Supervised Learning — Classification

In general, learning, as in statistical learning, machine learning, supervised learning, is to learn about output variables and output patterns from input variables. There are various naming conventions for the variables involved.

Output variable: response, outcome, dependent variables

Input variable: explanatory variables, independent variables, features

Supervised learning

A term popularized by the Machine Learning community, supervised learning infers the task or process of deriving a function mapping an input to an output, when the output variable is available in the training data. The modeling ($a.k.a.$ learning) process aims to develop a rule to predict the values of the output, based on the values of the input variables only.

Classification

Classification is a type of supervised learning when the output variable is categorical. Each category is conventionally called a class, or a sub-population.

Examples of classification problems:

- A patient has a set of symptoms, likely due to one of the four (non-overlapping) medical conditions. The medical examination process to identify the condition is expensive and takes time. Based on past experience and available demographic information of the patient, such as gender, age, income, and address, can we derive a quick prognosis rule to suggest which medical condition the patient most likely has?
- A person is applying for a loan from a bank. Given the applicant's demographic information and past credit history, the bank needs to give an answer of "Yes" or "No" based on the promise for the person to pay off the loan. In other words, the bank needs a rule to classify the applicant into one of the two classes.
- Researchers try to derive a rule to determine whether a person has disease condition D, using the person's DNA sequence. The rule is based on the patterns in DNA sequences of previous patients with or without the condition D. Quantitative measurements from DNA sequences are often on gene expressions, or on single nucleotide polymorphisms (SNPs) patterns.

Classification and Discrimination

The two terms Classification and Discrimination have been used almost interchangeably historically. The team Classification is more commonly used nowadays.

Both are common techniques used to separate observations into distinct groups and to allocate new observations into the "correct" class, when only the values of input variables are available. Classification is one of the basic aims of all scientific research. Our brains do classifications automatically in most aspects of life.

Traditionally, the goals of discrimination and classification are not identical but overlapping.

Goal of Discrimination

Given a population known to have several subpopulations, discrimination is to find a way (discriminant function) to characterize the nature of the differences, furthermore, to find discriminants whose numerical values are such that the populations are separated as much as possible. In the process, one may find characters that do not matter (in-discriminant). The key word is "separation" or rather "characterization of separation".

Goal of Classification

To sort observations into two or more labeled classes. To derive a rule that can be used to optimally assign new objects to the labeled classes. The key word: is"allocation" or "assignment".

There are no clear boundaries between discriminant analysis and classification. Sometimes discrimination implies the analysis part, thus more exploratory especially when the groups are not predefined, and classification is the application of the separation rule with clearly defined classes.

Very often the whole process is simply termed as classification.

Black box classification

As discriminant functions become highly complicated, the interpretability or explainability of the classification rules can be diminished.

Various artificial neural network based classification rules used in machine learning and AI are examples of so called black box algorithms. Researchers may know the layers and activation functions of the neural network, but could have little grasp of the resulting classification function that is modeled by training the neural network, and may not have a mathematical understanding why a certain sample would be incorrectly classified.

Nonetheless neural network based classification rules could be and often are the best performers in classification in practice, evaluated by the rate of correct categorizations, as long as there are enough training data.

Recent research attempts to characterize the effectiveness of the neural network at a deeper mathematical level and to shed lights on how the process works. Black box procedures can become more and more transparent.

Classifiers

A classification or a discriminant rule, also called a **learner** in Machine Learning, allocates objects into classes, often in the form of a decision function.

Data for supervised learning

In general, supervised learning starts the learning process by using available data with output variable information, also called "labeled" instances. A dataset of size n can be written as $\{(\boldsymbol{x}_i, y_i), i = 1, \cdots, n\}$, where vector $\boldsymbol{x}_i$'s are inputs, y_i's are responses.

Here we consider the simple case of input variable vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, response variable $y_i \in \mathbb{R}$.

For example, a sample of size n may look like the following, where each row is one observation point.

X_1	X_2	$\cdots$	X_p	output
x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

The objective is to use the data in the above form to develop a rule to predict outcome for future data when only input values are available. Future data may look like the following.

X_1	X_2	$\cdots$	X_p	outcome
x_{11}	x_{12}	$\cdots$	x_{1p}	?
x_{21}	x_{22}	$\cdots$	x_{2p}	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Data for classification

For classification data, the output variable is categorical. The categories are called classes or sub-populations.

The classes are often labeled numerically, e.g., the classes may be indexed as $1, 2, \cdots, g$. However, in general, the numbers denoting the classes are symbols only, the numerical order in the label has no (numerical) meaning.

In supervised learning, classifiers are developed (i.e. trained) using labeled data, which consist of past observations containing information of known class membership, as well as the set of relevant characters or properties of explanatory variables. Sorted data used for developing classifiers may look like the following.

X_1	X_2	$\cdots$	X_p	class membership
$x_{1,11}$	$x_{1,12}$	$\cdots$	$x_{1,1p}$	1
$x_{1,21}$	$x_{1,22}$	$\cdots$	$x_{1,2p}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1,n_11}	x_{1,n_12}	$\cdots$	x_{1,n_1p}	1
$x_{2,11}$	$x_{2,12}$	$\cdots$	$x_{2,1p}$	2
$x_{2,21}$	$x_{2,22}$	$\cdots$	$x_{2,2p}$	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{2,n_21}	x_{2,n_22}	$\cdots$	x_{2,n_2p}	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{g,11}$	$x_{g,12}$	$\cdots$	$x_{g,1p}$	g
$x_{g,21}$	$x_{g,22}$	$\cdots$	$x_{g,2p}$	g
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{g,n_g1}	x_{g,n_g2}	$\cdots$	x_{g,n_gp}	g

The key question in classification problems can be stated as the following: given future observations each with a set of characters or properties (without knowing their class membership), which class does each of the new observations belong?

Given n_t to-be-classified future observations, these new data can be formulated in the following data form.

X_1	X_2	$\cdots$	X_p	class
x_{11}	x_{12}	$\cdots$	x_{1p}	?
x_{21}	x_{22}	$\cdots$	x_{2p}	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n_t1}	x_{n_t2}	$\cdots$	x_{n_tp}	?

Data partitions for classification learning

To develop and evaluate a classifier, the available data with known class labels are often split into sub-datasets for the modeling and learning process.

- Training data

  A set of data with known class membership used to develop classifiers.

- Validating data

  A set of data with known class membership used to evaluate and tune candidate classifiers which are developed using the training data.

  The observations are classified by candidate classifiers with the class membership masked.

  The true membership is used to compare and select candidate classifiers.

- Testing data

  Ideally, a set of data with known class membership can be set aside to obtain performance measurements of the classifier, such as accuracy and sensitivity.

  Due to limited size of data, setting testing data aside may not be affordable. Then, instead of setting testing data aside during model development, cross validation is often used in practice.

Training and validation of classifiers

- Training

  The training process includes estimation of parameters on the training datasets, tuning hyperparameters on validation dataset, and reporting results on the testing dataset.

- Evaluation

  There are various ways to evaluate a classification rule. The criteria used to evaluate the goodness of a classifier should be the same or close to the criteria used to train the classifier.

- Statistical concerns

  The objective is to develop a classifier to do well on testing dataset. A common concern is overfitting, when the classifier works and fits closely on the training data but does not generalize well. A classifier should come with realistic estimates of its accuracy.

The state-of-the-art classifiers are neural network methods (such as deep learning) which are topics covered in details in other machine learning courses.

This course covers the basics related to probabilistic classifications, while briefly mentioning a few non-probabilistic methods such as nearest neighbor, SVM and tree methods.

1 Probabilistic classification basics (two populations)

The main target for a classification procedure is to assign an observation to its class as accurate as possible. There are various metrics to measure the accuracy of a classifier.

In probabilistic classification, probability language is used to quantify uncertainty and to describe the population that the sample data came from and from which future observations will be drawn; and statistics methods are used to evaluate the goodness of the classification assignment.

In this section, we introduce the basic concepts, terminologies, and criteria in probability classification processes, illustrated with the two-population classification case.

Let $\boldsymbol{X} = (X_1, \cdots, X_p) = [X_1 \ \cdots \ X_p]'$ denote a random vector of measurements on objects possibly from several populations. In other words, the observed values of $\boldsymbol{X}$ are supposedly of different distributions for objects belonging to different populations or sub-populations. Classifiers, that is, classification rules, attempt to caption the differences

between the populations. A new observation $\boldsymbol{x} = (x_1, \cdots, x_p)$ without knowledge of the original population it belongs to will be assigned to a class by a classifier.

In the case of two populations, an object with associated input variable measurements $\boldsymbol{x} = (x_1, \cdots, x_p)$ will be assigned to one of the two classes according to the allocation rule.

1.1 Notations of classification

A common convention (as in the textbook by Johnson and Wichern) is to denote the two sub-populations as π_1 and π_2 (not to be confused with the mathematical constant).

Classification regions

For a classifier which defines a classification rule,

- $\Omega = \{$the collection of all possible values of $\boldsymbol{x}\}$ is the sample space.
- $R_1 =$ the set or the region of all $\boldsymbol{x}$ values for which we classify objects as π_1,
  $R_2 = \Omega - R_1$ be the remaining $\boldsymbol{x}$ values for which we classify objects as π_2,
  R_i's are mutually exclusive and exhaustive: $R_1 \cap R_2 = \varnothing$, $R_1 \dot\cup R_2 = \Omega$.

Conditional probabilities and classification errors

Let $f_i(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^p$, denote the probability density function of random vector $\boldsymbol{X}$ for the ith population, where $i = 1, 2$, is the index for the two populations.

Given random events A and B, recall the probability of the intercept event and conditional event,

$$\mathbb{P}(A \cap B) = \mathbb{P}(AB) = \mathbb{P}(A|B)\,\mathbb{P}(B) = \mathbb{P}(B|A)\,\mathbb{P}(A)$$

where $\mathbb{P}(A|B)$ is the conditional probability of event A given the presence of event B, when $\mathbb{P}(B) > 0$.

Joint probability mass function of two discrete random variables has similar relations.

$$\mathbb{P}(Y = y, Z = z) = \mathbb{P}(Y = y|Z = z)\,\mathbb{P}(Z = z) = \mathbb{P}(Z = z|Y = y)\,\mathbb{P}(Y = y)$$

Analogously, joint probability density function of continuous random variables can be expressed via conditional densities,

$$f_{X,Y}(x, y) = f_{X|Y}(x|y)\, f_Y(y) = f_{Y|X}(y|x)\, f_X(x)$$

Mixed cases have similar, adapted expressions.

Define the **conditional probabilities of classification and misclassification** as

$$P(i|j) = \mathbb{P}(\text{classifying an object as in population } \pi_i \mid \text{in fact the object is from population } \pi_j)$$

The **classification errors** can be stated in terms of conditional probabilities.

