Statistical Pattern Recognition

Assoc. Prof. Dr. Sathit Intajag
Faculty of Engineering
KMITL

Introduction

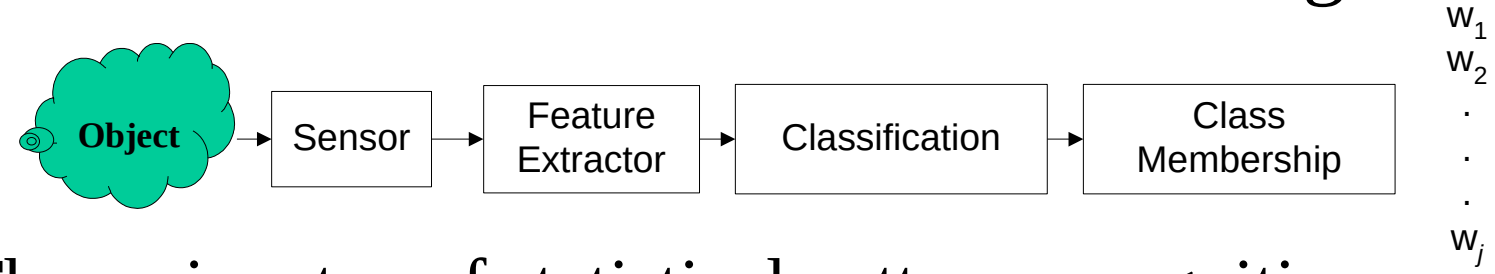
- Statistical pattern recognition is an application in computational statistics to **decide** the weighting density of data and to **recognize** (**Classify**).
- Applications of the statistical pattern recognition, such as:
  - A medical image: a doctor **diagnoses** a patient's illness based on the symptoms and test results.
  - A remote sensing: including a lot of works such as natural resource, a military, an agriculture, weather etc.
  - A financial: a loan manager at a bank must **decide** whether a customer is a good credit risk based on their income, past credit history and other variables.
  - A quality Control: a manufacturer must **classify** the quality of materials before using them in their products.
  - Etc.

Topics

- Statistical Inference: Bayes classifiers and pattern recognition in an hypothesis testing.
- Evaluation of the classifier.
- Clustering or unsupervised classification.

Supervised Learning

- The observations to recognize must be know which class each feature belongs to.



The major step of statistical pattern recognition.

1. Selecting features to distinguish between the classes, each observed set of feature measurements,
2. Training sets: {{input features}, {no. of classes}},
3. Classification methods are used to create weights from the training sets.

Iris Data

- Iris data is a standard data that use to test the classification algorithm.
- The data consist of three species of iris: Iris setosa, Iris versicolor and Iris virginica. These data were used by Fisher to develop a classifier.
- Four features use to distinguish the species of iris are sepal length, sepal width, petal length, and petal width.

Bayes Decision Theory

Bayes decision theory assumes the classification problem in terms of probabilities; therefore, all of the probabilities must be known or estimated from the data.

Let the class membership be represented by ω_j, j=1,...,J for a total of J classes, such as Iris data, J=3 classes:

ω_1 : Iris setosa

ω_2 : Iris versicolor

ω_3 : Iris virginica

The Iris Features

The features using for classification are denoted by *d*-dimensional vector **x**, *d*=1,2,...

From the iris data,we have four measurements, so $d = 4$.

In supervised learning method, the classification method has four inputs and three outputs.

Posterior Probability

The posterior probability is given by

$$P(\omega_j|\mathbf{x}). \qquad (1)$$

Eq. (1) represents the probability that the case belong to j-th class given the observed feature vector **x**.

To use this rule, we would evaluate all of the posterior probabilities, and the one with the highest prob. would be the class we choose.

By Bayes' theorem, the posterior prob. is defined by

$$P(\omega_j|\mathbf{x}) = \frac{P(\omega_j)P(\mathbf{x}|\omega_j)}{P(\mathbf{x})} \qquad (2)$$

where

$$P(\mathbf{x}) = \sum_{j=1}^{J} P(\omega_j)P(\mathbf{x}|\omega_j) \qquad (3)$$

Prior Probability

Form Eq. (2), we must know the **prior prob**. that it would be in class j given by

$$P(\omega_j); \qquad j=1,2,3,...,J. \qquad (4)$$

$P(\omega_j)$ represents the density assigned to ω_j before observing the data.

