Classification

Classification

Given $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^d$, $y_i \in \{1, \ldots, M\}$, design a classifier that accurately predicts the y corresponding to an x observed in the future.

Examples

1. Medical decision making

   x_i = vector of various health predictors obtained from patient # i (age, weight, blood pressure, etc...)

   $y_i = \begin{cases} 1 & \text{if patient } i \text{ has a certain disease} \\ 2 & \text{" " " doesn't have " "} \end{cases}$

2. Handwritten digit recognition

   x_i = vector of pixel values from a digital scan of a handwritten digit

   y_i = some number $0, 1, 2, \ldots, 9$

3. Speech recognition

x_i = frequencies of the 3 largest spectral peaks in the i^{th} speech signal

y_i = vowel being spoken

Formal Problem Statement

Assume

(X, Y) is a random pair, $(X, Y) \sim P_{X,Y}$

$X \in \mathbb{R}^d$

$Y \in \{0, 1\}$ ← binary Classification

A <u>classifier</u> is a function

$$f : \mathbb{R}^d \longrightarrow \{0, 1\}$$

Remarks

1. X is called a pattern, signal, feature vector

   The coordinates of X are called features, attributes, predictors

2. Y is called a label

3. $M = 2$ is taken for simplicity. Some of the methods we will discuss generalize easily to multiclass problems, while others do not.

4. $\mathbb{R}^d$ is called the input space, feature space, pattern domain, etc.

   Categorical predictors (e.g., color) are not allowed, but some of the methods we'll discuss can handle them.

5. Y is <u>not</u> necessarily a deterministic function of X.

Learning from data

Training data for classification refers to a random sample from $P_{X,Y}$:

$$\underbrace{(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{IID}}{\sim} P_{X,Y}}_{D_n}$$

Let $\mathcal{F} = \{f : \mathbb{R}^d \to \{0,1\}\}$, the set of all classifiers.

A discrimination rule is a family of functions $\{A_n\}_{n=1}^{\infty}$ such that

$$A_n : (\mathbb{R}^d \times \{0,1\})^n \longrightarrow \mathcal{F}$$
$$D_n \longmapsto \hat{f}_n$$

That is, a discrimination rule constructs classifiers from training data.

Supervised Learning (labels given)

Classification : $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$

Regression : $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$

Goal: predict Y from X

Unsupervised Learning (labels not given)

Density estimation : $X \in \mathbb{R}^d$

Goal: Estimate density of X

Clustering : $(X, Y) \in \mathbb{R}^d \times \{1, \ldots, M\}$

Goal: predict Y from X

↑ training data has no labels!

Although classification is a special case of regression, it is given special attention because

1. It is a very important problem in its own right
2. It is easier in general
3. Performance is measured differently

Classification:

$$R(f) = P_{x,y}(f(x) \neq Y)$$

Regression:

$$R(f) = E_{x,y}\{|f(x) - Y|^2\}$$

The Bayes Classifier

Define the Bayes classifier

$$f^*(x) = \begin{cases} 1 & \text{if} \quad \eta(x) \geq \frac{1}{2} \\ 0 & \text{if} \quad \eta(x) < \frac{1}{2} \end{cases}$$

where

$$\eta(x) = P\{Y = 1 \mid X = x\}.$$

Theorem:

$$R(f^*) = \inf_{f \in \mathcal{F}} R(f)$$

That is, the Bayes classifier is optimal - it has minimal probability of error.

Note: If $\eta(x) = \frac{1}{2}$, the label may be assigned arbitrarily and the result still holds.

Discrete case

Suppose X is discrete

$$p(x) = P\{X = x\}$$

$$p_0(x) = P\{X = x \mid Y = 0\}$$

$$p_1(x) = P\{X = x \mid Y = 1\}$$

$$\pi = P\{Y = 1\}$$

Exercise | Apply Bayes rule to express the Bayes classifier in terms of the above quantities.

