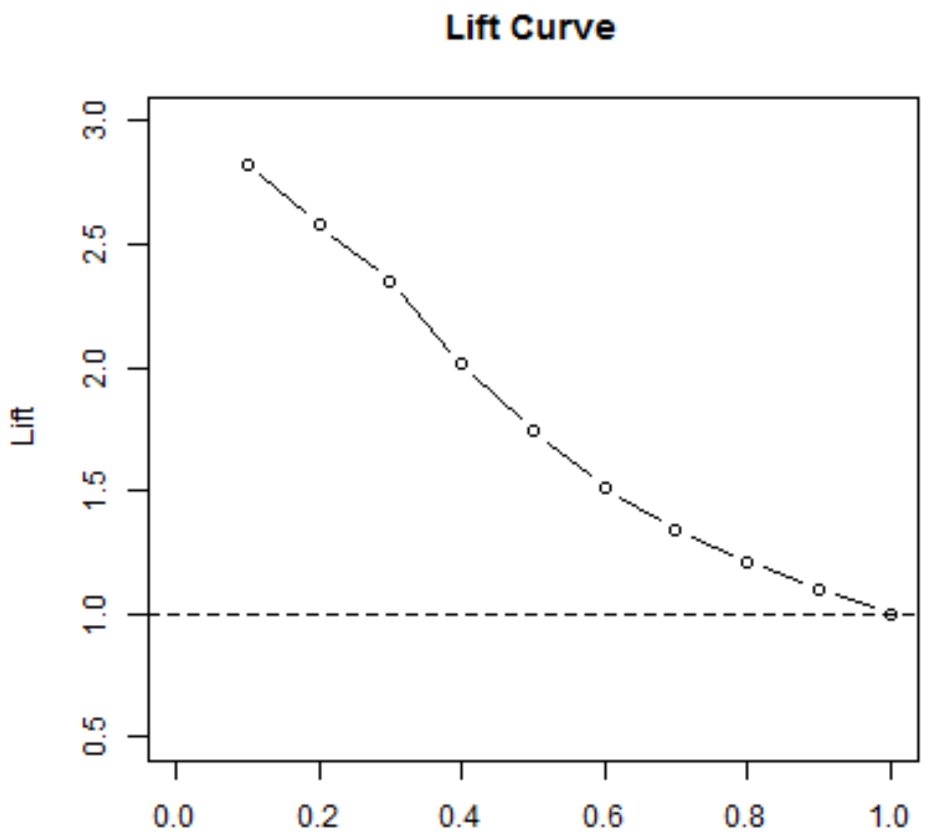
Repeat the procedure for a large number of values of $p \in (0, 1)$, and plot the curve

$$p \mapsto \ell_p.$$

Typically, this curve will be decreasing and such that $\ell_1 = 1$.

The lift curve is a variation on the receiver operating characteristic (ROC) curve, and is also known in econometrics as the Lorenz or power curve.

Lift Curve (3–3)



Cumulative Gain Curve (1–2)