- $P(2|1) = \mathbb{P}(\boldsymbol{X} \in R_2|\pi_1) = \displaystyle\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}$.
  $P(1|1) = 1 - P(2|1)$.
- $P(1|2) = \mathbb{P}(\boldsymbol{X} \in R_1|\pi_2) = \displaystyle\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$.
  $P(2|2) = 1 - P(1|2)$.

Suppose the probability of a randomly sampled observation belonging to population $\pi_i, i = 1, 2$ is

$$\mathbb{P}(\pi_i) = p_i, \qquad p_1 + p_2 = 1$$

Then

$$\begin{aligned}
&\mathbb{P}(\text{a randomly sampled observation correctly classified as } \pi_i) \\
&= \mathbb{P}(\text{the observation is from } \pi_i, \text{and it is classified as } \pi_i) \\
&= \mathbb{P}(\text{the observation is from } \pi_i) \cdot \mathbb{P}(\text{the observation is classified as } \pi_i \mid \text{the observation is from } \pi_i) \\
&= \mathbb{P}(\pi_i) \cdot P(i|i) = P(i|i)\ p_i
\end{aligned}$$

For two populations, the **unconditional probabilities of classification** are

- $\mathbb{P}(\text{a randomly sampled observation correctly classified as } \pi_1) = P(1|1)\, p_1$
- $\mathbb{P}(\text{a randomly sampled observation incorrectly classified as } \pi_1) = P(1|2)\, p_2$
- $\mathbb{P}(\text{a randomly sampled observation correctly classified as } \pi_2) = P(2|2)\, p_2$
- $\mathbb{P}(\text{a randomly sampled observation incorrectly classified as } \pi_2) = P(2|1)\, p_1$

Prior distributions

The true population membership probabilities (the ideal p_i's) are theoretical values, unknown in practical classification problems.

In probability derivation, p_i's are treated as the true population membership probabilities. In real data applications, they are best guess at the moment, often assigned at the beginning based on prior knowledge or other properties of the sub-populations, and are updated later based on observed data. Therefore the p_i's are called prior probability distributions or simply priors of the sub-populations.

Cost of misclassification

Often there is an associated cost $c(i|j)$ for wrongly classifying an object from class j to class i. The **misclassification cost** can be described by a cost matrix.

	Classified as	
True population	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

To obtain a useful expression for expected cost of misclassification, recall that, for a discrete random variable Y, its expectation is defined as

$$\mathbb{E}(Y) = \sum_{possible\ y's} y\,\mathbb{P}(Y = y)$$

Given the knowledge of another random variable, say W, the expectation of Y can be formulated as the expected value of the conditional expected value given the other random variable, as long as the expectations are well defined. This property is known as the law of total expectation, or more descriptively, the law of iterated expectations, expressed as

$$\mathbb{E}(Y) = \mathbb{E}[\mathbb{E}(Y|W)] = \sum_{possible\ w's} \mathbb{E}(Y|W = w)\mathbb{P}(W = w)$$

where $\mathbb{E}(Y|W = w)$ is the conditional expectation of Y given $W = w$, assuming W is a discrete random variable.

Using the law of iterated expectations, the **expected cost of misclassification** (ECM) can be broken down as

$$\begin{aligned} ECM &= \mathbb{E}(\text{cost of misclassifying the object into } \pi_2 \mid \text{object is misclassified as } \pi_2)\, \mathbb{P}(\text{object is misclassified as } \pi_2) \\ &\quad + \mathbb{E}(\text{cost of misclassifying the object into } \pi_1 \mid \text{object is misclassified as } \pi_1)\, \mathbb{P}(\text{object is misclassified as } \pi_1) \\ &= c(2|1)P(2|1)\, p_1 + c(1|2)P(1|2)\, p_2 \end{aligned}$$

Naturally, the smaller the expected cost of misclassification, the better.

1.2 Classification criteria

Misclassification can be described in terms of probability.

Minimum ECM - the Expected Cost of Misclassification

It is reasonable to desire for a classification rule that minimizes the overall cost of misclassification, which can be formulated as minimizing the average cost or the expected cost of misclassification, the ECM.

The optimal classifier can be described by its classification regions.

Recall that R_i denotes the region of all $\boldsymbol{x}$ values that are classified as in population π_i by the classifier, $c(j|i)$ is the cost of a π_i object being misclassified into π_j. where the definition of classification regions

$$R_i = \{\boldsymbol{x} :\ \boldsymbol{x} \text{ is assigned to class } \pi_i \text{ by the classification rule } \}, \qquad i = 1, \cdots, g.$$

As shown in the proof below, in the case of two classes, the optimal classification regions R_1 and R_2 that minimize the expected cost of misclassification ECM have the form

$$\begin{cases} R_1 &= \{\boldsymbol{x} : c(2|1)\, p_1 f_1(\boldsymbol{x}) \geq c(1|2)\, p_2 f_2(\boldsymbol{x})\} \\ R_2 &= \{\boldsymbol{x} : c(2|1)\, p_1 f_1(\boldsymbol{x}) < c(1|2)\, p_2 f_2(\boldsymbol{x})\} \end{cases} \tag{1}$$

The regions can be written in the ratio form, which is often easier to use.

$$R_1 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}, \qquad R_2 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\} \tag{2}$$

when the denominators are non-zero, otherwise the reciprocal version or definition (1) can be used instead.

Proof. In the following we prove the optimality of the classification rule in (1), equivalently, (2).

Recall that R_1 is the region of all $\boldsymbol{x}$ values that are classified as in population π_1 by the classification rule. We are to find the region R_1 which minimize the expected cost of misclassification ECM. Recall that Ω is the sample space of all possible $\boldsymbol{x}$ values, $R_2 = \Omega \setminus R_1$.

$$\begin{aligned} ECM &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \\ &= p_1 c(2|1) \int_{\Omega \setminus R_1} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 c(1|2) \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= p_1 c(2|1) \int_{\Omega} f_1(\boldsymbol{x})d\boldsymbol{x} - p_1 c(2|1) \int_{R_1} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 c(1|2) \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \\ &= p_1 c(2|1) + \int_{R_1} \left[p_2\, c(1|2) f_2(\boldsymbol{x}) - p_1\, c(2|1) f_1(\boldsymbol{x})\right] d\boldsymbol{x} \end{aligned}$$

The term $p_1 c(2|1) \geq 0$. To minimize ECM is to make the integral over R_1 as small as possible, or rather, as negative as possible. Note that

- including any $\boldsymbol{x}$ in R_1 with the integrand $p_2 c(1|2) f_1(\boldsymbol{x}) - p_1 c(2|) f_2(\boldsymbol{x}) > 0$ would increasing the ECM,
- including any $\boldsymbol{x}$ in R_1 with $p_2 c(1|2) f_1(\boldsymbol{x}) - p_1 c(2|) f_2(\boldsymbol{x}) < 0$ would reduce the ECM.
- including any $\boldsymbol{x}$ in R_1 with $p_2 c(1|2) f_1(\boldsymbol{x}) - p_1 c(2|) f_2(\boldsymbol{x}) = 0$ would not change the ECM.

Therefore, the minimum ECM is reached if and only if all $\boldsymbol{x}$ making the integrand $<$ will be assigned into R_1. Therefore, we may choose

$$R_1 = \{\boldsymbol{x} :\ p_2\, c(1|2) f_2(\boldsymbol{x}) \ -\ p_1\, c(2|1) f_1(\boldsymbol{x}) \leq 0\ \} = \{\boldsymbol{x} :\ p_2\, c(1|2) f_2(\boldsymbol{x}) \ <\ p_1\, c(2|1) f_1(\boldsymbol{x})\ \}$$

This is the definition in (1). In the non-degenerated case of $f_2(\boldsymbol{x}) > 0, c(2|1) > 0, p_1 > 0$, (1) is equivalent to

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

as shown in (2).

□

Remarks

- By the continuity of the distribution, $\mathbb{P}\{\boldsymbol{x} :\ p_2\, c(1|2) f_2(\boldsymbol{x}) \ <\ p_1\, c(2|1) f_1(\boldsymbol{x})\ \} = 0$. Thus it is valid to assign boundary points to either region. For example, we may choose $R_1 = \{\boldsymbol{x} :\ p_2\, c(1|2) f_2(\boldsymbol{x}) \ <\ p_1\, c(2|1) f_1(\boldsymbol{x})\ \}$.
- The formulation of R_i in (2) is to show case the ratio relationship, without addressing minor technical details. For example, The formulation assumes implicitly that $f_2(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x}$ in the sample space. Otherwise, whenever $f_2(\boldsymbol{x}) = 0$ and $f_1(\boldsymbol{x}) \neq 0$, the terms should be expressed in $f_2(\boldsymbol{x})/f_1(\boldsymbol{x})$ instead.
- The rule used to assign an observation to R_1 or R_2 in (2) is stated as comparing three ratios,

$$\binom{density}{ratio}, \qquad \binom{cost}{ratio}, \qquad \binom{prior}{ratio}$$

- In practice it is often much easier to compute the ratio of two quantities, such as the ratio of costs and the ratio of probability densities, than to obtain the exact values of the two quantities themselves. This advantage is another reason to express the classification regions in terms of ratios as in (2).

Minimum TPM classification

When cost is unknown or homogeneous, another criterion can be considered in terms of **total probability of misclassification** (TPM).

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

When the density functions (rather than only their ratios) are known, the TPM can be evaluated directly.

In fact,, minimizing the TPM is the same as minimizing the expected cost of misclassification ECM in the trivial case that the misclassification costs $c(j|i), j \neq i$ are the same for all classes. Then the optimal classification regions are simplified to

$$\begin{cases} R_1 = & \left\{\boldsymbol{x} : \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \dfrac{p_2}{p_1}\right\} = \left\{\boldsymbol{x} : \dfrac{p_1}{p_2}\dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1\right\} \\ R_2 = & \left\{\boldsymbol{x} : \dfrac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \dfrac{p_2}{p_1}\right\} = \left\{\boldsymbol{x} : \dfrac{p_2}{p_1}\dfrac{f_2(\boldsymbol{x})}{f_1(\boldsymbol{x})} > 1\right\} \end{cases}$$

when the denominators are nonzero.

Bayesian method — Maximum posterior probability classification

A (naïve) Bayesian approach is to allocate new observation $\boldsymbol{x}_o$ to the class with the largest posterior probability $P(\pi_i|\boldsymbol{x}_o)$.

