Introduction to Classification

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- At the end of the last class, we touched on the idea of classification.
- In this "lecture", we will look at some basics.
- Then we will move to classification and regression trees (CART).
- We will also look at bagging, boosting, and random forests.

Classification

- Suppose we observe p-dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that $k < n$ of them are from one of G known classes, i.e., $k < n$ are labelled.

- ... the estimation of labels for the $n - k$ unlabelled observations is a classification problem.

- Suppose we observe p-dimensional data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that each one of them is from one of G known classes, i.e., all n are labelled.

- ... the construction of a classification (discriminant) rule for future use is a classification problem.

Discriminant Analysis

- In common parlance, the word "discrimination" has extremely negative connotations.

- Discrimination, in the way we shall use the word, simply refers to the act of discriminating observations of one type from those of another (or others).

- These types are most often referred to as classes.

- Discriminant analysis is a classification technique, also known as a form of "supervised classification" or "supervised learning".

Discriminant Analysis contd.

- Again, suppose there are G classes.
- In general, discriminant analysis assumes that

$$\begin{aligned} p(\mathbf{x} \mid \text{class } g) &= \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \\ &= (2\pi)^{-p/2}|\boldsymbol{\Sigma}_g|^{-p/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right\}. \end{aligned}$$

- Let π_g be the (*a priori*) probability that an observation is from class g.
- Then,

$$\mathbb{P}[\text{class } g \mid \mathbf{x}] = \frac{\pi_g p(\mathbf{x} \mid \text{class } g)}{\sum_{h=1}^{G} \pi_h p(\mathbf{x} \mid \text{class } h)} = \frac{\pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\sum_{h=1}^{G} \pi_h \phi(\mathbf{x} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}.$$

Linear versus Quadratic

- Linear discriminant analysis supposes that $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ for $g = 1, \ldots, G$.
- Quadratic discriminant analysis makes no such supposition.
- Consider two classes, g and h. An observation $\mathbf{x}$ is assigned to class g rather than class h if

$$\mathbb{P}[\text{class } g \mid \mathbf{x}] > \mathbb{P}[\text{class } h \mid \mathbf{x}]$$

i.e., if

$$\pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) > \pi_h \phi(\mathbf{x} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h).$$

- Expanding from here, we can see where the term “quadratic” comes from.

Some Terminology

- Classification is often carried out using a training set and a test set (and sometimes a validation set).

- The **training set** contains labelled observations and is used to build ("train") the model.

- The observations in the **test set** are unlabelled or a treated as such; the test set is used to assess the efficacy of the model.

- Doing the training/test split in a **stratified** way means that the **proportion of each class** in the test set is (or very almost is) the same as in the training set.

Comments

- Let's look at some examples of LDA and QDA in R.

- Then, we will look at k-nearest neighbours classification.

- Then a word or two about comparing partitions.

- And then onto CART, etc.

k-Nearest Neighbours

- A very simple "statistical learning" or "machine learning" approach.
- An unlabelled observation is classified based on the labels of the k closest labelled points.
- Specifically, an unlabelled observation is assigned to the class that has the most labelled observations in its neighbourhood (which is of size k).
- In the training/test lingo, the labelled points are in the training set — but this lingo might not be considered strictly correct here.

Choosing k

- Important: k is not the number of classes but rather the size of the neighbourhood.
- The values of k are chosen based on the labelled points.
- In practice, this might mean choosing k based on the training set.
- Choosing k means running the kNN algorithm for different values of k.

Choosing k contd.

- The value of k that gives the best classification rate (on the labelled points/training set) is then chosen.
- If different k give similar classification rates, the smaller value of k is usually chosen.
- Then, using the chosen value of k, the unlabelled points can be classified.
- Let's look at some examples in R.

Comparing Partitions

- At the end of the last class, we briefly considered comparing partitions.
- Consider these classification tables (true classes are numbers and predicted classes are letters).

	A	B	C
Class 1	48	42	10
Class 2	0	0	100

	A	B	C
Class 1	48	10	42
Class 2	0	0	100

- Such tables can occur, e.g, in unsupervised classification.

Comparing Partitions II

- Which table depicts a better classifier?
- Both have a misclassification rate of 26%, but that does not tell the whole story.
- Consider pairwise agreements, i.e., pairs of observations that should be together (in the same class) and are plus pairs of observations that should be apart (in different classes) and are.
- We can think about a table of pairwise agreements and disagreements (rows can be thought of as true and columns as predicted classes, but it does not matter — it is just one partition versus another):

	Same group	Different groups
Same group	A	B
Different groups	C	D

STATS 780/CSE 780 Prof. Sharon McNicholas

Comparing Partitions III

- The Rand index (Rand, 1971) is the ratio of the pairwise agreements to the total number of pairs or, using the notion in the table on the last slide
$$\text{RI} = \frac{A+D}{N}, \tag{1}$$
where $N = A + B + C + D$ is the total number of pairs.
- Clearly, $\text{RI} = 1$ corresponds to perfect class agreement.
- Most other values (especially smaller values) of RI are difficult to interpret because chance agreement will tend to inflate the value in (1).
- So an adjustment is needed...

Comparing Partitions IV

- The adjusted Rand index (ARI; Hubert and Arabie, 1985) corrects the Rand index for agreement by chance and is given by

$$\text{ARI} = \frac{N(A+D) - [(A+B)(A+C) + (C+D)(B+D)]}{N^2 - [(A+B)(A+C) + (C+D)(B+D)]}, \qquad (2)$$

  where, again, $N = A + B + C + D$ is the total number of pairs.

- The ARI, (2), has expected value 0 under random classification and a value of 1 for perfect class agreement. This is apparent from the general form of the correction, i.e.,

$$\text{corrected index} = \frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}.$$

Comparing Partitions V

- Negative values of the ARI are also possible and can be interpreted as classifications that are worse than would be expected under random classification.

- Note that the ARI has no well-defined (general) lower bound; Hubert and Arabie (1985) comment that "the required normalization would offer no practical benefits".

- The ARI can be computed using the `classAgreement()` function of the `e1071` library in R.

- The material on comparing partitions is based on McNicholas (2016, Section 1.4)[a]

[a] McNicholas, P.D. (2016), *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.