For every fixed $p \in (0, 1)$, compute

$$g_p = N_p^c / n_{\text{valid}}^{(2)}$$

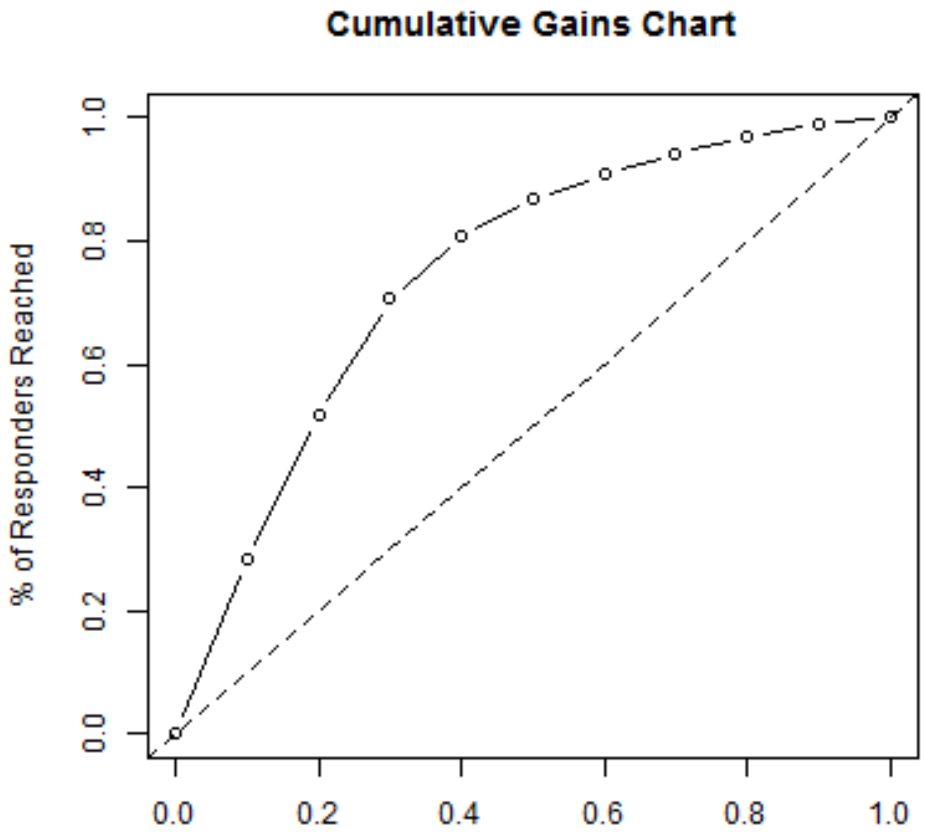
and plot the curve

$$p \mapsto g_p.$$

Typically, the curve should joint $(0, 0)$ and $(1, 1)$ and stand above the line $p \mapsto p$ if the classification method is any good.

Note that $\ell_p = g_p / p$ because $\mathrm{E}(N_p^r) / n_{\text{valid}}^{(2)} = p$.

Cumulative Gain Curve (2–2)



ROC Curve (1–3)

The Receiver Operating Characteristic (ROC) curve is also related to the cumulative gains.

1. Let $\pi_i^{(2)}$ be the probability that observation i belongs to Group 2 according to the model.
2. Choose $u \in [0, 1]$ and predict that observation i belongs to Group 2 if $\pi_i^{(2)} \geq u$ and to Group 1 otherwise.
3. Repeat Step 2 for all $i \in \{1, \ldots, n\}$ and compute the resulting sensitivity $\text{SEN}(u)$ and the specificity $\text{SPE}(u)$.
4. Repeat Steps 2 and 3 for all $u \in [0, 1]$, plot the pairs $(\text{SEN}(u), \text{SPE}(u))$ on a graph and connect the points.

ROC Curve (2–3)

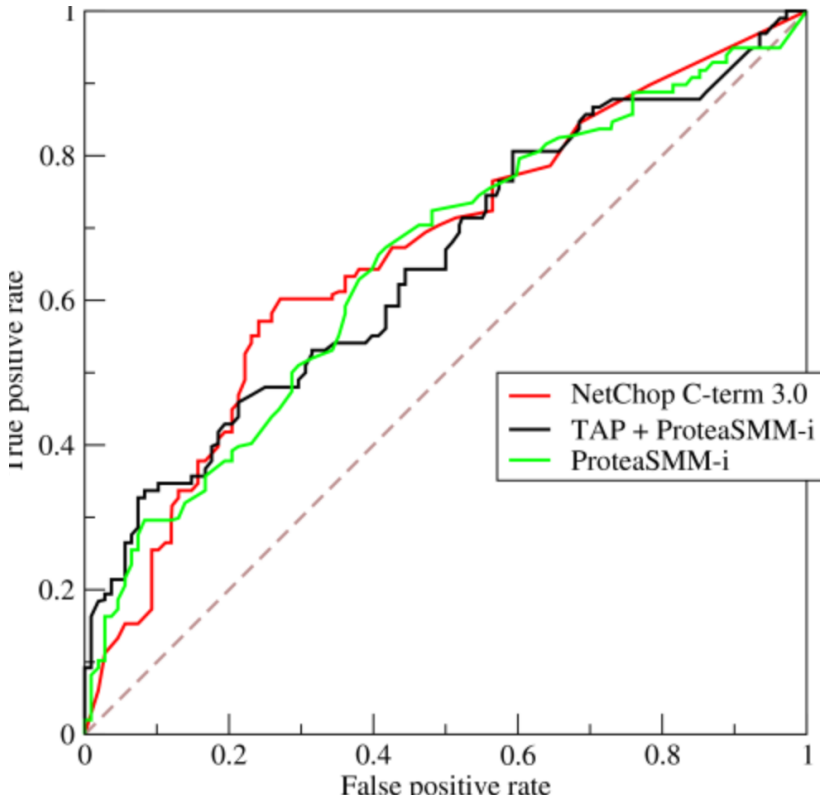
The curve resulting from point 4 above is the ROC curve. It starts at $(0, 0)$ and ends at $(1, 1)$.

When points are classified randomly, one gets a 45-degree line.

A measure of the predictive power is given by the area under the ROC curve (AUC, Area Under the Curve).

An AUC of 0.5 is what one can expect when classification is done randomly. The closer AUC to 1, the better.