By the Bayes' Rule, the posterior probability for two-class case is defined as

$$\begin{aligned} P(\pi_1|\boldsymbol{x}_o) &= \frac{\mathbb{P}(\text{we observe } \boldsymbol{x}_o, \text{ and } \boldsymbol{x}_o \text{ is from } \pi_1)}{\mathbb{P}(\text{we observe } \boldsymbol{x}_o)} \\ &= \frac{\mathbb{P}(\text{we observe } \boldsymbol{x}_o|\pi_1)\,\mathbb{P}(\pi_1)}{\mathbb{P}(\text{we observe } \boldsymbol{x}_o|\pi_1)\,\mathbb{P}(\pi_1) + \mathbb{P}(\text{we observe } \boldsymbol{x}_o|\pi_2)\,\mathbb{P}(\pi_2)} \end{aligned}$$

Assume that the prior probability of being in class i is p_i and the mass function or density function for population i is f_i. Then the posterior probabilities can be expressed as

$$\begin{aligned} P(\pi_1|\boldsymbol{x}_o) &= \frac{p_1 f_1(\boldsymbol{x}_o)}{p_1 f_1(\boldsymbol{x}_o) + p_2 f_2(\boldsymbol{x}_o)} \\ P(\pi_2|\boldsymbol{x}_o) &= \frac{p_2 f_2(\boldsymbol{x}_o)}{p_1 f_1(\boldsymbol{x}_o) + p_2 f_2(\boldsymbol{x}_o)} \end{aligned}$$

Often as an default, the (naïve) Bayesian rule may classifies an observation $\boldsymbol{x}_o$ to population π_1 if the posterior $P(\pi_1|\boldsymbol{x}_o) \geq P(\pi_2|\boldsymbol{x}_o)$. So the classification region

$$\begin{aligned} R_1 &= \{\boldsymbol{x} : P(\pi_1|\boldsymbol{x}) \geq P(\pi_2|\boldsymbol{x})\} \\ &= \left\{\boldsymbol{x} : \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} \geq \frac{p_2 f_2(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})}\right\} = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1}\right\} \end{aligned}$$

and classified as in π_2 if otherwise,

$$R_2 = \{\boldsymbol{x} : P(\pi_2|\boldsymbol{x}) \geq P(\pi_1|\boldsymbol{x})\} = \left\{\boldsymbol{x} : \frac{f_2(\boldsymbol{x})}{f_1(\boldsymbol{x})} \geq \frac{p_1}{p_2}\right\}$$

Therefore maximizing posterior probability is equivalent to minimizing the Total Probability of Misclassification TPM.

While the criteria of minimizing total probability of misclassification and the Bayesian rule are the same, the posterior distribution contains a lot more information, thus more useful but harder to obtain.

1.3 Probability classification approaches

Probabilistic classification can be carried out in many different approaches, and the criteria of goodness of the classifier can be at various levels. Common classification methods include

Use discriminant functions

Model the conditional probability $P(i|j)$

Model the posterior distribution $P(\pi_i|\boldsymbol{x})$

Some are simpler to obtain and less computationally intensive. On the other hand, more complex approaches often provide more information.

For example, the values of the posterior probability $P(\pi_i|x)$ provide more information than simple yes-or-no class boundaries, especially in the not so clear-cut cases. The prior-posterior point of view is also very useful in many situations.

2 Classification with two multivariate normal populations

Under multivariate normal distributions, the abstract classification rule (2) has concrete analytic expressions.

Recall the optimal classification rule is

$$R_1 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}, \qquad R_2 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}.$$

2.1 Linear classification rules for normal populations when $\Sigma_1 = \Sigma_2 = \Sigma$

Let the populations $\pi_i, i = 1, 2$, be described by the density functions of $N_p(\boldsymbol{\mu}_i, \Sigma)$.

The allocation rule that minimizes the expected cost of misclassification ECM is to allocate $\boldsymbol{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right)$$

The rule allocates $\boldsymbol{x}_0$ to π_2 if otherwise.

Proof. Under the normality and equal variance assumptions, $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_i, \Sigma)$ if $\boldsymbol{x}$ is from population π_i. Thus the density ratio in the abstract classification rule (2) has an explicit expression

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = e^{-\frac{1}{2}[(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)]}$$

The expression inside the square brackets in the exponent can be regrouped as follows.

$$\begin{aligned} &(\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&(\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&(\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&(\boldsymbol{x} - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&-(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1 + \boldsymbol{x} - \boldsymbol{\mu}_2) \\ =&-(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}[2\boldsymbol{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \end{aligned}$$

Therefore,

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = e^{-\frac{1}{2}\{-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}[2\boldsymbol{x}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)]\}} = e^{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)}$$

The inequality

$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}$$

becomes

$$e^{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}$$

Since the costs and p_i are all positive, we may take logarithm to obtain

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x}_0 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right)$$

□

Sample classification rule (two normal samples, equal variance)

In practice, sample estimates are used in the place of the the unknown population measures $\boldsymbol{\mu}_i$ and Σ. The population covariance matrix Σ is estimated by the pooled sample covariance matrix $\boldsymbol{S}_{pool}$, and the mean $\boldsymbol{\mu}_i$ of sub-population π_i is estimated by the sample mean $\bar{\boldsymbol{x}}_i$ of class i.

Consequently, the sample classification rule is to allocate an observation $\boldsymbol{x}_0$ to π_1 if

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_{pool}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right)$$

and to allocate $\boldsymbol{x}_0$ to π_2 if otherwise.

The above classification rule of minimum ECM for two multivariate normal populations of equal covariance structure is a linear function of the new observation $\boldsymbol{x}_0$. The coefficients of the linear functions are determined by the training sample data.

2.2 Quadratic classification rule for normal populations when $\Sigma_1 \neq \Sigma_2$

When two subpopulations do not have a common covariance matrix, $\Sigma_1 \neq \Sigma_2$, their population density ratio has the form

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} e^{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]}$$

Taking logarithm,

$$\ln \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{1}{2}\left[\boldsymbol{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} - \boldsymbol{x}'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{x}'\Sigma_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{x} + \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2\right]$$

The minimum ECM classification regions

$$R_1 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}, \qquad R_2 = \left\{\boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right\}.$$

in (2) become

$$\begin{aligned} R_1: \ & -\frac{1}{2}\boldsymbol{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1})\boldsymbol{x} - k \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right) \\ R_2: \ & -\frac{1}{2}\boldsymbol{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1})\boldsymbol{x} - k < \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right) \end{aligned} \tag{3}$$

with

$$k = \frac{1}{2}(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2) + \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|}$$

Now the terms involving $\boldsymbol{x}$ are quadratic in $\boldsymbol{x}$. So the classification regions R_1 and R_2 can be in more than two connected pieces in general. One region, say, R_2, could be in the middle of two R_1 sections, or vice versa. The boundary surface of the two classes are quadratic.

In $p = 2$ case, the class borders are quadratic curves, or conic sections, which include parabolas, hyperbolas, ellipses, circles, and degenerate cases such as lines.

If the rule applies to the case $\Sigma_1 = \Sigma_2$, the quadratic term disappears, then rule is identical to the linear minimum ECM classification rule again.

Sample quadratic classification rule (two normal samples, unequal variance)

To classify a new observation $\boldsymbol{x}_o$, use sample statistics $\bar{\boldsymbol{x}}_i$ and $\boldsymbol{S}_i$ as estimates of $\boldsymbol{\mu}_i$ and Σ_i. The sample quadratic classification rule of normal populations can be stated as the following.

Allocate $\boldsymbol{x}_o$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_o'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_o + (\bar{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1} - \bar{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1})\boldsymbol{x}_o - \hat{k} \geq \ln\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right),$$

allocate $\boldsymbol{x}_o$ to π_2 otherwise, where $\hat{k}$ is the sample estimate

$$\hat{k} = \frac{1}{2}(\bar{\boldsymbol{x}}_1'\boldsymbol{S}_1^{-1}\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2'\boldsymbol{S}_2^{-1}\bar{\boldsymbol{x}}_2) + \frac{1}{2}\ln\frac{|\mathbf{S_1}|}{|\boldsymbol{S}_2|}$$

3 Fisher's linear discriminant function

R. A. Fisher considered linear projections of a multivariate random vector, that is, a linear combinations of the component variables, so as to transform multivariate distributions into univariate ones. The linear projection is chosen to achieve **separation** or discrimination of the transformed sample distributions as much as possible.

> **Remarks**: Recall that in Principal Component Analysis and Canonical Correlation Analysis, we projecte multivariate distributions into simpler, univariate distributions. Hence achieving dimension reduction. Fisher's idea in discrimination analysis bears similarity: finding the direction $\boldsymbol{a}$ which maximizes the ratio of the between-class variance versus the within-class variance of $\boldsymbol{a}'\boldsymbol{X}$. Then the value of the univariate $Y = \boldsymbol{a}'\boldsymbol{X}$ is used to separate the classes of higher dimensional, p-variate data.

Fisher's linear discriminant analysis (LDA) is usually presented in the sample version.

Assume that an $(n_1 + n_2) \times p$ data matrix X consists of

n_1 p-variate vectors $\{\boldsymbol{x}_{1i}\}_{i=1,\cdots,n_1}$ observed from subpopulation π_1 with sample mean vector $\bar{\boldsymbol{x}}_1$, and

n_2 p-variate vectors $\{\boldsymbol{x}_{2i}\}_{i=1,\cdots,n_2}$ observed from subpopulation π_2 with sample mean vector $\bar{\boldsymbol{x}}_2$.

Let

$$\{y_{1i} = \boldsymbol{a}'\boldsymbol{x}_{1i}, i = 1, \cdots, n_1\}, \qquad \{y_{2i} = \boldsymbol{a}'\boldsymbol{x}_{2i}, i = 1, \cdots, n_2\}$$

be the two transformed univariate samples, with sample means

$$\bar{y}_1 = \boldsymbol{a}\bar{\boldsymbol{x}}_1, \qquad \bar{y}_2 = \boldsymbol{a}\bar{\boldsymbol{x}}_2$$

respectively.

Assuming equal covariance structure in the subpopulations, and X has sample covariance matrix $\boldsymbol{S}_x$, which can be estimated by the pooled covariance matrix $\mathbf{S}_{pool}$ from the data. the pooled sample variance of the two transformed samples is

$$s_y^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2} = \boldsymbol{a}'\hat{\boldsymbol{S}}_{\boldsymbol{x}}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}$$

The desired separation, measured by the statistical distance $(\bar{y}_1 - \bar{y}_2)^2/s_y^2$, is to be maximized.

Fisher's result

Of all $\boldsymbol{a} \in \mathbb{R}^p$, the linear transformation $\hat{y} = \hat{\boldsymbol{a}}'\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{x}$ maximizes the ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2))^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}}$$

The attained maximum is

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = \max_{\boldsymbol{y}_i = \boldsymbol{a}'\boldsymbol{x}_i} \frac{(\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2))^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}} = \max_{\boldsymbol{y}_i = \boldsymbol{a}'\boldsymbol{x}_i} \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

Proof. By the Maximization Lemma in linear algebra (ref. equation (2-50) in J&W), if B is a $p \times p$ symmetric positive definite matrix, then for a given $\boldsymbol{w} \in \mathbb{R}^p$,

$$\max_{\boldsymbol{v} \neq 0} \frac{(\boldsymbol{v}'\boldsymbol{w})^2}{\boldsymbol{v}'B\boldsymbol{v}} = \boldsymbol{w}'B^{-1}\boldsymbol{w} \tag{4}$$

and the maximum is attained if and only if $\boldsymbol{v} = cB^{-1}\boldsymbol{w}$ for some constant $c \neq 0$.