Solution

$$\eta(x) = P\{Y=1 \mid X=x\}$$

$$= \frac{P\{Y=1\} \cdot P\{X=x \mid Y=1\}}{P\{X=x\}}$$

$$= \frac{\pi\, p_1(x)}{p(x)}$$

$$= \frac{\pi\, p_1(x)}{\pi p_1(x) + (1-\pi) p_0(x)}$$

Therefore,

$$\eta(x) \gtrless \frac{1}{2} \iff 2\pi p_1(x) \gtrless \pi p_1(x) + (1-\pi) p_0(x)$$

$$\iff \pi p_1(x) \gtrless (1-\pi) p_0(x)$$

$$\iff \frac{p_1(x)}{p_0(x)} \gtrless \frac{1-\pi}{\pi}$$

likelihood ratio test

Proof of Theorem (discrete case)

Notation: $G_f = \{x : f(x) = 1\}$.

$$R(f) = P\{f(X) \neq Y\}$$

$$= \pi P\{f(X) = 0 \mid Y = 1\} + (1-\pi) P\{f(X) = 1 \mid Y = 0\}$$

$$= \pi \sum_{x \notin G_f} p_1(x) + (1-\pi) \sum_{x \in G_f} p_0(x).$$

This expression is minimized by taking $x \in G_f$ ($f(x) = 1$) when

$$\pi p_1(x) \geq (1-\pi) p_0(x). \qquad \square$$

Continuous case

Notation: $g(x)$ = density of X

$g_0(x)$ = " " $X \mid Y=0$

$g_1(x)$ = " " $X \mid Y=1$

Then

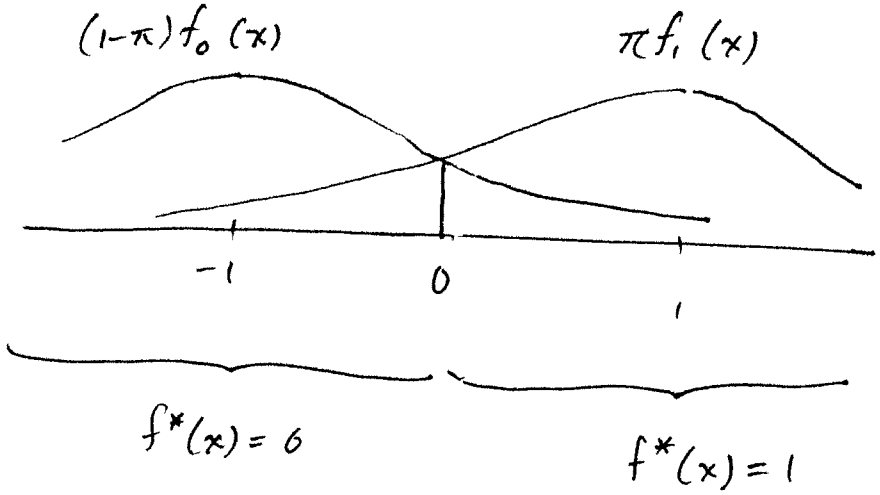
$$\eta(x) =$$

Exercise Prove the Bayes classifier is optimal for the case of continuous data.

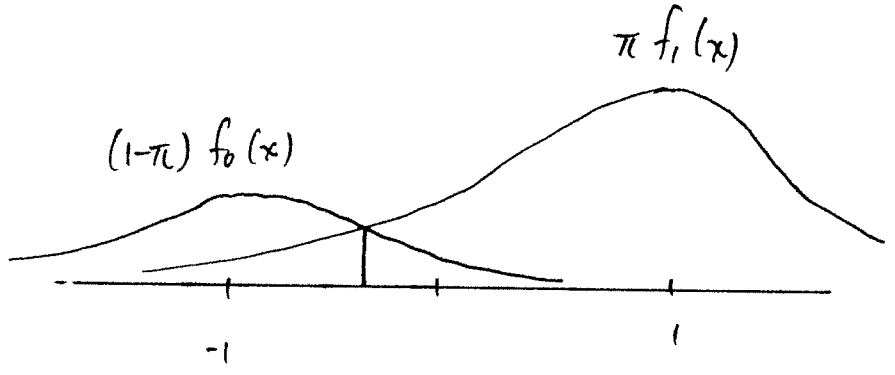
Example : $X|Y=0 \sim N(-1,1)$

$X|Y=1 \sim N(1,1)$

Case 1 : $\pi = \frac{1}{2}$



Case 2 $\pi > \frac{1}{2}$



Multiple Classes

$$f^*(x) = \arg\max_y P\{Y = y \mid X = x\}$$

$$= \begin{cases} \arg\max_y \pi_y \cdot P_y(x) & \text{discrete case} \\ \arg\max_y \pi_y \, g_y(x) & \text{continuous case} \end{cases}$$

Big Picture

f^* is the gold standard.

However, when learning from data,

f^* is unknown. Note: $R^* := R(f^*)$ is not necessarily 0 (see previous example).