It may be based on previous data and analyses (e.g., pilot studies),

it may represent a purely subjective personal belief, or it may be chosen in a way intended to have limited influence on final inference.

Class-Conditional Probability

The class-conditional prob. (state-conditional prob.)

$$P(\mathbf{X}|\omega_j); \quad j=1,2,3,...,J, \qquad (5)$$

$P(\mathbf{X}|\omega_j)$ represents the prob. distribution of the features for each class.

The process of estimating both the class-conditional and prior probabilities is how we build the classifier.

Constructing the Classifier

- Define prior probabilities

  These can either be inferred from prior knowledge of the application, estimated from the data or assume to be equal.

- Estimating class-conditional probabilities:
  - Parametric method

    This method a distribution for the class-conditional probability densities is assumed or estimated them by estimating the corresponding distribution parameters.

  - Nonparametric method

    This method includes the averaged shifted histogram, the frequency polygon, kernel densities, finite mixtures and adaptive mixtures, etc.

Bayes Decision Rule

When we have the classifier; then, we can use Bayes' theorem to obtain posterior probabilities.

Bayes Decision Rule:

Given a feature vector **X,** assign it to class ω_j if

$$P(\omega_j|\,\mathbf{X}) > P(\omega_i|\,\mathbf{X}); \qquad i=1,2,3,...,J;\ i\neq j. \quad (6)$$

This state an observation **X** is classified belonging to class that has the highest posterior probability.

Alternative Decision Rule

From Eq. (2), we have

$$P(\omega_j|\mathbf{X}) = \frac{P(\omega_j)P(\mathbf{X}|\omega_j)}{P(\mathbf{X})}$$

$$P(\omega_i|\mathbf{X}) = \frac{P(\omega_i)P(\mathbf{X}|\omega_i)}{P(\mathbf{X})}$$

and from the decision rule, $P(\omega_j|\mathbf{X}) > P(\omega_i|\mathbf{X})$, so that

$$P(\omega_j)P(\mathbf{X}|\omega_j) > P(\omega_i)P(\mathbf{X}|\omega_i) \qquad (7)$$

Minimum Error

It is known [1] that the decision rule given by Eq. (6) yields a classifier with the minimum probability of error.

An error is made when we classify an observation as class ω_i when it is really in the j-th class.

To get the probability of error, we calculate the following integral over all values of **X**.

$$P(error) = \sum_{i=1}^{J} \int_{\Omega_i^c} P(\mathbf{x}|\omega_i) P(\omega_i) d\mathbf{x}. \qquad (8)$$

[1] Duda O. Richard and Peter E. Hart, Pattern Classification and Scene Analysis, New York: John Wiley & Son, 1973.

Ex. Bayes Decision Rule

Look at a univariate classification problem of two classes. The class-conditionals are given by the normal distributions as follows:

$P(x|\omega_1) = \phi(\mathbf{X}; -1, 1)$

$P(x|\omega_2) = \phi(\mathbf{X}; 1, 1)$.