Here let $\boldsymbol{w} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 = \boldsymbol{d}$, $\boldsymbol{v} = \boldsymbol{a}$, and $B = \boldsymbol{S}_{pool}$. Then $\boldsymbol{v} = \boldsymbol{a} = \boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ (or its constant multiple) attains the maximum

$$\max_{\boldsymbol{a} \neq 0} \frac{(\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2))^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}} = \max_{\boldsymbol{v} \neq 0} \frac{(\boldsymbol{v}'\boldsymbol{w})^2}{\boldsymbol{v}'B\boldsymbol{v}} = \boldsymbol{w}'B^{-1}\boldsymbol{w} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

□

For a new observation $\boldsymbol{x}_0$, its transformed position on the line $y = \boldsymbol{a}'\boldsymbol{x}_0$ is evaluated to assign its class membership, using the midpoint of the transformed means $\frac{1}{2}(\bar{y}_1 + \bar{y}_2)$ as the partition point.

Fisher's discriminant rule for sample data can be stated as following.

Fisher's sample linear discriminant rule ($a.k.a.$ Fisher Linear discriminant function)

Compute $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$ and $\boldsymbol{S}_{pool}$ from sample training data. Allocate new observation $\boldsymbol{x}_o$ to π_1 if

$$\hat{y}_o \geq \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

Equivalently, in terms of $\boldsymbol{x}$'s, assign new observation $\boldsymbol{x}_o$ to π_1 if

$$\hat{y}_o = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}\boldsymbol{x}_o \geq \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \hat{m}$$

and allocate $\boldsymbol{x}_o$ to π_2 if otherwise. We may check that

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{S}_{pool}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\hat{\boldsymbol{a}}'\bar{\boldsymbol{x}}_1 + \hat{\boldsymbol{a}}'\bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

It is equivalent to said that assigning $\boldsymbol{x}_o$ to class i if $|\hat{y}_o - \bar{y}_i|$ is smaller than $|\hat{y}_o - \bar{y}_j|$, $j \neq i$, that is, if the distance of y_o to the class i mean $\bar{y}_i$ is the closer one.

Remarks

- The intuition of Fisher's linear discriminants is to project high dimensional data in $\mathbb{R}^p$ to one dimensional lines (linear combinations of the original variables) in $\mathbb{R}$. Fisher's rule picks the line that yields **maximum separation** of between class and within class variance ratio.

- The linear discriminant for maximum separation can be used to classification. The classification boundary is a hyperplane in $\mathbb{R}^p$ consisting of points with equal distance to the transformed means $\bar{y}_1$ and $\bar{y}_2$.
- The maximum ratio D^2 can be viewed as the normalized square of distance between the class means of the transformed variable y, as well as the square of the statistical distance ($a.k.a.$ Mahalanobis distance) between the vector means of the original data.
- Recall that D^2 is used in Hotelling's T^2 to test if the two means are equal. The test can be used here to check if the separation of the two population is significant enough to apply classification. Under the assumption of normal distribution of equal variance for the two populations, the test statistic is

  $$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \cdot \frac{n_1 n_2}{n_1 + n_2} D^2 \sim F_{p, n_1 + n_2 - p - 1} \qquad under \quad H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

  There is not much point to conduct classification if the difference of the class means is not significant in the first place. On the other hand, significant difference of the class means is not sufficient to guarantee a good classification.

- Note that Fisher arrived at similar conclusions as the normal classification under the ECM criteria, however from a different perspective without the normality assumption, which also testifies the ubiquitous nature of normal distributions.

4 How good is a classifier?

To judge the performance of a classification rule, it is natural to look at the extend of misclassification, or the error rates.

There are many classification error rates, many may be tailored to the specific classification task. We introduce several commonly considered error rates, some of them may be under different names.

- Theoretical optimum error rate, or minimum total probability of misclassification; and minimum expected cost of misclassification when unequal cost occurs.
- Actual error rate, estimated by testing error rate.
- Apparent error rate, or training error rate.
- Expected actual error rate and its sample estimates, estimated by validation and testing error rate.

To estimate the classification errors, typically we need training data, validation data, and testing data.

Optimum error rate

Recall that if the prior population probabilities p_i and the population density distributions f_i are known, the total probability of misclassification (TPM) is

$$TPM = p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \tag{5}$$

The smaller the total probability of misclassification, the better. The choices of R_1, R_2 that minimize TPM achieve the **optimum error rate** (OER).

In fact we have already derived that the optimal classifier minimizing expected cost of misclassification is rule (2), with classification regions

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \right\}, \qquad R_2 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \right\}$$

Thus under equal cost $c(1|2) = c(2|1)$, the desirable optimal error rate OER is achieved by the minimum TPM classification rule.

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1} \right\}, \qquad R_2 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1} \right\}$$

For $\boldsymbol{x}$ of multivariate distribution $N(\boldsymbol{\mu}_i, \Sigma)$, that is, the populations are of equal covariance structure, then under equal cost $c(1|2) = c(2|1)$ and equal prior $p_1 = p_2 = \frac{1}{2}$, the above classification rule can be explicit and simplified. The allocation is to assign $\boldsymbol{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x}_0 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \ \geq 0$$

If in fact $\boldsymbol{x}_0$ is from π_2, $\boldsymbol{x}_0 \sim N(\boldsymbol{\mu}_2, \Sigma)$, then the univariate variable

$$y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x}_0 \ \sim \ N\left(\ (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_2, \ (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \ \right)$$

Then the optimal error rate OPM can be simplified. The last integral in (5) is

$$\begin{aligned}
\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} &= P(1|2) = \mathbb{P}(\boldsymbol{x} \text{ is incorrectly classified as } \pi_1) \\
&= \mathbb{P}\left(y_0 > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \\
&= \mathbb{P}\left(\frac{y_0 - E(y_0)}{\sqrt{var(y_0)}} > \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_2}{\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}} \right) \\
&= \mathbb{P}\left(Z > \frac{1}{2}\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} \right) \\
&= \mathbb{P}\left(Z \leq -\frac{1}{2}\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} \right) \\
&= \Phi\left(-\frac{\Delta}{2} \right)
\end{aligned}$$

where $Z \sim N(0,1)$ is a univariate standard normal random variable, Φ is the cumulative distribution function of Z, and

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

is the square of standardized distance of the two population means. Similarly the first integral in (5) can be evaluated.

$$\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} = \cdots = \Phi\left(-\frac{\Delta}{2} \right)$$

Thus the optimal error rate minimizing the total probability of misclassification is

$$OER = \min(TPM) = \frac{1}{2}\Phi\left(-\frac{\Delta}{2} \right) + \frac{1}{2}\Phi\left(-\frac{\Delta}{2} \right) = \Phi\left(-\frac{\Delta}{2} \right) = 1 - \Phi\left(\frac{\Delta}{2} \right)$$

and Δ can be estimated from sample statistics as

$$\hat{\Delta}^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'S_{pool}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$$

The result is for normal populations with equal variance, under equal prior and equal costs.

In practice, the population densities are unknown, the above estimate also has errors that can be hard to evaluate.

Actual error rate

The **actual error rate** (AER) of a classification rule is

$$AER = p_1 \int_{\hat{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{\hat{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

where $\hat{R}_i$ are the classification regions derived from sample data. The AER is supposed to indicate how the sample classification rule will perform in future observations. But it cannot be calculated using sample data, since the unknown population density functions f_1, f_2 still appear in the above expression. However we will look at sample approximations of the expected value of AER.

Apparent error rate (training error rate)

A measure that does not depend on the form of populations is the **training error rate** or the **apparent error rate** (APER), which is defined as the fraction of observations in the **training** sample that are misclassified by the sample classification function. APER can be easily calculated from the following (confusion) matrix obtained from the training data classification:

	Predicted membership		
True membership	π_1	π_2	total
π_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
π_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

where

n_{1c} = number of π_1 items correctly classified as π_1 items
n_{1m} = number of π_1 items incorrectly classified as π_2 items
n_{2c} = number of π_2 items correctly classified as π_2 items
n_{2m} = number of π_2 items incorrectly classified as π_1 items

The training error rate or the apparent error rate is

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2}$$

which is intuitive and easy to calculate.

Estimation of expected Actual Error Rate (training error rate)

We need to estimate Actual Error Rate AER, or $E(AER)$, the expected true misclassification rate, based on available data.

Is the Apparent Error Rate (APER) $\frac{n_{1m} + n_{2m}}{n_1 + n_2}$ a good measure of misclassification?

The APER is easy to calculate and appealing. However using Apparent Error Rate to estimate the expected actual error rate E(AER) is often too optimal, since APER is evaluated on the data that the classification rule comes from.

In practice, the estimation is via validation or cross-validation error rates.

A common practice to estimate misclassification rate is to split the total sample data into a training set and a validation set. The training set is used to construct the classification rule, the validation set is use to evaluate the misclassification rate. If the sample size is not large, this method would be wasteful and not practical.

A modification of the validation process is the omit-one-at-a-time approach, or Lachenbruch's "holdout" procedure. In this approach, one observation is omitted during the development of the classification rule, then this left out observation is classified by the rule. The process is repeated with every observation. Using $n_{im}^{(H)}$ denote the number of misclassifications of the **holdout** observations in π_i, we may obtain estimates of the misclassification probabilities

$$\hat{P}(2|1) = \frac{n_{1m}^{(H)}}{n_1}, \qquad \hat{P}(1|2) = \frac{n_{2m}^{(H)}}{n_2}.$$

The total proportion can be used to estimate the **expected apparent error rate** or the **training error rate**:

$$\hat{E}(AER) = \frac{n_{1m}^{(H)} + n_{2m}^{(H)}}{n_1 + n_2}$$

The approximation method can be generalized to classifications of $g > 2$ populations.

$$\hat{E}(AER) = \frac{n_{1m}^{(H)} + \cdots + n_{gm}^{(H)}}{n_1 + \cdots + n_g} = \frac{\sum_{i=1}^{g} n_{im}^{(H)}}{\sum_{i=1}^{g} n_i}$$

5 Classification of more than two populations

Many concepts in above can be generalized into the classification of more than two populations.