ROC Curve (3–3)



Typical ROC curves

Decision Trees

Decision trees are an important type of algorithm for predictive modeling.

Classical decision tree algorithms have been around for decades; modern variations like random forest are among the most powerful techniques available.

This segment will focus on the classical supervised learning technique known as CART, which stands for Classification And Regression Trees.

Given: Observations $(X_1, \ldots, X_p) \in \mathbb{R}^p$ are available.

A classification tree tries to assign an observation to one of K groups on the basis of $(X_1, \ldots, X_p)$. In contrast, a regression tree tries to predict the value of a continuous response variable Y from $(X_1, \ldots, X_p)$.

General Strategy

In discriminant analysis, a score $f(x_1, \ldots, x_p) \in \mathbb{R}$ was computed for each observation. This score is used to determine to which group an individual belongs. This induces a partition of $\mathbb{R}^p$.

In the construction of classification or regression trees, the space $\mathbb{R}^p$ is also partitioned but this is done through successive binary divisions as a function of one of the variables.

The choice of variable used for the division is determined by some criterion. It can vary from iteration to iteration.

When the algorithm stops, all observations in the same "leaf" or "terminal node" get the same prediction, which is the most frequent class among the observations within that node.

Example (1–3)

As an illustration, consider the Heart Disease Data Set which is available at UC Irvine's Machine Learning Repository. See

`https://archive.ics.uci.edu/ml/datasets/heart+Disease`

Data are available for 303 patients with chest pain.

For each patient, the data at hand consist of

- ✓ a variable Y indicating the presence or absence of a cardiac problem;
- ✓ 13 covariates such as age, gender, cholesterol level

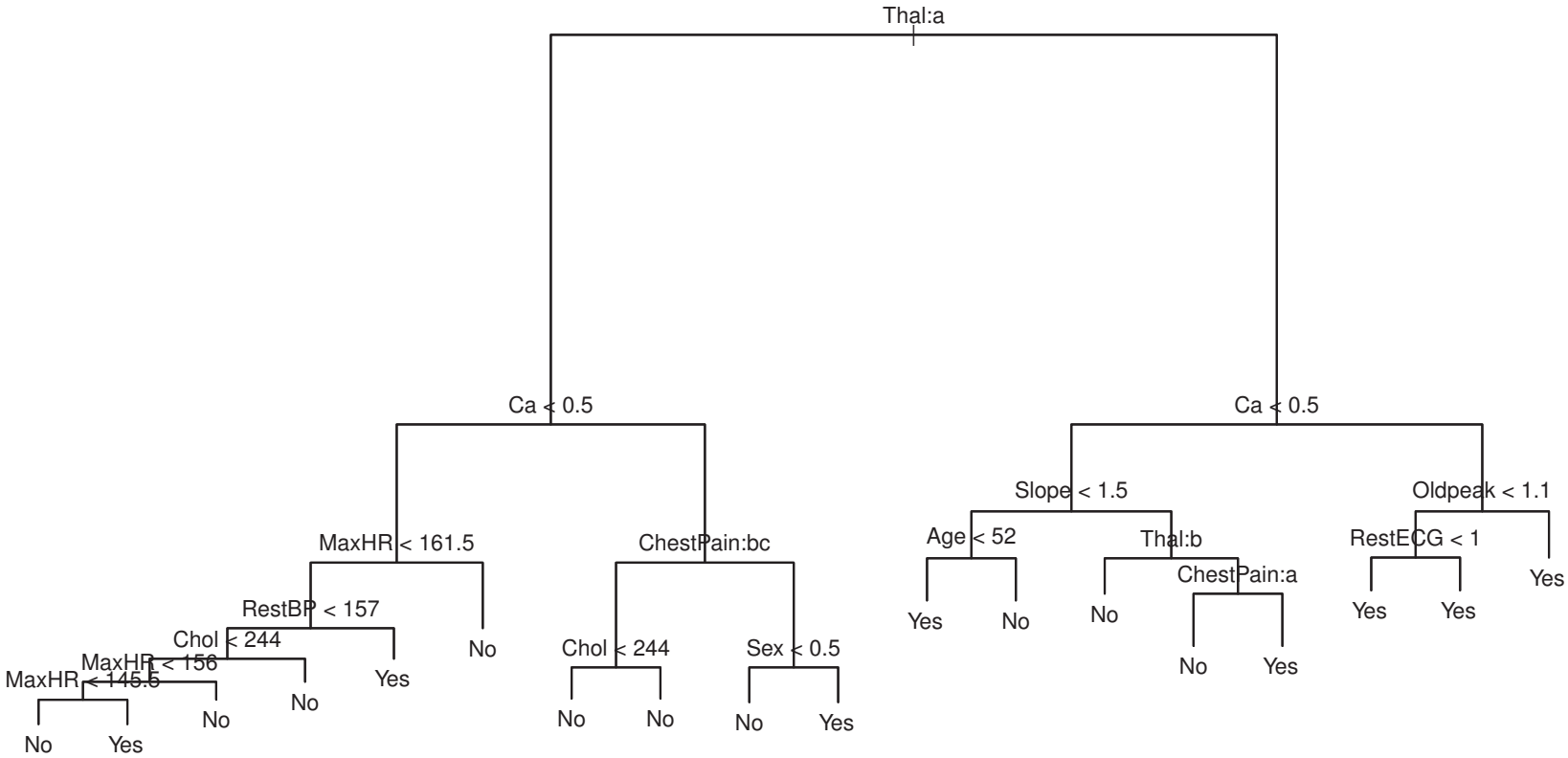
Example (2–3)

