Classification

Ali Ghodsi

University of Waterloo

September 17, 2015

Classification

The problem of prediction a discrete random variable $\mathbf{y}$ from another random variable $\mathbf{x}$ is Called classification. Consider *iid* data $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ where

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \in \mathcal{X} \subset \mathbb{R}^d$$

is a $d - dimensional$ vector and $\mathbf{y}_i$ takes values in some finite set $\mathcal{Y}$. A classification rule is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new $\mathbf{x}$ we predict $\mathbf{y}$ to be $h(\mathbf{x})$.

Error rate

Definition: The true error rate of a classifier h is

$$L(h) = Pr(h(\mathbf{x}) \neq \mathbf{y})$$

and the empirical error rate or training error rate is:

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n} I(h(\mathbf{x}_i) \neq \mathbf{y}_i)$$

Bayes Classifier

Consider the special case where $\mathcal{Y} = \{0, 1\}$ then

$$
\begin{aligned}
r(x_0) &= P(\mathbf{y} = 1 | \mathbf{x} = x_0) \\
&= \frac{P(\mathbf{x} = x_0 | \mathbf{y} = 1) P(\mathbf{y} = 1)}{P(\mathbf{x} = x_0)} \\
&= \frac{P(\mathbf{x} = x_0 | \mathbf{y} = 1) P(\mathbf{y} = 1)}{P(\mathbf{x} = x_0 | \mathbf{y} = 1) P(\mathbf{y} = 1) + P(\mathbf{x} = x_0 | \mathbf{y} = 0) P(\mathbf{y} = 0)}
\end{aligned}
$$

Bayes Classifier

Definition: The Bayes classification rule h^* is

$$h(x_0) = \begin{cases} 1 & \textit{if } \hat{r}(x_0) > 1/2 \\ 0 & \textit{otherwise} \end{cases}$$

Decision Boundary

The set $D(h) = \{\mathbf{x} : P(\mathbf{y} = 1|\mathbf{x} = x_0) = P(\mathbf{y} = 0|\mathbf{x} = x_0)\}$ is called the decision boundary.

$$h(x_0) = \begin{cases} 1 & \textit{if } P(\mathbf{y} = 1|\mathbf{x} = x_0) > P(\mathbf{y} = 0|\mathbf{x} = x_0) \\ 0 & \textit{otherwise} \end{cases}$$

Bayes Classifier

Theorem: The Bayes rule is optimal, that is if h is any other classification rule then $L(h^*) \leq L(h)$.

Bayes Classifier

Why do we need any other method?

Bayes Classifier

The Bayes rule depends on unknown quantities, so we need to use the data to find some approximation to the Bayes rule.

Three main approaches

1. Empirical Risk Minimization: Choose a set of classifier H and find $h^* \in H$ that minimizes some estimate of $L(h)$.

Three main approaches

2. Regression: Find an estimate $\hat{r}$ of the function r and define

$$h(x) = \begin{cases} 1 & \text{if } \hat{r}(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Three main approaches

3. Density Estimation: Estimate $P(\mathbf{x} = x|\mathbf{y} = 0)$ from the $\mathbf{x}_i$'s for which $\mathbf{y}_1 = 0$, estimate $P(\mathbf{x} = x|\mathbf{y} = 1)$ from the $\mathbf{x}_i$'s for which $\mathbf{y}_i = 1$, and let

$$P(\mathbf{y} = 1) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$$

Define $\hat{r}(x) = \hat{P}(\mathbf{y} = 1|\mathbf{x} = x)$ and

$$h(x) = \begin{cases} 1 & \textit{if } \hat{r}(x) > 1/2 \\ 0 & \textit{otherwise} \end{cases}$$

Multi-class Classification

Generalize to the case that $\mathbf{y}$ take on more than two values
Theorem: Suppose that $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, K\}$, the optimal rule is

$$h^*(x_0) = argmax_k \{P(\mathbf{y} = k | \mathbf{x} = x_0)\}$$

Multi-class Classification

where

$$P(\mathbf{y} = k|\mathbf{x} = x_0) = \frac{f_k(x_0)\pi_k}{\sum_r f_r(x_0)\pi_r}$$

$$= \frac{P(\mathbf{x} = x_0|\mathbf{y} = 1)P(\mathbf{y} = 1)}{P(\mathbf{x} = x_0|\mathbf{y} = 1)P(\mathbf{y} = 1) + P(\mathbf{x} = x_0|\mathbf{y} = 0)P(\mathbf{y} = 0)}$$

Toward LDA and QDA

The simplest approach to classification is to use the third approach and assume a parametric model for the densities.

The simplest method is to use approach 3 (above) and assume a parametric model for densities. Assume class conditional is Gaussian.