Suppose the probability of a randomly sampled observation belonging to population $\pi_i, i = 1, \cdots, g$ is

$$\mathbb{P}(\pi_i) = p_i, \qquad \sum_{i=1}^{g} p_i = 1.$$

Then, just as in the two-population case,

$$\begin{aligned}
&\mathbb{P}(\text{a randomly sampled observation correctly classified as } \pi_i) \\
&= \mathbb{P}(\text{the observation is from } \pi_i, \text{and it is classified as } \pi_i) \\
&= \mathbb{P}(\text{the observation is from } \pi_i) \cdot \mathbb{P}(\text{the observation is classified as } \pi_i \mid \text{the observation is from } \pi_i) \\
&= \mathbb{P}(\pi_i) \cdot P(i|i) = P(i|i)\ p_i
\end{aligned}$$

The **unconditional probabilities of misclassification** need to be in a summation of all cases,

$$\begin{aligned}
&\mathbb{P}(\text{a randomly sampled observation incorrectly classified as } \pi_i) \\
&= \mathbb{P}(\text{the observation is from } \pi_j \text{ for some } j \neq i, \text{and it is classified as } \pi_i) \\
&= \sum_{j,\ j\neq i} \mathbb{P}(\text{the observation is from } \pi_j) \cdot \mathbb{P}(\text{the observation is classified as } \pi_i \mid \text{the observation is from } \pi_j) \\
&= \sum_{j,\ j\neq i} \mathbb{P}(\pi_j) \cdot P(i|j) = \sum_{j,\ j\neq i} P(i|j)\ p_j
\end{aligned}$$

5.1 The minimum expected cost of misclassification for multiple classes

Notations

The notations are analogous to that for two population classification, now adjusted to $g \geq 2$ subpopulations.

- $p_i = P(\pi_i) =$ prior probability of population π_i, $i = 1, \cdots, g$.
- $c(k|i) =$ the cost of allocating an item to π_k when if in fact it belongs to π_i. $c(i|i) = 0$.
- $R_i =$ the region or the set of $\boldsymbol{x}$ classified as π_i, satisfying the properties

$$\cup_{i=1}^{g} R_i = \Omega, \qquad R_i \cap R_j = \varnothing \quad for \quad i \neq j.$$

- $P(k|i) = \mathbb{P}(\text{classifying item as in } \pi_k \mid \text{ the item is from } \pi_i) = \int_{R_k} f_i(\boldsymbol{x}) d\boldsymbol{x}$, where f_i the density function of population π_i.
- Expected cost of misclassifying an item $\boldsymbol{x}$ from π_i:

$$\begin{aligned}
ECM(i) &= \mathbb{E}(\text{the cost of classifying an object into } \pi_j, j \neq i \mid \text{the object is from } \pi_i) \\
&= \sum_{j=1}^{g} (\text{the cost of classifying an object into } \pi_j \mid \text{the object is from } \pi_i) P(j|i) \\
&= \sum_{j=1}^{g} c(j|i) P(j|i), \qquad with \quad c(i|i) = 0, \quad P(i|i) = 1 - \sum_{j\neq i} P(j|i).
\end{aligned}$$

- The expected cost over all items by the classification rule:

$$\begin{aligned}
ECM &= \mathbb{E}(\text{ Overall cost of misclassification }) \\
&= \sum_{i=1}^{g} \mathbb{E}(\text{the cost of classifying an object into } \pi_j, j \neq i \mid \text{the object is from } \pi_i)\, p_i \\
&= \sum_{i=1}^{g} p_i ECM(i) \\
&= \sum_{i=1}^{g} p_i \left(\sum_{j=1}^{g} P(j|i) c(j|i) \right) = \sum_{i=1}^{g} p_i \left(\sum_{j=1, j\neq i}^{g} P(j|i) c(j|i) \right)
\end{aligned}$$

Results

- Classification regions for more than two classes

  The classification regions R_k $(k = 1, \cdots, g)$ that minimize the overall expected cost of misclassification ECM are defined by allocating observation $\boldsymbol{x}$ to population π_k for which the cost of misclassification R_k $\sum_{i=1, i\neq k}^{g} p_i f_i(\boldsymbol{x}) c(k|i)$ is smallest.

$$R_k = \left\{ \boldsymbol{x} : \sum_{i=1, i\neq k}^{g} p_i f_i(\boldsymbol{x}) c(k|i) \leq \sum_{i=1, i\neq \ell}^{g} p_i f_i(\boldsymbol{x}) c(\ell|i), \quad \ell = 1, \cdots, g. \right\}$$

which can be written as

$R_k = \{\boldsymbol{x} : E(\text{cost of being misclassified into } \pi_k) \le E(\text{cost of being misclassified into } \pi_\ell) \text{ for any } \ell = 1, \cdots, g.\}$

In other words, an observation is classified into the class where the misclassification penalty is minimized.

- Comparison and consistency with two population classification

  Recall that in the case of two classes, the classification regions can be written as

$$R_1 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \ge \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \right\} = \left\{ \boldsymbol{x} : \ p_2 f_2(\boldsymbol{x}) c(1|2) \ \le \ p_1 f_1(\boldsymbol{x}) c(2|1) \ \right\}$$
$$R_2 = \left\{ \boldsymbol{x} : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1} \right\} = \left\{ \boldsymbol{x} : \ p_1 f_1(\boldsymbol{x}) c(2|1) \ < \ p_2 f_2(\boldsymbol{x}) c(1|2) \ \right\}$$

  They can be redescribed in terms of expected cost of misclassification of each class:

$$R_1 = \{\boldsymbol{x} : E(\text{cost of misclassified into } \pi_1) \le E(\text{cost of misclassified into } \pi_2)\}$$
$$R_2 = \{\boldsymbol{x} : E(\text{cost of misclassified into } \pi_2) \le E(\text{cost of misclassified into } \pi_1)\}$$

  which are consistent with the multiple population case.

- If a tie occurs, $\boldsymbol{x}$ can be assigned to any of the tied populations.

- If the misclassification costs $c(j|i), j \ne i$ are the same, then the classification rule is simplified.

  The region that an observation classified to π_k becomes

$$R_k = \left\{ \boldsymbol{x} : \sum_{i=1, i\ne k}^{g} p_i f_i(\boldsymbol{x}) \le \sum_{i=1, i\ne \ell}^{g} p_i f_i(\boldsymbol{x}), \quad \ell = 1, \cdots, g. \right\}$$

  which is equivalent to

$$R_k = \left\{ \boldsymbol{x} : \ p_k f_k(\boldsymbol{x}) \ge p_i f_i(\boldsymbol{x}), \quad for\ all \ \ i \ne k \right\}$$

  Sometimes it is more convenient to use the expression of logarithm:

$$R_k = \left\{ \boldsymbol{x} : \ \ln\left[p_k f_k(\boldsymbol{x})\right] \ge \ln\left[p_i f_i(\boldsymbol{x})\right], \quad for\ all \ \ i \ne k \right\} \tag{6}$$

- Equivalence to Bayesian maximum posterior probability classifier

  Bayesian approach allocates new observation $\boldsymbol{x}_o$ to the class with the largest posterior probability $P(\pi_i|\boldsymbol{x}_o)$. The Bayes' Rule for the posterior probability is

$$P(\pi_i|\boldsymbol{x}_o) = \frac{\mathbb{P}(\text{observe } \boldsymbol{x}_o, \text{ and } \boldsymbol{x}_o \text{ is from } \pi_i)}{\mathbb{P}(\text{observe } \boldsymbol{x}_o)} = \frac{\mathbb{P}(\text{observe } \boldsymbol{x}_o|\pi_i)\mathbb{P}(\pi_i)}{\sum_{j=1}^{g} \mathbb{P}(\text{observe } \boldsymbol{x}_o|\pi_j)\mathbb{P}(\pi_j)}$$

  The prior probability of being in class i is $\mathbb{P}(\pi_i) = p_i$. Therefore,

$$\text{Posterior probability that } \boldsymbol{x} \text{ is from } \boldsymbol{\pi}_k \text{ given that } \boldsymbol{x} \text{ is observed}$$
$$= P(\pi_i|\boldsymbol{x}_o) = \frac{p_i f_i(\boldsymbol{x}_o)}{\sum_{j=1}^{g} p_j f_j(\boldsymbol{x}_o)} = \frac{p_i f_i(\boldsymbol{x}_o)}{p_1 f_1(\boldsymbol{x}_o) + \cdots + p_g f_g(\boldsymbol{x}_o)}$$

As in developing classification rules and evaluations in the two population situation, the above definitions and results describe the basic principles of optimal classification functions. In applications, the quantities p_i and f_i are often not known given data. Estimation procedures have to be developed to apply these criteria.

5.2 Classification of more than two multivariate normal populations

Now consider the case when each subpopulation is of p-variate normal distribution with density $f_i(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_i, \Sigma_i)$.

The log classification regions in (6) have the form

$$R_k = \left\{ \boldsymbol{x} : \ \ln\left[p_k f_k(\boldsymbol{x})\right] \ge \ln\left[p_i f_i(\boldsymbol{x})\right], \quad for\ all \ \ i \ne k \right\}$$

The normal distribution corresponds to

$$\ln(p_i f_i(\boldsymbol{x})) = \ln(p_i) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)$$

from which comes the **quadratic discrimination score**

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln(p_i), \qquad i = 1, \cdots, g.$$

The classification regions in (6) becomes

$$R_k = \left\{ \boldsymbol{x} : \ d_k^Q(\boldsymbol{x}) \ge d_i^Q(\boldsymbol{x}), \quad for\ all \ \ i \ne k \right\} = \left\{ \boldsymbol{x} : \ d_k^Q(\boldsymbol{x}) = \max\{d_1^Q(\boldsymbol{x}), \cdots, d_g^Q(\boldsymbol{x})\} \right\}$$

Sample quadratic classification rule (> 2 normal samples, unequal variance)

In practice, $d_i^Q(\boldsymbol{x})$ is replaced by the sample estimate $\hat{d}_i^Q(\boldsymbol{x})$, and the theoretical values Σ_i and $\boldsymbol{\mu}_i$ in $d_i^Q(\boldsymbol{x})$ replaced by sample estimates $\boldsymbol{S}_i$ and $\bar{\boldsymbol{x}}_i$. The sample classification rule allocates $\boldsymbol{x} \to \pi_i$ if

$$\hat{d}_i^Q(\boldsymbol{x}) = \max\{\hat{d}_1^Q(\boldsymbol{x}), \cdots, \hat{d}_g^Q(\boldsymbol{x})\}$$

Equal-covariance multivariate normal distributions

If the g subpopulations have the same covariance matrix,

$$\Sigma_1 = \cdots = \Sigma_g = \Sigma$$

then

$$\begin{aligned} d_i^Q(\boldsymbol{x}) = & \ -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}\boldsymbol{x}\Sigma^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i\Sigma^{-1}\boldsymbol{\mu}_i + \ln(p_i) \\ = & \ C(\boldsymbol{x}, \Sigma) + \boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \ln(p_i) \end{aligned}$$

where

$$C(\boldsymbol{x}, \Sigma) = -\frac{1}{2}\ln|\Sigma| - \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x}$$

in $d_i^Q(\boldsymbol{x})$ is a common term for all i.

