Topic 1: Concepts of classification

Welcome to Week 5

Dr Pavel Krivitsky gives you a brief overview of topics and concepts we'll be covering in this week.

Transcript

Weekly learning outcomes

- Define, calculate for a classifier, and interpret the basic measures of classification, including confusion matrices, true/false positive/negatives, expected cost of misclassification, and related concepts.
- Define various optimal classification rules.
- Use linear and quadratic discriminant analysis to classify observations.
- Use hypothesis testing to determine whether the assumptions of linear discriminant analysis are satisfied.
- Fit, tune, and assess a support vector machine to a specified dataset and use it to predict new observations.
- Explain the assumptions underlying the above inferential procedures and check them.

Topics we will cover are:

- Topic 1: Concepts of classification
- Topic 2: Linear discriminant analysis
- Topic 3: Quadratic discriminant analysis
- Topic 4: Support vector machines concepts and overview of estimation
- Topic 5: Tuning support vector machines

Optional readings

An alternative presentation of the concepts for this week can be found in:

Johnson, R. A., & Wichern, D. (2008). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall.

- 11.1–11.6.

All readings are available from the course Leganto reading list. Please keep in mind that you will need to be logged into Moodle to access the Leganto reading list.

Questions about this week's topics?

This week's topics were prepared by Dr P. Krivitsky. If you have any questions or comments, please post them under Discussion or email directly: p.krivitsky@unsw.edu.au

Discrimination and classification

Introduction: Separation and classification for two populations

Discriminant analysis and classification are widely used in multivariate techniques. The goal is either *separating sets of objects* (in discriminant analysis terminology) or *allocating new objects to given groups* (in classification theory terminology).

Basically, discriminant analysis is more exploratory in nature than classification. However, the difference is not significant especially because very often a function that separates may sometimes serve as an allocator, and, conversely, a rule of allocation may suggest a discriminatory procedure. In practice, the goals in the two procedures often overlap.

We will consider the case of two populations (classes of objects) first. Typical examples include: an anthropologist wants to classify a skull as a male or female; a patient needs to be classified as needing surgery or not needing surgery etc.

Denote the two classes by π_1 and π_2. The separation is to be performed on the basis of measurements of p associated random variables that form a vector $\boldsymbol{X} \in \mathbb{R}^p$. The observed values of $\boldsymbol{X}$ belong to different distributions when taken from π_1 and π_2 and we shall denote the densities of these two distributions by $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, respectively.

Allocation or classification is possible due to the fact that one has a *learning sample* at hand, i.e., there are some measurement vectors that are known to have been generated from each of the two populations. These measurements have been generated in earlier similar experiments. The goal is to partition the sample space into 2 mutually exclusive regions, say R_1 and R_2, such that if a *new* observation falls in R_1, it is allocated to π_1 and if it falls in R_2, it is allocated to π_2.

Classification errors

There is always a chance of an erroneous classification (misclassification). Our goal will be to develop such classification methods that in a suitably defined sense minimise the chances of misclassification.

It should be noted that one of the two classes may have a greater likelihood of occurrence because one of the two populations might be much larger than the other. For example, there tend to be a lot more financially sound companies than bankrupt companies. These *prior probabilities* of occurrence should also be taken into account when constructing an optimal classification rule if we want to perform optimally.

In a more detailed study of optimal classification rules, cost is also important. If classifying a π_1 object to the class π_2 represents a much more serious error than classifying a π_2 object to the class π_1 then these cost differences should also be taken into account when designing the optimal rule.

The **conditional** probabilities for misclassification are defined naturally as:

$$\Pr(2|1) = \Pr(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} \tag{5.1}$$

$$\Pr(1|2) = \Pr(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \tag{5.2}$$

Summarising

We turn briefly to the question of how to summarise a classifier's performance. Each object has a true class membership and the one predicted by the classifier, and for a given dataset for which true memberships are known, we may summarise the counts of the four resulting possibilities in a contingency table called a *confusion matrix*, i.e.,

A confusion matrix can be produced when there are more than two classes as well.