The prior are

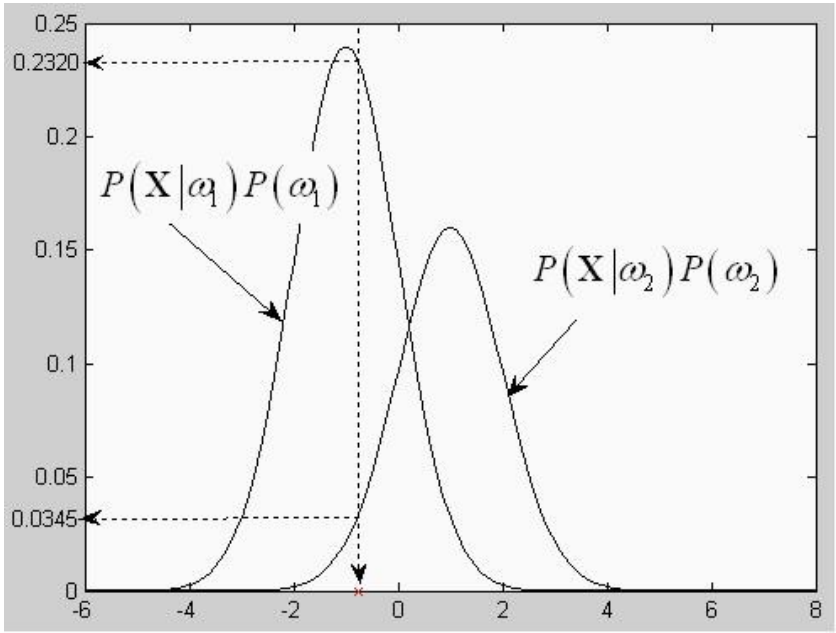
$P(\omega_1) = 0.6$

$P(\omega_2) = 0.4$.

If x = -0.75

$P(-0.75|\omega_1)\ P(\omega_1)=0.2320$

$P(-0.75|\omega_2)\ P(\omega_2)=0.0354$.



Cont. Ex.

If we change the decision boundary, then the error will be greater, illustrating that Bayes Decision Rule in one that minimize the probability of misclassification.

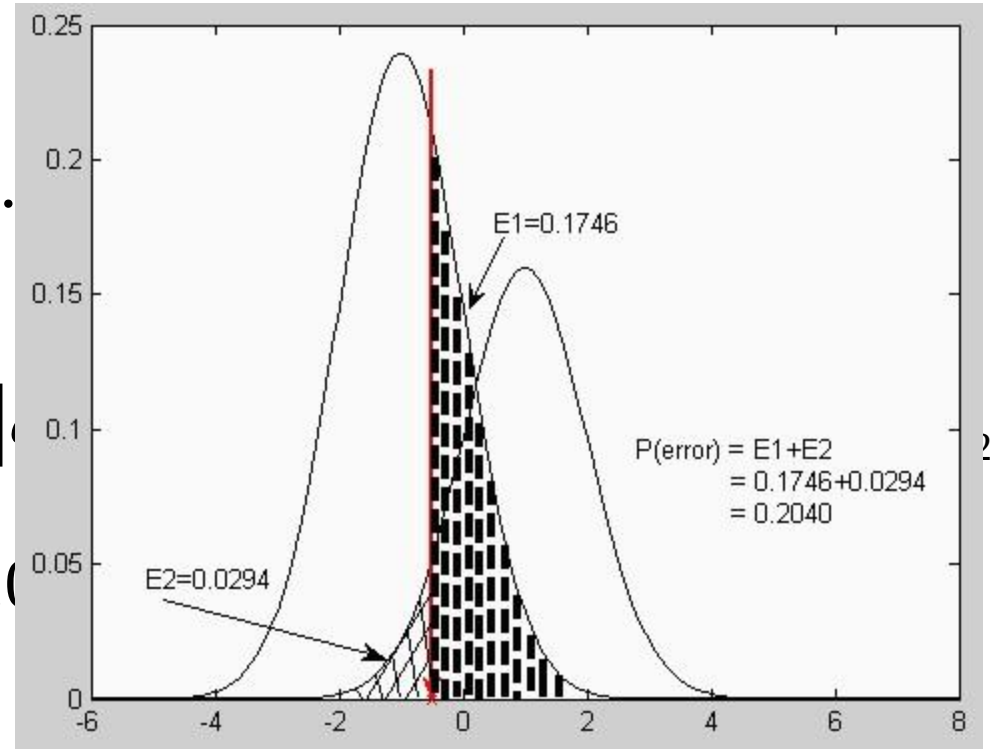
Such as x = -0.5.

$$P(error) = \sum_{i=1}^{2} \int_{\Omega_i^c} P(-0.$$

$$= \int_{-0.5}^{\infty} P(-0.5|$$

$$= 0.1746 + 0.0$$

$$)d\mathbf{X}$$



Likelihood Ratio Approach

The likelihood ratio technique address the issue of variable misclassification costs in a hypothesis testing framework.