The remaining terms in $d_i^Q(\boldsymbol{x})$ are linear in $\boldsymbol{x}$, which leads to the **linear discriminant score**

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i + \ln(p_i), \qquad i = 1, \cdots, g.$$

The classification regions become

$$R_k = \left\{ \boldsymbol{x} : \ d_k(\boldsymbol{x}) = \max\{d_1(\boldsymbol{x}), \cdots, d_g(\boldsymbol{x})\} \right\}$$

Sample classification rule (for $g > 2$ normal classes, equal variance)

Using sample estimates, the sample classification regions are defined as

$$\hat{R}_k = \left\{ \boldsymbol{x} : \ \hat{d}_k(\boldsymbol{x}) = \max\{\hat{d}_1(\boldsymbol{x}), \cdots, \hat{d}_g(\boldsymbol{x})\} \right\}$$

where the kth linear discriminant score for an observation $\boldsymbol{x}$ is

$$\hat{d}_k(\boldsymbol{x}) = \bar{\boldsymbol{x}}_k' \boldsymbol{S}_{pool}^{-1} \boldsymbol{x} - \frac{1}{2} \bar{\boldsymbol{x}}_k' \boldsymbol{S}_{pool}^{-1} \bar{\boldsymbol{x}}_k + \ln(p_k), \qquad k = 1, \cdots, g,$$

and

$$\boldsymbol{S}_{pool} = \frac{1}{n-g} \Big[(n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_g - 1)\boldsymbol{S}_g \Big], \qquad n = n_1 + \cdots + n_g.$$

Boundaries of classification regions

The boundaries of classification regions can be obtained by equating pairwise discriminants and finding the intersection points. For example, the boundary of R_1 and R_2 consists of $\boldsymbol{x}$ satisfying $d_1(\boldsymbol{x}) = d_2(\boldsymbol{x})$.

To find all boundaries,

$$\text{Setting} \quad d_i(\boldsymbol{x}) = d_j(\boldsymbol{x}), \qquad i, j = 1, \cdots, g.$$

Note that it is equivalent to setting

$$d_i^Q(\boldsymbol{x}) = d_j^Q(\boldsymbol{x})$$

$$\implies \quad \binom{g}{2} \text{ hyperplanes forming the boundaries of the classical regions } R_1, \cdots, R_g \subset R^p.$$

The boundaries can be estimated by the sample discriminants $\hat{d}_k(\boldsymbol{x}), k - 1, \cdots, g$, thus form the borders of the sample classification regions as

$$\hat{d}_i(\boldsymbol{x}) = \hat{d}_j(\boldsymbol{x}), \qquad i, j = 1, \cdots, g.$$

Remarks The maximum property (6) guarantees the unambiguity of the classification regions described above.

Error rate estimations

Using the Lachenbruch's holdout procedure as in the two-population case, we may use the omit-one-at-a-time misclassification rates to estimate the expected actual error rate.

$$\hat{E}(AER) = \frac{\sum_{i=1}^{g} n_{im}^{(H)}}{\sum_{i=1}^{g} n_i}$$

5.3 Fisher's linear discriminants for serval populations

Fisher extended his linear discriminant function to several populations, still without the normality assumption that the populations are of multivariate normal distribution. However equal covariance assumption $\Sigma_1 = \cdots = \Sigma_g$ is implied, often also the full-rank assumption that $rank(\Sigma) = p$. Fisher's method can provide a useful low dimensional representation of the g populations.

Recall that Fisher's method for two populations is to find a univariate $y = \boldsymbol{a}'\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^p$, which maximally separates the two populations. The optimal vector $\boldsymbol{a}$ maximizes the sample ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{\text{between group variance}}{\text{within group variance}} \Big|_{y=\boldsymbol{a}\boldsymbol{x}} = \frac{(\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2))^2}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}}$$

The generalization of the method to more than two population still seeks an optimal direction to achieve maximum separation of the populations. Rewrite the original ratio as

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{\boldsymbol{a}'(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}} \tag{7}$$

To generalize, we show that we can rewrite the matrix in the numerator of (7) in a more generalizable form,

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' = 2\left\{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})'\right\} = 2\sum_{i=1}^{2}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})' = 2\boldsymbol{B}$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2).$$

and

$$\boldsymbol{B} = \sum_{i=1}^{g}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})' \qquad with \quad g = 2.$$

Proof.

First, note that we can write

$$\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_1 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2), \quad and \qquad \bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}_2 = \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) - \bar{\boldsymbol{x}}_2 = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

Thus

$$\begin{aligned}
&(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \\
=&\left[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}_2)\right]\left[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}_2)\right]' \\
=&(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})' + 2(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}} - \bar{\boldsymbol{x}}_2)' \\
=&(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})' + 2\frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \\
=&(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})' + \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'
\end{aligned}$$

$$\implies \quad (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})'$$

$$\implies \quad (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' = 2\left\{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})' + (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})'\right\}$$

□

Thus maximizing the ratio in (7) is equivalent to maximizing the more generalizable form

$$\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}} = \frac{\sum_{i=1}^{g}\boldsymbol{a}'(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})'\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{S}_{pool}\boldsymbol{a}} = \frac{\sum_{i=1}^{g}(\bar{y}_i - \bar{y})^2}{s_y^2}$$

Recall the summation form of the matrix $\boldsymbol{S}_{pool}$ in the denominator,

$$\boldsymbol{S}_{pool} = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}(n_i - 1)\boldsymbol{S}_i = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)' = \frac{1}{n-2}\boldsymbol{W}$$

For general $g > 2$ populations, the generalized version of (7) is proportional to the ratio

$$\frac{\boldsymbol{a}'\{\text{between group variance matrix}\}\boldsymbol{a}}{\boldsymbol{a}'\{\text{within group variance matrix}\}\boldsymbol{a}} = \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}}$$

where

$$\boldsymbol{B} = \sum_{i=1}^{g}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})'$$

is the **sample between groups matrix** (note the slight difference from the B matrix used in MANOVA), and

$$\boldsymbol{W} = \sum_{i=1}^{g}(n_i - 1)\boldsymbol{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'$$

is the **sample within groups matrix**. Here

$$\bar{\boldsymbol{x}} = \frac{1}{g}\sum_{i=1}^{g}\bar{\boldsymbol{x}}_i$$

is the **overall average vector**, or more descriptively, **sample-mean average** vector, which is the average of the individual subsample average of subpopulations.

Notice that, when the class sizes differ, it may happen that

$$\text{“sample mean average (overall average)”} = \frac{1}{g}\sum_{i=1}^{g}\bar{\boldsymbol{x}}_i \qquad \neq \qquad \frac{1}{n}\sum_{i=1}^{g} n_i\bar{\boldsymbol{x}}_i = \text{“grand average”}$$

That is, when the subsample sizes $n_1, \cdots, n_g$ are not equal, the above definition of **sample mean average** (overall average) $\frac{1}{g}\sum_{i=1}^{g}\bar{\boldsymbol{x}}_i$ can slightly differ from the **grand average** of all observations $\frac{1}{n}\sum_{i=1}^{g} n_i\bar{\boldsymbol{x}}_i$, where $n = \sum_{i=1}^{g} n_i$.

Solutions for Fisher's LDA for g classes

Now the objective is to maximize the ratio

$$\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}}$$

The vector $\boldsymbol{a}$ that maximizes the ratio is the eigenvector of the largest eigenvalue of matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$, by a result from the maximization inequality of quadratic forms (more details below).

The second, the third “best” directions $\boldsymbol{a}$'s can be found consecutively, if needed.

These vector $\boldsymbol{a}_i$'s are called **discriminant coordinates**. They give consecutive directions of maximum separation of the classes, they are not discriminant function themselves.

Another name for the $\boldsymbol{a}_i$'s is canonical variates, come from an alternative derivation via canonical correlation analysis (CCA) on predictor variable matrix and response variable matrix (omitted).

Fisher's sample linear discriminants

Let $\hat{\lambda}_1, \cdots, \hat{\lambda}_r > 0$ denote the nonzero eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}$, where

$$r = rank(\boldsymbol{W}^{-1}\boldsymbol{B}) \leq \min(g-1, p)$$

Let $\hat{\boldsymbol{e}}_i$ be the corresponding eigenvectors, $\boldsymbol{W}^{-1}\boldsymbol{B}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i\hat{\boldsymbol{e}}_i$, with normalization $\hat{\boldsymbol{e}}_i\boldsymbol{S}_{pool}\hat{\boldsymbol{e}}_i = 1$.

Then the vector of coefficient $\hat{\boldsymbol{a}}$ that maximizes the ratio

$$\frac{\hat{\boldsymbol{a}}'\boldsymbol{B}\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}'\boldsymbol{W}\hat{\boldsymbol{a}}}$$

is given by $\hat{\boldsymbol{a}} = \hat{\boldsymbol{e}}_1$, with $\boldsymbol{W}^{-1}\boldsymbol{B}\hat{\boldsymbol{e}}_1 = \hat{\lambda}_1\hat{\boldsymbol{e}}_1$. Thus

$$\max_{\boldsymbol{a}}\frac{\hat{\boldsymbol{a}}'\boldsymbol{B}\hat{\boldsymbol{a}}}{\hat{\boldsymbol{a}}'\boldsymbol{W}\hat{\boldsymbol{a}}} = \frac{\hat{\boldsymbol{e}}_1'\boldsymbol{B}\hat{\boldsymbol{e}}_1}{\hat{\boldsymbol{e}}_1'\boldsymbol{W}\hat{\boldsymbol{e}}_1} = \hat{\lambda}_1.$$

The p-vector $\boldsymbol{e}_1$ give the optimal direction to distinguish the g subpopulations. The direction maximizes the “signal to noise” ratio by maximizing the ratio of between group and within group variance.

The matrix $\boldsymbol{W}^{-1}\boldsymbol{B}$ has $r = rank(\boldsymbol{W}^{-1}\boldsymbol{B})$ non-zero eigenvalues $\hat{\lambda}_i$, each of corresponding eigenvector $\hat{\boldsymbol{e}}_i$, $\boldsymbol{W}^{-1}\boldsymbol{B}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i\hat{\boldsymbol{e}}_i$. For $\hat{\boldsymbol{a}}_i = \hat{\boldsymbol{e}}_i$ or any of its constant multiple,

$$\frac{\hat{\boldsymbol{a}}_i'\boldsymbol{B}\hat{\boldsymbol{a}}_i}{\hat{\boldsymbol{a}}_i'\boldsymbol{W}\hat{\boldsymbol{a}}_i} = \hat{\lambda}_i, \qquad i = 1, \cdots, r.$$

Therefore the above method can produce up to r linear discriminants.

Classify new observation using Fisher's sample linear discriminants

Although the main objective of Fisher's discriminant analysis was to separate the subpopulations, the linear discriminants can be used to form class boundaries and to classify new observations.

In the case of $g = 2$ classes, The allocation rule based on Fisher's linear discriminants is equivalent to assigning a new observation $\boldsymbol{x}_o$ to class i if the distance $|\hat{y}_o - \bar{y}_i|$ is smaller, where $i = 1, 2$, and the distance is often Euclidean distance.

