Chapter 8

Friday, Jan 27, 06

Two approaches to classification:

- Generative Approach

  Uses a parametric family of models and obtains a classifier by first estimating the class conditional density, then classifying each new data point to the class with the highest probability. It is a way to generate **x** from y. Ex: Naive Bayes

- Discriminative Approach

  Depends only on the conditional density $p(y|x)$. Discriminative methods model the conditional without making any assumptions about the imput **x**. Here **x** is always observed. We don't need to generate it. Ex: Logitic Regression

Gaussian Class-conditional densities

Let

$$P(\mathbf{x}|Y = j) = N(\mu_j, \Sigma_j)$$

$$P(Y = j) = \pi_j$$

Recall

$$P(Y = j|\mathbf{x}) = \frac{P(\mathbf{x}|Y = j)P(Y = j)}{\sum_{k=1}^{C} P(\mathbf{x}|Y = k)P(Y = k)}$$

and the Gaussian pdf

$$p(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{d}/\mathbf{2}}|\mathbf{\Sigma}|^{\mathbf{1}/\mathbf{2}}} \exp\left[-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x} - \mu)\right]$$

Remember, under 0-1 loss, Bayes decision rule will pick the class j that maximizes the discriminant function, which we saw in section 6.2

$$g_j(x) = \log P(\mathbf{x}|Y = j) + \log P(Y = j)$$

so it will pick g_j if $g_j > g_k$

Expanding the previous equations and considering the following scenarios:

- $\Sigma_j = \Sigma$, tied across all classes
- Σ_j is diagonal (The Naive Bayes assumption)
- Y is binary, $Y \in \{0, 1\}$
- The general case

Case 1: $\Sigma_j = \Sigma$, $Y \in \{0, 1\}$

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

$$= \frac{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)\right]\pi_1}{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\left(\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)\right]\pi_1+\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_0)'\Sigma^{-1}(\mathbf{x}-\mu_0)\right]\pi_0\right)}$$

$$= \frac{\pi_1 e^{a_1}}{\pi_1 e^{a_1}+\pi_0 e^{a_0}}$$

Where

$$a_j = -\frac{1}{2}(\mathbf{x} - \mu_j)'\Sigma^{-1}(\mathbf{x} - \mu_j)$$

$$\begin{aligned} a_0 - a_1 &= -\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma^{-1} (\mathbf{x} - \mu_0) + \frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1) \\ &= -(\mu_1 - \mu_0)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 + \mu_0) \end{aligned}$$

By dividing the numerator and the denominator by $\pi_1 e^{a_1}$, we get:

$$\begin{aligned} P(Y = 1|\mathbf{x}) &= \frac{1}{1+\exp\left[-\log\frac{\pi_1}{\pi_0}+a_0-a_1\right]} \\ &= \frac{1}{1+\exp[-\beta'\mathbf{x}-\gamma]} \\ &= \sigma(\beta'\mathbf{x}+\gamma) \end{aligned}$$

Where

$$
\begin{aligned}
\beta &= \Sigma^{-1}(\mu_1 - \mu_0) \\
\gamma &= -\tfrac{1}{2}(\mu_1 - \mu_0)^T(\mu_1 + \mu_0) + \log \tfrac{\pi_1}{\pi_0} \\
\sigma(z) &= \tfrac{1}{1+e^{-z}} = \tfrac{e^z}{e^z+1}
\end{aligned}
$$

$\sigma(z)$ is called the **logistic function** or **sigmoid function**

The sigmoid has the following property:

$$p = \sigma(z) \iff z = \log \frac{p}{1-p}$$

where $\log \frac{p}{1-p}$ is called a **log-odds ratio**.

It is also easy to show:

$$1 - \sigma(z) = \sigma(-z)$$

Effect of β

consider the case where

$$\sigma(\beta'\mathbf{x})$$

Then

$$\frac{P(Y=1|\mathbf{x}, x_j=1)}{P(Y=1|\mathbf{x}, x_j=0)} = \frac{\exp\left(\beta_0 + \sum_{i\neq j}\beta_i x_i + \beta_j\right)}{\exp\left(\beta_0 + \sum_{i\neq j}\beta_i x_i\right)} = e_j^{\beta}$$

β_j controls the steepness with which the probability increases.

Decision Boundary

Points of equal posteriors all lie on the line between the two means.

$$P(Y = 1|\mathbf{x}) = P(Y = 0|\mathbf{x}) = 0.5$$

To find the decision boundary, we need to solve for:

$$\begin{aligned}
\sigma(z) &= 0.5 \\
z &= \log \tfrac{p}{1-p} \\
&= \log \tfrac{0.5}{0.5} \\
&= \log 1 \\
&= 0
\end{aligned}$$

If we consider the case where $\pi_1 = \pi_0 = 0.5$, we have:

$$z = \beta'\mathbf{x} + \gamma = (\mu_1 - \mu_0)' \left(x - \frac{(\mu_1 +, u_0)}{2} \right)$$

The boundary line is orthogonal to $\mu_2 - \mu_1$ and is equidistance from the two means. If the priors are non-uniform the the decision boundary shifts:

- if $\pi_1 > \pi_2$, the boundary shifts right.

- if $\pi_1 < \pi_2$, the boudary shifts left.

Effect of Σ: If Σ is not spherical, the decision boundary is no longer orthogonal to $\mu_2 - \mu_1$.