This method does not assign an explicit cost to make an error as in the Bayes approach, but it enables us to set the amount of error we will tolerate for misclassifying one of the classes.

Ex. Of two Classes in Likelihood Ratio

First we determine a class which corresponds to the null hypothesis and call this the **non-target class, ω_2**. The other class is denoted as the **target class, ω_1**.

In a military command, taking features from images of objects and classify them as targets or non-targets. If an object is classified as a target (Tanks or Military Trucks), then we will destroy it. Non-target objects are such things as school buses or automobiles etc.

H_0: Object is a school bus, automobile, etc.

H_1: Object is a tank, military vehicle, etc.

False Alarms

Error from classification process of pattern recognition is called as **false alarm** or **false positives**. It is wrongly classifying something as a target ($\boldsymbol{\omega_1}$), when it should be classified as non-target ($\boldsymbol{\omega_2}$).

The probability of making a false alarm (or the probability of making a Type I error) is denoted as

$$P(FA) = \alpha.$$

False Alarms (continue)

Bayes Decision Rule gives a rule that yields the minimum probability of incorrectly classifying observed patterns. We can change this rule to obtain the desired probability of false alarm.

In two class case, we can put our Bayes Decision Rule in a different form. From Eq. (7), we have our decision as

$$P(\mathbf{X}|\omega_1)P(\omega_1) > P(\mathbf{X}|\omega_2)P(\omega_2) \Rightarrow \mathbf{x} \text{ is in } \omega_1, \qquad (9)$$

or else we classify **X** as belonging to ω_2.

Likelihood Ratio

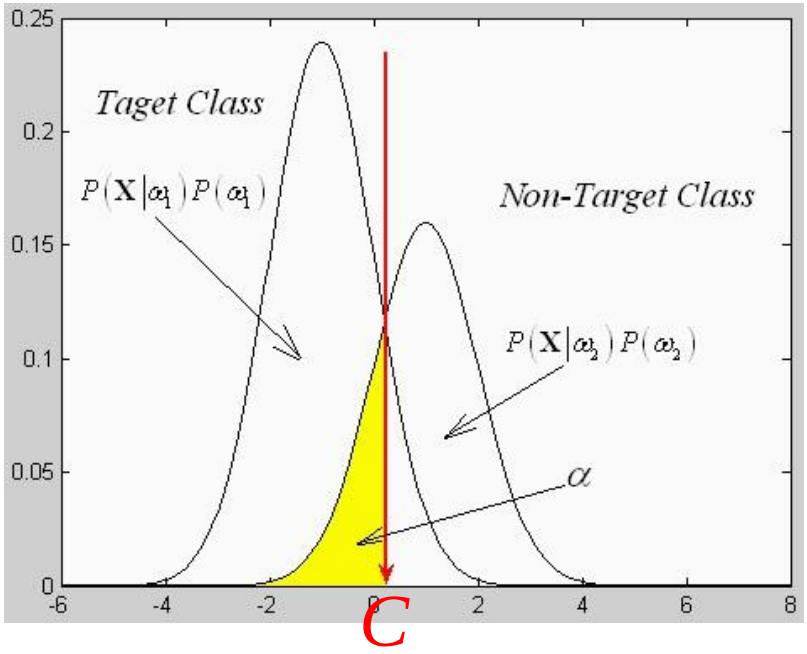
Rearranging Eq. (9) yields the following decision rule

$$L_R = \frac{P(\mathbf{X}|\omega_1)}{P(\mathbf{X}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} = \tau_C \Rightarrow \mathbf{X} \text{ is in } \omega_1. \qquad (10)$$

The ratio of this Eq. is called the **likelihood ratio**, and τ_C is the **threshold**.

If $L_R > \tau_C$, then we decide that the case belongs to class ω_1. If $L_R < \tau_C$, then the observation is classified to ω_2.