In the special case where there are two classes that can be meaningfully labelled as Negative/Positive, False/True, No/Yes, Null/Alternative, or similar, it is common to use the following terminology for them:

One can then define various performance metrics such as

sensitivity (a.k.a. recall, true positive rate (TPR)): $\Pr(\text{Predicted positive}|\text{Actual positive}) = \frac{\text{TP}}{\text{TP}+\text{FN}}$

specificity (a.k.a. selectivity, true negative rate (TNR)):
$\Pr(\text{Predicted negative}|\text{Actual negative}) = \frac{\text{TN}}{\text{TN}+\text{FP}}$

false positive rate (a.k.a. (FPR), fall-out): $\Pr(\text{Predicted positive}|\text{Actual negative}) = \frac{\text{FP}}{\text{TN}+\text{FP}} = 1 - \text{TNR}$

precision (a.k.a. positive predictive value): $\Pr(\text{Actual positive}|\text{Predicted positive}) = \frac{\text{TP}}{\text{TP}+\text{FP}}$

negative predictive value: $\Pr(\text{Actual negative}|\text{Predicted negative}) = \frac{\text{TN}}{\text{TN}+\text{FN}}$

F1 score: $\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$

Many classifiers return a continuous score that needs to be thresholded to produce a binary decision (e.g., predict "Yes" if the score exceeds some constant k and "No" otherwise), it is a common practice to plot a *receiver operating characteristic* (ROC) curve by varying the threshold and then plotting the TPR (on the vertical axis) against FPR (on the horizontal axis) that result. Both of which decrease as k increases. A perfect classifier would have a threshold for which the curve achieves the $(0, 1)$ point, whereas classifier close to the $y = x$ line is no better than chance.

Optimal classification rules

Rules that minimise the expected cost of misclassification (ECM)

Lemma 5.1. *Denote by p_i the **prior** probability of $\pi_i, i = 1, 2, p_1 + p_2 = 1$. Then the **overall** probabilities of incorrectly classifying objects will be:* $\Pr(\text{misclassified as } \pi_1) = \Pr(1|2)p_2$ *and* $\Pr(\text{misclassified as } \pi_2) = \Pr(2|1)p_1$. *Further, let $c(i|j), i \neq j, i, j = 1, 2$ be the misclassification costs. Then the **expected cost of misclassification** is*

$$ECM = c(2|1)\Pr(2|1)p_1 + c(1|2)\Pr(1|2)p_2 \tag{5.3}$$

The regions R_1 and R_2 that minimise ECM are given by

$$R_1 = \left\{ x : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right\} \tag{5.4}$$

and

$$R_2 = \left\{ x : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right\}. \tag{5.5}$$

Proof: It is easy to see that $\text{ECM} = \int_{R_1} \left[c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x}) \right] d\boldsymbol{x} + c(2|1)p_1$. Hence, the ECM will be minimised if R_1 includes those values of $\boldsymbol{x}$ for which the integrand $\left[c(1|2)p_2 f_2(\boldsymbol{x}) - c(2|1)p_1 f_1(\boldsymbol{x}) \right] \leq 0$ and excludes all the complementary values.

Note, the significance of the fact that in Lemma 5.1 **only ratios** are involved. Often in practice, one would have a much clearer idea about the cost ratio rather than for the actual costs themselves.

For your own exercise, consider the partial cases of Lemma 5.1 when $p_2 = p_1$, $c(1|2) = c(2|1)$ and when both these equalities hold. Comment on the soundness of the classification regions in these cases.

Rules that minimise the total probability of misclassification (TPM)

If we ignore the cost of misclassification, we can define the total probability of misclassification as

$$\text{TPM} = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

Mathematically, this is a particular case of Lemma 5.1 when the costs of misclassification are equal—so nothing new here.

Bayesian approach

Here, we try to allocate a new observation $\boldsymbol{x}_0$ to the population with the larger posterior probability $\Pr(\pi_i|\boldsymbol{x}_0), i = 1, 2$. According to Bayes's formula we have

$$\Pr(\pi_1|\boldsymbol{x}_0) = \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}, \ \Pr(\pi_2|\boldsymbol{x}_0) = \frac{p_2 f_2(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}$$

Mathematically, the strategy of classifying an observation $\boldsymbol{x}_0$ as π_1 if $\Pr(\pi_1|\boldsymbol{x}_0) > \Pr(\pi_2|\boldsymbol{x}_0)$ is again a particular case of Lemma 5.1 when the costs of misclassification are equal. But note that the calculation of the posterior probabilities themselves is in itself a useful and informative operation.