

# 2.9 Bayes Decision Theory

## Discrete Features

Yuta Akizuki

## Recap of 2.2 - Bayesian Formula

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

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

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

## Recap of 2.3 - Zero-one Loss Function

The conditional risk is

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{j=1}^c \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \\ &= 1 - P(\omega_i | \mathbf{x}) \end{aligned}$$

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

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

where

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

## Recap of 2.4 - Discriminant Functions

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

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

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

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

# 2.9 Bayes Decision Theory

## Discrete Features

## 2.9 Bayes Decision Theory - Discrete Features

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

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

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

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

## 2.9 Bayes Decision Theory - Discrete Features

The posterior probability is

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

Where the evidence is

$$P(\mathbf{x}) = \sum_{j=1}^c P(\mathbf{x}|\omega_j)P(\omega_j)$$



## 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.

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

## 2.9.1 Independent Binary Features

Consider the two-category problem:

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

## 2.9.1 Independent Binary Features

The class-conditional probabilities can be written as:

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

The likelihood ratio is given by

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

## 2.9.1 Independent Binary Features

The discriminant function is

$$\begin{aligned} g(\mathbf{x}) &= \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \frac{P(\omega_1)}{P(\omega_2)} \\ &= \sum_{i=1}^d \left[ x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)} \end{aligned}$$

## 2.9.1 Independent Binary Features

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

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

where

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

## 2.9.1 Independent Binary Features

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

- If  $p_i > q_i$ , then  $\frac{p_i}{q_i} > 1$ ,  $\frac{1 - q_i}{1 - p_i} > 1 \Rightarrow w_i > 0$
- For any fixed  $q_i < 1$ ,  $w_i$  gets larger as  $p_i$  gets larger
- $P(\omega_i)$  give biases the decision in favor of  $\omega_i$  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  $\mathbf{x}$  appear in d-dimensional hypercube, and the decision surface defined by  $g(\mathbf{x}) = 0$  is a hyperplane.

# Bayesian Decisions for 3D Binary Data

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

The decision surface<sup>1</sup> is

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

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

---

<sup>1</sup> Visualized decision surface for example 3 on [GitHub](#).