Ex. of False Alarm



From figure, false alarm is

$$P(\alpha) = \int_{-\infty}^{C} P(x|\omega_2)P(\omega_2)\,dx$$

$$= P(\omega_2)\int_{-\infty}^{C} P(x|\omega_2)dx$$

We then have to find the value for C such that

$$\int_{-\infty}^{C} P(x|\omega_2)\,dx = \frac{P(FA)}{P(\omega_2)}.$$

If we know that $P(\omega_2)$, such as $P(\omega_2)=0.4$ and $P(x|\omega_2)\sim N(1, 1)$. If our desired $P(FA) = 0.05$, then

$$\int_{-\infty}^{C} P(x|\omega_2)\,dx = \frac{0.05}{0.40} = 0.125.$$

Evaluating The Classifier

Classifier evaluation is usually measured by the percentage of observations that we correctly classify.

This yields an estimate of the probability of correctly classifying cases.

It is also important to report the probability of false alarms, when the application requires it (when there is a target class).

Hereinafter, two methods for estimating the probability of correctly classifying cases and the probability of false alarm, that are the use of an **independent test sample** and **cross-validation**, are described.

Independent Test Sample

This method needs to use a large sample. The sample is divided into **a training set** and **a testing set**.

The training set is used to build a classifier and **the testing set** is used to verify the classifier.

The proportion of correctly classified observations is the ***estimated classification rate.***

Note that if the classifier has not seen the patterns in the test set, then the classification rate estimated in this way is **unbiased**.

Probability of Correct Classification- Independent Test Sample

1. Randomly separate the sample into two sets of size n_{TEST} and n_{TRAIN}, where $n_{TEST} + n_{TRAIN} = n$.
2. Build the classifier using the training set.
3. Present each pattern from the test set to the classifier and count the number of correct class (N_{CC}).
4. The Probability of Correct Classification, $P(CC) = N_{CC}/n_{TEST}$.

The higher this proportion, the better the classifier.

Cross-Validation

The concept of cross-validation is the same as the independent test sample; **whereas, the cross-validation provides to relatively small data set.**

The data are separated into testing sets of size *k*. The *n*-*k* observations are used to build the classifier, and the remaining *k* partitions are used for test it.

Probability of Correct Classification –Cross-Validation

1. Set the number of correctly classified patterns to 0, $N_{CC} = 0$.
2. Keep out one observation, call it, $\mathbf{X}_i$.
3. Build the classifier using the remaining n-1 observations.
4. Present the observation $\mathbf{X}_i$ to the classifier and obtain a class label using the classifier from the previous step.
5. If the class label is correct, then increment the number correctly classified using $N_{CC} = N_{CC} + 1$
6. Repeat steps 2 through 5 for each pattern in the sample.
7. The probability of correctly classifying an observation is given by $P(CC) = N_{CC}/n$.

Using Cross-Validation

How to use cross-validation to evaluate a classifier by using the likelihood method with varying decision thresholds τ_C is described.

It would be useful to understand how the classifier performs for various thresholds of the likelihood ratio.

This will tell us what performance degradation we have if we limit the probability of false alarm to some level.

Cont.

We return to two classes by dividing the sample into two sets as following

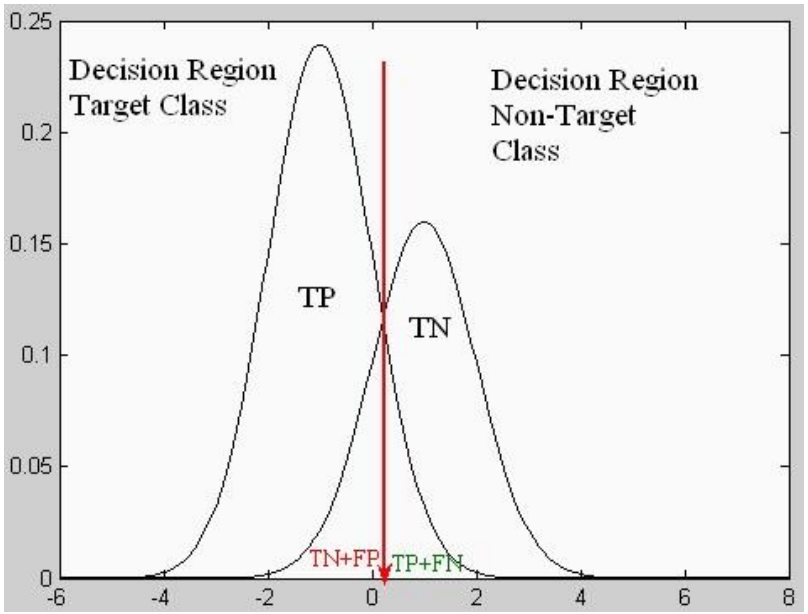
$\mathbf{X1}_i$: Target pattern (ω_1, with the number of observation n_1)