This view can be generalized to $g > 2$: For $i = 1, \cdots, g$, a new observation $\boldsymbol{x}_o$ will be assigned to class i where $y_o = a'x_o$ is closest to the mean of class i. Using Euclidean distance, observation x_o is to be assigned to π_k if

$$|a'(x_o - \bar{x}_k)|^2 \leq |a'(x_o - \bar{x}_j)|^2, \quad for\ any\ j = 1, \cdots, g.$$

In the case of two discriminants, allocate x_o to π_k if

$$|a_1'(x_o - \bar{x}_k)|^2 + |a_2(x_o - \bar{x}_k)|^2 \leq |a_1'(x_o - \bar{x}_j)|^2 + |a_2'(x_o - \bar{x}_j)|^2, \quad for\ any\ j = 1, \cdots, g$$

In the case of r discriminants, allocate x_o to π_k if

$$\sum_{i=1}^{r}|a_i'(x_o - \bar{x}_k)|^2 \leq \sum_{i=1}^{r}|a_i'(x_o - \bar{x}_j)|^2, \quad for\ any\ j = 1, \cdots, g$$

The classification rule by Fisher's linear discriminant creates a Voronoi diagram type of partition of the sample space into p regions with respect to the p sample means.

Obtain more Fisher's sample discriminants

More discriminants can be obtained, for visualization or other purpose.

- First sample discriminant: $y_1 = \hat{\boldsymbol{a}}_1\boldsymbol{x} = \hat{\boldsymbol{e}}_1\boldsymbol{x}$
- Second sample discriminant: $y_2 = \hat{\boldsymbol{a}}_2\boldsymbol{x} = \hat{\boldsymbol{e}}_2\boldsymbol{x}$

- ……
- kth sample discriminant: $y_k = \hat{\boldsymbol{a}}_k \boldsymbol{x} = \hat{\boldsymbol{e}}_k \boldsymbol{x} \quad (k \le r)$
- $\boldsymbol{W}^{-1}\boldsymbol{B}\hat{\boldsymbol{a}}_i = \hat{\lambda}_i \hat{\boldsymbol{a}}_i, \quad \hat{\boldsymbol{a}}_i \boldsymbol{S}_{pool} \hat{\boldsymbol{a}}_k = \begin{cases} 1 & \text{if } i = k \le r \\ 0 & otherwise \end{cases}$

A low dimensional representation can be obtained using the first few discriminants.
For example, displaying the data on (y_1, y_2) plane aids visual inspection of distinct classes.

Remarks on Fisher's linear discriminant functions:

- Analogous to the dimension-deduction role of principal components in PCA, hopefully the first couple of sample discriminants show a good separation of the classes and can be used to assign classes to new observations.
- Geometrically, Fisher's linear discriminant function aims to find consecutive (orthogonal) directions that maximized the normalized covariance between groups.
- A key assumption is that the subpopulations are of equal covariance matrix, which could often be violated in practice, albeit a relatively robust assumption.
- Normal distribution is not assumed.
- In LDA, some type of difference in the component X's is utilized as part of the discrimination.

  This part is comparable to MANOVA and linear regression models, where the response Y is modeled as a linear combination of features, or X's.

  In LDA for classification, the response Y, the index of the classes, is categorical.
- LDA, PCA and FA all seek linear combination of X's that best explains some characteristics of the data.

  PCA seeks directions representing maximum variation of the data as a whole.

  LDA looks for the difference between the classes, treating X as predictors.

  FA treats X's as dependent and interdependent variables, creating unobservable explanatory variables.

Appendix* — Remarks on maximum inequality

In the derivation of Fisher's discriminants and in a few other derivations in this course, we applied the **maximization inequality of quadratic forms**,

$$\max_{x \in \mathbb{R}^p} \frac{x'Cx}{x'x} = u_1' C u_1 = \lambda_1, \tag{8}$$
$$\min_{x \in \mathbb{R}^p} \frac{x'Cx}{x'x} = u_p' C u_p = \lambda_p.$$

where

- $C = C'$ is any $p \times p$ symmetric, positive semi-definite matrix,
- $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$ are eigenvalues of C,
- u_i's are orthonormal eigenvectors such that

  $$Cu_i = \lambda_i u_i, \qquad for \quad i = 1, \cdots, p$$

  and

  $$u_i^T u_j = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

Notes

- The above results are stated in the lecture notes Matrix Algebra Basics, as corollaries of the Maximization Lemma, which is an extension of Cauchy-Schwarz Inequality.
- Application in LDA

  We may claim that, by (8),

  $$\max_a \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}} = \lambda_1, \qquad \min_a \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}} = \lambda_p$$

  are achieved when $\boldsymbol{a}$ are constant multiples of e_1 and e_p, where e_i's are eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$ with eigenvalue λ_i, and the e_i's are scaled by $e_i'\boldsymbol{W}e_j = \delta_{ij}$.

  Proof.

  Assume that the $p \times p$ matrix $\boldsymbol{W}$ is symmetric positive definite. Then $\boldsymbol{W}^{-1}$ exists, and $\boldsymbol{W}$ can be expressed as the square product of some symmetric positive semi-definite matrix, which we may call a root-matrix of $\boldsymbol{W}$ and denote it as $\boldsymbol{W}^{1/2}$. So we have

  $$\boldsymbol{W} = \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2} \qquad \boldsymbol{W}^{-1} = \boldsymbol{W}^{-1/2}\boldsymbol{W}^{-1/2}$$

  Define

  $$x = \boldsymbol{W}^{1/2}\boldsymbol{a}, \qquad C = \boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2}$$

  Since $\boldsymbol{B}$ is symmetric and positive semi-definite, using its root-matrix analogously, it follows that matrix C is also symmetric and positive semi-definite (exercise).

Then we can express the ratio in Fisher's discriminant derivation as

$$\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}} = \frac{x'Cx}{x'x}$$

The maximum and the minimum of the ratio are λ_1 and λ_p, which are the largest and smallest eigenvalues of the matrix $C = \boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2}$.

The maximum and the minimum are achieved when vector $x = \boldsymbol{W}^{1/2}\boldsymbol{a} = v_1$ and v_p, where $v_i \in \mathbb{R}^p, i = 1, \cdots, p$ are orthonormal eigenvectors of C corresponding to eigenvalues $\lambda_1, \cdots, \lambda_p$.
In other words, the maximum and the minimum of the ratio $\frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{W}\boldsymbol{a}}$ are achieved when
$\boldsymbol{a} = \boldsymbol{W}^{-1/2}v_1$ and $\boldsymbol{a} = \boldsymbol{W}^{-1/2}v_p$.

Define $e_i = \boldsymbol{W}^{-1/2}v_i$, then

$$(\boldsymbol{W}^{-1}\boldsymbol{B})e_i = \boldsymbol{W}^{-1/2}(\boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2})(\boldsymbol{W}^{1/2}e_i) = \boldsymbol{W}^{-1/2}Cv_i = \boldsymbol{W}^{-1/2}\lambda_i v_i = \lambda_i e_i, \quad i = 1, \cdots, p.$$

Therefore e_i's are eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$ with eigenvalues λ_i,

$$\boldsymbol{W}^{-1}\boldsymbol{B}e_i = \lambda_i e_i, \qquad for \quad i = 1, \cdots, p$$

Note that, the eigenvectors are normalized with respect to the $\boldsymbol{W}$ matrix,

$$e_i'\boldsymbol{W}e_j = v_i'v_j = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \qquad i, j = 1, \cdots, p.$$

□

- Recall $\boldsymbol{W} = (n-g)\boldsymbol{S}_{pool}$. In practice the scaling may become $e_i'\boldsymbol{S}_{pool}e_j = \delta_{ij}$ in the place of $e_i'\boldsymbol{W}e_j = \delta_{ij}$.
- The ratio is also called Rayleigh quotient, where the matrix can be a complex Hermitian matrix.
- In some software, a multiplicative factor may be added to the $e_i'\boldsymbol{S}_{pool}e_j = \delta_{ij}$ scaling normalization.

6 The Nearest-Neighbor Classifier

Now we take a look at a classifier with classification rules very different from the ones developed under multivariate normal distribution theory.

Nearest Neighbor classifier

The nearest-neighbor classifier is simple yet effective in practice.

The classification rule consists of the following two steps.

1. Define distance function on the input variables.

   $$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \qquad \boldsymbol{x}_i \in \mathbb{R}^p.$$

   For example, Euclidean distance can be used with component variables as axes.

   $$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

2. The Nearest Neighbor classifier h classifies a new input $\boldsymbol{x}_o$ by looking at the class label y of the closest sample point in the training dataset, then copy the label.

   $$h(\boldsymbol{x}_o) = \hat{y}(\boldsymbol{x}_o) = y(\boldsymbol{x}^*), \qquad \boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \; d(\boldsymbol{x} - \boldsymbol{x}_o)$$

A Problem

However this simple rule is easy to overfit, that is, the pattern in training set is followed closely, including noise.
For example, assume we know that the two classes should have a linear boundary, and the data are measured with some random noise. Then the one-nearest neighbor rule is likely to follow the noise closely, producing a wiggly boundary.

A Solution

To smooth the boundary, k nearest neighbors, instead of just one neighbor, and a majority rule can be used to make an assignment of a new observation.

Modify the above Step 2, in its place use

2*. Classify a new input $\boldsymbol{x}_o$ by looking at the class labels of the closest k sample points in the training dataset, take a vote, then use the most frequent label among the k nearest neighbors.

This is the **k-Nearest Neighbor** classifier, known as **kNN**. The classification function for a new observation $\boldsymbol{x}_o$ is

$$h(\boldsymbol{x}_o) = \hat{y}(\boldsymbol{x}_o) = \text{most frequent class label among } k \text{ closest training data point to } \boldsymbol{x}_o$$

In the case of two classes, let $y(\boldsymbol{x}) = 1, -1$ be the class membership function of input $\boldsymbol{x}$.
For a new observation $\boldsymbol{x}_o$, the kNN estimator $h(\boldsymbol{x}_o)$ assigned class label can be expressed as the sign of average of the k labels.

$$h(\boldsymbol{x}_o) = \hat{y}(\boldsymbol{x}_o) = sign\left\{\frac{1}{k}\sum_{\boldsymbol{x}_i \in N_k(\boldsymbol{x}_o)} y_(\boldsymbol{x}_i)\right\}$$

where the sign function is defined as

$$sign(a) = \begin{cases} 1, & a > 0 \\ 0, & a = 0 \\ -1, & a < 0 \end{cases}$$

How to choose k

Validation data are used to decide k.

- Split sample data into training set and validation set.
  For example, allocate 70% of the data as the training set, 30% as the validation set.
- Training kNN on training set for different values of k.
  For each k, using k-NN on the training set, using the learned rules to classify data in the validation set.
  Obtain performance evaluation, e.g., classification error rate, for each k.
- Pick the k with the best performance on the validation set.
- Random splitting and training may repeat in order to obtain estimation variance.