List of covariates:

1. Age (in years)
2. Sex (1 = male)
3. CP: chest pain type (4 values)
4. trestbps: resting blood pressure (in mm Hg on admission)
5. chol: serum cholestorol in mg/dl
6. fbs: fasting blood sugar > 120 mg/dl (1 = true)
7. restecg: resting electrocardiographic results (3 values)
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes)
10. oldpeak: ST depression induced by exercise relative to rest
11. slope: slope of the peak exercise ST segment (1 = up; 2 = flat; 3 = downsloping)
12. ca: number of major vessels (0–3) colored by flourosopy
13. thal: Thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

The tree construction algorithm, available through R with the command `rpart`, is time consuming because at each step, one must select the best possible binary division. Yet the resulting tree may not be optimal.

Example (3–3)


Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Slope < 1.5
Oldpeak < 1.1
RestBP < 157
No
Chol < 244
Age < 52
Thal:b
RestECG < 1
Yes
MaxHR < 156
Yes
ChestPain:a
MaxHR < 145.5
No
No
Chol < 244
Sex < 0.5
Yes
No
No
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
Yes
Yes

Division Criteria

Denote by $\hat{p}_{mk}$ the proportion of observations from the mth partition set in $\mathbb{R}^p$ associated with the kth class.

Three criteria are typically used to determine the optimal binary division at each step of the construction, viz.

1. Classification error rate, viz. $E = 1 - \max_k(\hat{p}_{mk})$.
2. Gini's index $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$.
3. Cross Entropy, viz. $D = -\sum_{k=1}^{K} \hat{p}_{mk} \ln(\hat{p}_{mk})$.

Cross entropy is called information in R.

Information Gain

The information gain is what we try to maximize when we look for a way of splitting a node in two parts.

For example with Gini's index, the information gain is given by

$$\text{Gini's index before division} - \sum_{j=1}^{2} \frac{n_j}{n_1 + n_2} G_j,$$

where $j \in \{1, 2\}$ denotes the two segments created by the division,

n_j is the number of observations in segment j, and

G_j is Gini's index for segment j.

For entropy, simply replace G_j by D_j.

Numerical Example (1–2)

Consider the following artificial dataset:

	Student							
Question	1	2	3	4	5	6	7	8
Hours Studied (X_1)	10	4	5	20	0	9	12	7
Problems Solved (X_2)	30	25	5	10	0	35	40	22
Verdict	S	F	F	S	F	F	S	F

(a) Compute Gini's index.

(b) Suppose the first division is based on $X_1 < 8$. Compute Gini's index for both segments.

(c) What is the resulting information gain?

Numerical Example (2–2)

	Student							
Question	1	2	3	4	5	6	7	8
Hours Studied (X_1)	10	4	5	20	0	9	12	7
Problems Solved (X_2)	30	25	5	10	0	35	40	22
Verdict	S	F	F	S	F	F	S	F

If $k = 1$ for fail, one has $\hat{p}_1 = 5/8$, $\hat{p}_2 = 3/8$ and hence

$$G = \frac{5}{8}\left(1 - \frac{5}{8}\right) + \frac{3}{8}\left(1 - \frac{3}{8}\right) \approx 0.46875,$$

$$G_{X_1<8} = \frac{4}{4}\left(1 - \frac{4}{4}\right) + \frac{0}{4}\left(1 - \frac{0}{4}\right) = 0, \quad G_{X_1\geq 8} = \frac{1}{4}\left(1 - \frac{1}{4}\right) + \frac{3}{4}\left(1 - \frac{3}{4}\right) = 3/8,$$

with $n_{X_1<8} = 4$, $n_{X_1\geq 8} = 4$ and hence the gain is

$$0.4687 - \left(\frac{4}{4+4} \times 0 + \frac{4}{4+4} \times 3/8\right) \approx 0.28125.$$