$\mathbf{X2}_i$: Non-target pattern (ω_2, with the number of observation n_2).

Some terminology for any boundary we might set for the decision regions of making mistakes in classifying cases.

- True Positive (TP) is the fraction of patterns correctly classified as target cases.
- False Positive (FP) is the fraction of non-target patterns incorrectly classified as target cases.
- True Negative (TN) is the fraction of non-target cases correctly classified as non-target.
- False Negative (FN) is the fraction of target cases incorrectly classified as non-target.

Cont.



There will be some target patterns that we correctly classify as target (TP) and some we misclassify as non-targets (FN).

Similarly, there will be non-target patterns that are correctly classified as non-targets (TN) and some that are misclassified as targets (FP).

Receiver Operating Characteristic (ROC) Curve

A ROC curve is a plot of the rate of TP against FP. ROC curves are used primarily in signal detection and medical diagnosis.

In these application, the TP rate is also called the sensitivity.

- ***Sensitivity*** is the probability that a classifier will classify a pattern as a target when it really is a target.
- ***Specificity*** is the probability that a classifier will correctly classify the true non-target cases.

Therefore a ROC curve is also a plot of sensitivity against 1 minus specificity.

ROC cont.

One of the major purposes of a ROC curve is to measure the **discriminating power of the classifier (power of test)**.

From ROC curve, we can understand the following about a classifier:

- It shows the trade-off between the *P*(*CC*) the target class (*sensitivity*)and the false alarm rate (1-*specificity*).
- The area under the ROC curve can be used to compare the **performance of classifiers**.

Cross-Validation for Specified False Alarm Rate

1. Given observations with class label ω_1 (target) and ω_2 (non-target), set desired probabilities of false alarm and a value for *k*.
2. Leave *k* points out of the non-target class to form a set of test cases denoted by *TEST*. We denote cases belonging to class ω_2 as $\mathbf{X2}_i$.
3. Estimate the class-conditional probabilities using the remaining n_2-*k* non-target cases and the n_1 target cases.
4. For each of those *k* observations, form the likelihood ratios

$$L_R\left(\mathbf{X2}_i\right) = \frac{P\left(\mathbf{X2}_i \middle| \omega_1\right)}{P\left(\mathbf{X2}_i \middle| \omega_2\right)}; \quad \mathbf{X2}_i \text{ in } TEST.$$

Continue

5. Repeat steps 2 through 4 using all of the non-target cases.
6. Order the likelihood ratios for the non-target class.
7. For each $P(FA)$, find the threshold that yields that value.
8. Leave k points out of the target class to form a set of test cases denote by *TEST*. We denote cases belonging to ω_1 by $\mathbf{X1}_i$.
9. Estimate the class-conditional probabilities using the remaining n_1-k target cases and the n_2 non-target cases.
10. For each of those k observations, form the likelihood ratios

$$L_R\left(\mathbf{X1}_i\right) = \frac{P\left(\mathbf{X1}_i \middle| \omega_1\right)}{P\left(\mathbf{X1}_i \middle| \omega_2\right)}; \quad \mathbf{X1}_i \text{ in } TEST.$$

Continue

11. Repeat steps 8 through 10 using all of the target cases.
12. Order the likelihood ratios for the target class.
13. For each threshold and $P(FA)$, find the proportion of target class that are correctly classified to obtain the $P(CC_{Target})$.