$\mathcal{Y} = \{0, 1\}$ assumed (i.e., 2 labels)

$$h(x) = \begin{cases} 1 & P(Y = 1|\mathbf{x} = x) > P(Y = 0|\mathbf{x} = x) \\ 0 & otherwise \end{cases}$$

$$P(Y = 1|\mathbf{x} = x) = \frac{f_1(x)\pi_1}{\sum_k f_k \pi_k} \quad \text{(denom = P(x))}$$

1) Assume Gaussian distributions

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{k})^T \Sigma_k^{-1} (\mathbf{x} - \mu_\mathbf{k})\Big)$$

must compare

$\frac{f_1(x)\pi_1}{p(x)}$ with $\frac{f_0(x)\pi_0}{p(x)}$

Note that the p(x) denom can be ignored:

$f_1(x)\pi_1$ with $f_0(x)\pi_0$

To find the decision boundary, set

$$f_1(x)\pi_1 = f_0(x)\pi_0$$

$$\frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{1})^T\Sigma_1^{-1}(\mathbf{x}-\mu_\mathbf{1}))\pi_1 =$$
$$\frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{0})^T\Sigma_0^{-1}(\mathbf{x}-\mu_\mathbf{0}))\pi_0$$

2) Assume $\Sigma_1 = \Sigma_0$, we can use $\Sigma = \Sigma_0 = \Sigma_1$.

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \Sigma^{-1}(\mathbf{x} - \mu_\mathbf{1}))\pi_1 =$$
$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{0})^T \Sigma^{-1}(\mathbf{x} - \mu_\mathbf{0}))\pi_0$$

3) Cancel $(2\pi)^{-d/2}|\Sigma|^{-1/2}$ from both sides.

$$exp(-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{1}})^T\Sigma^{-1}(\mathbf{x}-\mu_{\mathbf{1}}))\pi_1 =$$
$$exp(-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{0}})^T\Sigma^{-1}(\mathbf{x}-\mu_{\mathbf{0}}))\pi_0$$

4) Take log of both sides.

$$-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1))+\log(\pi_1)=$$
$$-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0))+\log(\pi_0)$$

5) Subtract one side from both sides, leaving zero on one side.

$$-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)+\log(\pi_1)-[-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0)+\log(\pi_0)]=0$$

$$\frac{1}{2}[-\mathbf{x}^T\Sigma^{-1}\mathbf{x}-\mu_1{}^T\Sigma^{-1}\mu_1+2\mu_1{}^T\Sigma^{-1}\mathbf{x}+\mathbf{x}^T\Sigma^{-1}\mathbf{x}+\mu_0{}^T\Sigma^{-1}\mu_0-2\mu_0{}^T\Sigma^{-1}\mathbf{x}]+\log(\frac{\pi_1}{\pi_0})=0$$

Cancelling out the terms quadratic in $\mathbf{x}$ and rearranging results in

$$\frac{1}{2}[-\mu_1{}^T\Sigma^{-1}\mu_1 + \mu_0{}^T\Sigma^{-1}\mu_0 + (2\mu_1{}^T\Sigma^{-1} - 2\mu_0{}^T\Sigma^{-1})\mathbf{x}] + \log(\frac{\pi_1}{\pi_0}) = 0$$

We can see that the first pair of terms is constant, and the second pair is linear in x. Therefore, we end up with something of the form

$$\boldsymbol{a}^T\boldsymbol{x} + b = 0$$

If we relax assumption 2 (i.e. $\Sigma_1 \neq \Sigma_0$) then we get a quadratic equation that can be written as:

$$x^T Ax + b^T x + c = 0$$

Theorem

Suppose that $Y \in \{1, \ldots, K\}$, if $f_k(\mathbf{x}) = Pr(\mathbf{x} = \mathbf{x}|Y = k)$ is Gaussian. The Bayes Classifier is

$$h^*(\mathbf{x}) = \arg\max_k \delta_k(\mathbf{x})$$

Where

$$\delta_k(\mathbf{x}) = -\frac{1}{2} log(|\Sigma_k|) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + log(\pi_k)$$

To compute this, we need to calculate the value for each class, and then take the one with the max value.

When Gaussians have the same covariance

$$\Sigma_0 = \Sigma_1 = \Sigma_2 = \cdots = \Sigma_{k-1}$$

(e.g. LDA), then

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \tfrac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + log(\pi_k)$$

In practice

In practice we don't know the parameters of Gaussian and will need to estimate them using our training data.

$$\hat{\pi}_k = \hat{Pr}(y = k) = \frac{n_k}{n}$$

where n_k is the number of class k observations.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k - k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

If we wish to use LDA we must calculate a common covariance, so we average all the covariances e.g.

$$\sum = \frac{\sum_{r=1}^{k}(n_r \Sigma_r)}{\sum_{r=1}^{k} n_r}$$

Where:

n_r is the number of data points in class r

$\sum_r$ is the covariance of class r

n is the total number of data points, and k is the number of classes.