BAYES CLASSIFIERS

Probabilistic Setting for Classification

Consider jointly distributed random variables X and Y where

$X \in \mathbb{R}^d$ — feature vector

$Y \in \{1, \dots, K\}$. — class label

Let's denote the joint distribution by P_{XY}. This is a function that maps subsets of $\mathbb{R}^d \times \{1, \dots, K\}$ (the sample space) to $[0, 1]$. We will conceptualize P_{XY} via two decompositions:

1. $P_{XY} = P_{X|Y} \cdot P_Y$
2. $P_{XY} = P_{Y|X} \cdot P_X$

Let's consider the first one:

$$P_{XY} = P_{X|Y} \cdot P_Y$$

P_Y: marginal distribution of Y, also called the <u>prior</u> class distribution

$P_{X|Y}$: conditional distribution of $X|Y$, aka the <u>class-conditional distribution</u>.

Note that P_Y is a discrete distribution, and can be represented by the <u>prior probabilities</u>

$$\pi_k = P_Y(Y=k), \quad k=1,\dots,K.$$

In practice, the features are usually either discrete or continuous, in which case $P_{X|Y}$ is represented by K different pmfs or pdfs, one for each class.

Example Suppose $K=2$, $d=1$, and $X|Y=k$ is Gaussian:

$$X|Y=1 \sim N(\mu_1, \sigma_1^2)$$

$$X|Y=2 \sim N(\mu_2, \sigma_2^2)$$

The decomposition $P_{XY} = P_{X|Y} \cdot P_Y$ has two primary uses.

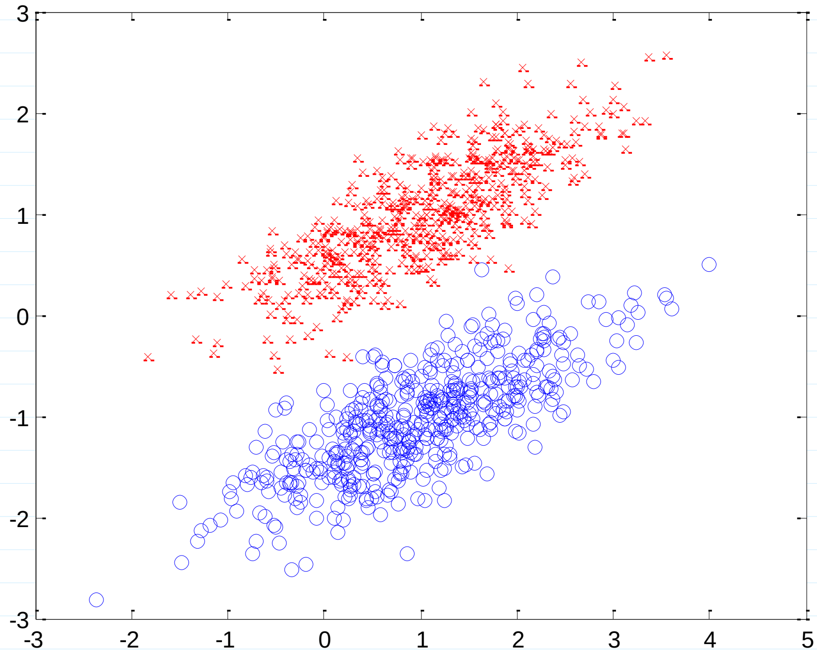
- ▷ Data generation: To simulate a realization of (X,Y), first generate a realization y of Y using P_Y, then generate a realization x of X using $P_{X|Y=y}$. Then (x,y) is a realization of (X,Y).

__Example__ Suppose $d=2$ and

$X | Y=1 \sim N(\mu_1, \Sigma)$ $\quad \mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

$X | Y=2 \sim N(\mu_2, \Sigma)$ $\quad \Sigma = \begin{bmatrix} .9 & .4 \\ .4 & .3 \end{bmatrix}$

If $\pi_1 = \pi_2 = \frac{1}{2}$, then training data would look like



▷ Calculating probabilities and expections using the laws of total probability and total expectation. We'll see an example below.

The other decomposition,

$$P_{XY} = P_{Y|X} \cdot P_X, \leftarrow \text{marginal distribution of } X$$

← posterior distribution of Y given X
(a discrete distrib. over labels
that depends on the observed x)

can also be used for data generation or calculating probabilities
and expectations.

Bayes Classifiers

Given a joint distribution P_{XY} of (X, Y), what is the best possible classifier?

A classifier is a function $f : \mathbb{R}^d \longrightarrow \{1, \ldots, K\}$.

The best classifier depends on the performance measure.
The most common performance measure is the probability of error, or risk, defined by

$$R(f) = P_{XY}(f(X) \neq Y),$$

i.e., the probability of the event

$$\{(x,y) \in \mathbb{R}^d \times \{1, \ldots, K\} \mid f(x) \neq y\}.$$

The Bayes risk is the smallest risk of any classifier, and is denoted R^*. If $R(f) = R^*$, f is called a Bayes classifier.

Let $\pi_k = P_Y(Y=k)$ denote the prior class probabilities, $g_k(x)$ the class-conditional pmfs/pdfs of $X|Y=k$, and

$$\eta_k(x) = P_{Y|X=x}(Y=k \mid X=x)$$

the posterior class probabilities. Notice that $\forall x, \sum_{k=1}^{K} \eta_k(x) = 1$.

Theorem The classifier

$$f^*(x) = \underset{k=1,\ldots,K}{\arg\max}\ \eta_k(x)$$

$$= \underset{k=1,\ldots,K}{\arg\max}\ \pi_k\, g_k(x)$$

is a Bayes classifier.

Proof For convenience, assume $X|Y=k$ has a continuous distribution for each k. Let f denote an arbitrary classifier. Denote the decision regions

$$\Gamma_k(f) = \{x \mid f(x) = k\}.$$

Then

$$1 - R(f) = P_{XY}\left(f(X) = Y\right)$$

[law of total probability]

$$= \sum_{k=1}^{K} P_Y(Y = k) \cdot P_{X|Y=k}\left(f(X) = k\right)$$

$$= \sum_{k=1}^{K} \pi_k \cdot \int_{\Gamma_k(f)} g_k(x)\,dx$$

Notice that $\Gamma_1(f), \ldots, \Gamma_K(f)$ form a partition of $\mathbb{R}^d$, i.e., every $x \in \mathbb{R}^d$ belongs to one and only one $\Gamma_k(f)$. Thus, to maximize $1 - R(f)$, we should choose $\Gamma_k(f)$ such that

$$x \in \Gamma_k(f) \iff \pi_k g_k(x) \text{ is maximal.}$$

So a Bayes classifier is

$$f^*(x) = \arg\max_k \; \pi_k g_k(x).$$

The proof is completed by observing

$$\eta_k(x) = \frac{\pi_k g_k(x)}{\underbrace{\sum_{\ell=1}^{K} \pi_\ell g_\ell(x)}_{\text{independent of } k}}$$

which follows by Bayes rule. ∎

In machine learning, we don't know P_{XY}, so we cannot find a Bayes classifier. As we will see, however, the above result motivates several classification methods.

To Learn More

Any textbook on machine learning with have a section on the above "decision theoretic" framework for classification.