Remarks on kNN

- The smooth or regularization parameter k in kNN is hyper-parameter with range $k < n_i, i = 1, \cdots, g$.
- The method is simple, good performance in practice, with reasonably good interpretability.
- KNN is a non-linear non-parametric method, can be used to express complex, non-linear, non-parametric boundaries. The complexity of classification boundary increases with the number of data points.
- KNN is sensitive to the scale of inputs. It is common to scale each component variable to a fixed range $[0, 1]$ or $[-1, 1]$, or to standardize to mean 0 variance 1.
- KNN is an instance-based learning algorithm. Instead of using explicit generalization, kNN compares new instances with instances seen in training, which have been stored in memory.
- KNN has large memory requirement for prediction, and not among the best classifiers in terms of accuracy.

7 SVM - Support vector machines

SVM is a classification method with two outstanding characters:

- Maximizing classification margin.
- Readily generalizable using the kernel method.

Distance of a point to a line

In $\mathbb{R}^p$, the distance of a point x_o to a line or linear hyperplane $w'x + b = 0$ can be written as

$$d = d(x_o, L) = \frac{|w'x_o + b|}{\|w\|} \tag{9}$$

where $|\cdot|$ is the absolute value of a scalar, $\|\cdot\|$ is the vector norm, here we consider the Euclidean norm, also called the 2-norm.

Proof.

Given any vector x with its endpoint on the line $L = \{x : w'x + b = 0\}$, the distance of vector $x_o \in \mathbb{R}^p$ to L is the length of the projection of vector $x_o - x$ to vector $w/\|w\|$, the unit normal vector of the line L.

$$\begin{aligned} d(x_o, L) &= |(x_o - x)'w|/\|w\| = |x_0'w - x'w|/\|w\| \\ &= |x_0'w - (-b)|/\|w\| = |x_0'w + b|/\|w\| \\ &= |w'x_o + b|/\|w\| \end{aligned}$$

In the above we used the fact that, for any vector x^* with endpoint on the line $w'x + b = 0$, we have $w'x^* + b = 0$, and $w'x^* = x^{*\prime}w = -b$. □

Signed distance to a line

The signed distance of a vector x to a line $w'x + b = 0$ is defined as

$$\frac{w'x + b}{\|w\|} \tag{10}$$

which is also called **directional distance** of point x to line $w'x + b = 0$.

In higher dimensional space with $x \in \mathbb{R}^p$, $p > 2$, the equation $w'x + b = 0$ represent a hyperplane.

Exercise: The distance of a point to a hyperplane in $\mathbb{R}^p$ has an expression analogous to (9).

SVM for linear separable 2-classes

First we consider the simplest separable case.

Assume there exists a linear classifier, that is, there is a line or hyperplane that completely separate the points in two classes.

SVM aims for the linear classifier maximizing the margin between the two classes.

SVM classifier formulation

Denote the class label of a training point x as y, with values $y = 1$ or $y = -1$.

The SVM classification hyperplane H: $w'x + b = 0$ should have the following properties.

- H divides the two classes.
- $w'x + b > 0$ for x in class $y = 1$, and $w'x + b < 0$ for $y = -1$.
- There is $c > 0$, such that there are supporting vectors on the margin hyperplanes $w'x + b = \pm c$:

  There are vectors x with $w'x + b = c, y = 1$, and
  there are vector x with $w'x + b = -c, y = -1$.
- Other vectors x should have $|w'x + b| > c$.

Conventional parameterization

The margin hyperplanes $w'x+b = \pm c$ is equivalent to $(w/c)'x+b/c = \pm 1$, which can be written as $w^{*\prime}x+b^* = \pm 1$.

We can rescale the vectors to express the margin hyperplanes as $w'x+b = \pm 1$. Then the SVM formulation becomes

$$w'x + b \begin{cases} \geq 1, & y = 1 \\ \leq -1, & y = -1 \end{cases}$$

Combine the two inequalities, the SVM classifier can be stated as

$$y(w'x + b) \geq 1 \tag{11}$$

with the objective to maximize the margin.

Margin size

If
x_1 is a supporting vector with $w'x_1 + b = 1$,
x_2 is a supporting vector with $w'x_2 + b = -1$,

then
the distance between the two margin hyperplane $w'x + b = \pm 1$ is

$$\left| \frac{w'}{\|w\|}(x_2 - x_1) \right| = \frac{|w'x_2 - w'x_1|}{\|w\|} = \frac{|(1-b) - (-1-b)|}{\|w\|} = \frac{2}{\|w\|}$$

This is the quantity that the SVM aims to maximize.

The Primal problem

To determine the SVM classifier parameters w and b is to minimize the margin of the linear classifier $2/\|w\|$, under the constraints $y_i(w^t x_i + b) \geq 1$, for $i = 1, \cdots n$, for each of the n training points.
(In fact, contributing cases are $y_i(w^t x_i + b) = 1$, the points on the margins, the supporting vectors.)

Maximizing the margin size $\frac{2}{\|w\|}$ is equivalent to minimizing $\|w\|$ or $\|w\|^2/2$. The optimization problem can be stated as

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \qquad \text{subject to} \qquad y_i(w^t x_i + b) \geq 1, \ \forall i = 1, \cdots, n. \tag{12}$$

Using the Lagrange multiplier method, we write

$$L = \|w\|^2/2 - \sum_{i=1}^{n} \lambda_i [y_i(w^t x_i + b) - 1]$$

Local extreme points should be critical points with partial derivatives zero.

$$\frac{\partial L}{\partial w} = 0_p \qquad \Longrightarrow \quad w - \sum_{i=1}^{n} \lambda_i y_i x_i = 0_p$$

$$\frac{\partial L}{\partial b} = 0 \qquad \Longrightarrow \quad \sum_{i=1}^{n} \lambda_i y_i = 0$$

The dual problem*

The primal problem is difficult to solve.
However the optimization problem of $\min_w L$ can be restated as a dual problem (derivation omitted).

The original primal optimization becomes

$$\max_{\lambda} \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j \right), \quad s.t. \quad \sum_{i=1}^{n} \lambda_i y_i = 0, \quad \lambda_i \geq 0.$$

which is called the dual problem.

Remarks on the dual problem

- The dual problem is a quadratic linear optimization problem, doable.
- The optimization achieves at the boundaries,

  $$\lambda_i [y_i(w^T x_i + b) - 1] = 0, \qquad i = 1, \cdots, n. \tag{13}$$

  (Note: These so-called KarushKuhnTucker (KKT) conditions are first-order necessary conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.)
- To satisfy $\lambda_i [y_i(w^T x_i + b) - 1] = 0$, we see that only support vectors corresponding to $\lambda_i \neq 0$, therefore contributing to the decision of max margin.
- In the dual problem, the input x's occur only in the inner product form

  $$\langle x_i, x_j \rangle = x_i^T x_j$$

  Which leads to the important tool of kernel method in SVM.

Kernel methods in SVM*

- Often nonlinear classification problems, when the classes are not linearly separable in the original feature space, can become linearly separable if the features are viewed as in a properly chosen higher dimensional space.

  A mapping

  $$x \in \mathbb{R}^p \to \psi(x) \in \mathbb{R}^q, \qquad q > p.$$

  can be constructed so that the nonlinear classification in the space of $x \in \mathbb{R}^p$ becomes a linear separable classification in the space of $\phi(x) \in R^q \subset \mathbb{R}^p$.

  As a simple example, consider the case that the two classes are

  $$C_1 = \{(x_1, x_2), x_1^2 + x_2^2 < 1\}, \qquad C_2 = \{(x_1, x_2), x_1^2 + x_2^2 > 1\}$$

  The class boundary is a circle, then the two classes C_1, C_2 are not linearly separable in the feature space $x = (x_1, x_2) \in \mathbb{R}^2$.

  To separate the two classes by a linear boundary in a higher dimensional space, construct a mapping

  $$\psi(x) = \psi(x_1, x_2) = (x_1,\ x_2,\ x_1^2 + x_2^2) = (x_1,\ x_2,\ x_3) = x^*$$

  The mapping is from a lower dimensional space to a higher dimensional one.

  $$\psi: \quad \mathbb{R}^2 \quad \to \quad \mathbb{R}^3$$

  Now the two classes is linearly separable in $\mathbb{R}^3$ by the hyperplane $w'x + b = 0$ with $w = [0\ 0\ 1]', b = -1$.

- In SVM, a mapping is constructed to simplify or to linearize the classification problem. Since the x's in the optimization condition is in the form of inner product $\langle x_i, x_j \rangle = x_i' x_j$, the mapping is selected so that

  $$\langle \psi(x_i), \psi(x_j) \rangle = \langle x_i, x_j \rangle$$

  Then the dual problem can be restated in the higher dimensional space $\mathbb{R}^q$, and only $\langle \psi(x_i), \psi(x_j) \rangle$ need to be calculated, not a full-ranged $\psi(x)$.

- The $\psi(x)$ function is called a **kernel**. Therefore by replacing the inner product $\langle x_i, x_j \rangle$ with a kernel in a higher dimensional space, SVM can linearize a non-linear boundary in the original, lower dimensional feature space.

SVM linear non-separable case*

In real world classification problems, perfect separation of classes are often impossible and unnecessary. A classification rule will allow misclassifications to a tolerable degree.

When the classes overlap in input feature variable space, there are no hyperplanes that can separate the classes perfectly.

One way to deal with the problem is to allow some points on the wrong side of the separating boundary.

The distance of the points on the wrong side to the margin hyperplane should be minimized. These margins with tolerance are called **soft margins**.

Formulation of soft margin with slack variable ξ *

For each feature point on the wrong side of the separation margins, let ξ denote the distance of the point to its margin. The objective is to

$$minimize_{w,b,\xi_i} \left(\|w\|^2/2 + C \sum_{i=1}^{n} \xi_i \right)$$

under the constraints

$$y_i(w^T x_i + b) \geq 1 - \sum_{i1}^{n} \xi_i, \qquad \xi_i \geq 0, \quad i = 1, \cdots, n.$$

The slack variable ξ_i allows the input point x_i to be closer to the hyperplane or even be on the wrong side. However there is a penalty in the objective function for allowing such slack.

If C is very large, the SVM becomes very strict and tries to get all points to be on the right side of the hyperplane. If C is very small, the SVM becomes very loose and may sacrifice some points to be on the wrong side to obtain a simpler solution.

Remarks about basic SVM

- Linear boundaries with some optimal-separation theoretical properties.
- Transform to higher dimensions to obtain linearly separation (kernel function).
- Based on a theoretical model of learning explicitly, with guaranteed performance.
- Not affected by local minima.
- Do not suffer from the curse of dimensionality.
- Quadratic program, doable.
- Optimization algorithm instead of greedy search.
- The kernel function has to be handpicked.
- Integratable into other high performers such as deep neural network.