2.9 Bayes Decision Theory

Discrete Features

Yuta Akizuki

Recap of 2.2 - Bayesian Formula

The posterior probability computed from $p(\boldsymbol{x}|\omega_j)$ is

$$P(w_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j)P(w_j)}{p(\boldsymbol{x})}$$

- The categories: $\omega_1, \ldots, \omega_c$
- The feature vector of $\boldsymbol{x} = (x_1, \ldots, x_d)^t$
- $x_i \in \boldsymbol{R}$

Recap of 2.3 - Zero-one Loss Function

The conditional risk is

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\boldsymbol{x})$$
$$= 1 - P(\omega_i|\boldsymbol{x})$$

where the zero-one loss function is assigned,

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & (i = j) \\ 1 & (i \neq j) \end{cases}$$

Recap of 2.4 - Discriminant Functions

Let $g_i(\boldsymbol{x}) = -R(\alpha_i|\boldsymbol{x})$ as a discriminant function, the classifier is said to assign a feature vector $\boldsymbol{x}$ to class ω_i if

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad for\ all\ j \neq i$$

where

$$g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(w_i)}{\sum_{j=1}^{c} p(\boldsymbol{x}|\omega_j)P(\omega_j)}$$

Recap of 2.4 - Discriminant Functions

Since the evidence is independent on ω_i and natural logarithm is a monotonically increasing function, $g_i(\boldsymbol{x})$ can be written as

$$g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i)$$

For the two category case, the discriminant function is defined as

$$\begin{aligned} g(\boldsymbol{x}) &\equiv g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ &= \ln \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

2.9 Bayes Decision Theory

Discrete Features

2.9 Bayes Decision Theory - Discrete Features

When the components of $\boldsymbol{x}$ are discrete values, sums of discrete probability distribution becomes

$$\sum_{x} P(\boldsymbol{x}|\omega_j)$$

instead of integrals of the probability density function for continuous features:

$$\int p(\boldsymbol{x}|\omega_j)d\boldsymbol{x}$$

2.9 Bayes Decision Theory - Discrete Features

The posterior probability is

$$P(w_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_j)P(w_j)}{P(\boldsymbol{x})}$$

Where the evidence is

$$P(\boldsymbol{x}) = \sum_{j=i}^{c} P(\boldsymbol{x}|\omega_j)P(\omega_j)$$

2.9 Bayes Decision Theory - Discrete Features

To minimize the error rate is to maximum the posterior probability when the loss function $\lambda(\alpha_i|\omega_j)$ is zero-one loss function.

$$\begin{aligned}
\alpha^* &= \arg\min_i R(\alpha_i|\boldsymbol{x}) \\
&= \arg\min_i \sum_{j=i}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\boldsymbol{x}) \\
&= \arg\min_i 1 - P(\omega_i|\boldsymbol{x}) \\
&= \arg\max_i P(\omega_i|\boldsymbol{x})
\end{aligned}$$

2.9.1 Independent Binary Features

Consider the two-category problem:

- The categories: ω_1, ω_2
- The independent feature vector of $\boldsymbol{x} = (x_1, \ldots, x_d)^t$
- $x_i = 0\ ,\ 1$
- $p_i = Pr[x_i = 1|\omega_1]$ (the probability of $x_i = 1$ under ω_1)
- $q_i = Pr[x_i = 1|\omega_2]$ (the probability of $x_i = 1$ under ω_2)

2.9.1 Independent Binary Features

The class-conditional probabilities can be written as:

$$P(\boldsymbol{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} \quad P(\boldsymbol{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}$$

The likelihood ratio is given by

$$\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1 - p_i}{1 - q_i}\right)^{1-x_i}$$

2.9.1 Independent Binary Features

The discriminant function is

$$g(\boldsymbol{x}) = ln\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} + \frac{P(\omega_1)}{P(\omega_2)}$$

$$= \sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\right] + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

2.9.1 Independent Binary Features

The function is linear in the x_i, and thus can be written

$$g(\boldsymbol{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

where

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad and \quad w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

2.9.1 Independent Binary Features

$$w_i = \ln \frac{p_i}{q_i} \frac{1-q_i}{1-p_i} \; and \; w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

- If $p_i > q_i$, then $\frac{p_i}{q_i} > 0, \frac{1-q_i}{1-p_i} > 0 \Rightarrow w_i > 0$
- For any fixed $q_i < 1$, ω_i gets larger as p_i gets larger
- $P(w_i)$ give biases the decision in favor of ω_i in the threshold weight of w_0

2.9.1 Independent Binary Features

- The simple classifier is obtained because of the condition of feature independence.
- The inter-dependent features needs a more complicated classifier.
- The possible values for $\boldsymbol{x}$ appear in d-dimensional hypercube, and the decision surface defined by $g(\boldsymbol{x}) = 0$ is a hyperplane.

Bayesian Decisions for 3D Binary Data

$$P(\omega_1) = P(\omega_2) = 0.5,\ p_i = 0.8\ and\ q_i = 0.5\ for\ i = 1, 2, 3$$

The decision surface[1] is

$$g(\boldsymbol{x}) = \sum_{i=1}^{3} 1.3863 x_i - 2.75 = 0$$

$$1.3863 x_1 + 1.3863 x_2 + 1.3863 x_3 - 2.75 = 0$$

[1] Visualized decision surface for example 3 on https://ytakzk.github.io/Bayes-Decision-Theory-Discrete-Features/.