Bayes classifier

In statistical classification the **Bayes classifier** minimizes the probability of misclassification.[1]

Definition

Suppose a pair $(\boldsymbol{X}, \boldsymbol{Y})$ takes values in $\mathbb{R}^d \times \{1, 2, \ldots, K\}$, where $\boldsymbol{Y}$ is the class label of $\boldsymbol{X}$. This means that the conditional distribution of X, given that the label Y takes the value r is given by

$$\boldsymbol{X} \mid \boldsymbol{Y} = r \sim \boldsymbol{P}_r \text{ for } r = 1, 2, \ldots, K$$

where "$\sim$" means "is distributed as", and where $\boldsymbol{P}_r$ denotes a probability distribution.

A classifier is a rule that assigns to an observation $X=x$ a guess or estimate of what the unobserved label $Y=r$ actually was. In theoretical terms, a classifier is a measurable function $\boldsymbol{C} : \mathbb{R}^d \to \{1, 2, \ldots, K\}$, with the interpretation that C classifies the point x to the class $C(x)$. The probability of misclassification, or risk, of a classifier C is defined as

$$\mathcal{R}(C) = \mathrm{P}\{C(\boldsymbol{X}) \neq \boldsymbol{Y}\}.$$

The Bayes classifier is

$$C^{\text{Bayes}}(\boldsymbol{x}) = \underset{r \in \{1,2,\ldots,K\}}{\operatorname{argmax}} \mathrm{P}(\boldsymbol{Y} = r \mid \boldsymbol{X} = \boldsymbol{x}).$$

In practice, as in most of statistics, the difficulties and subtleties are associated with modeling the probability distributions effectively—in this case, $\mathrm{P}(\boldsymbol{Y} = r \mid \boldsymbol{X} = \boldsymbol{x})$. The Bayes classifier is a useful benchmark in statistical classification.

The excess risk of a general classifier $\boldsymbol{C}$ (possibly depending on some training data) is defined as $\mathcal{R}(C) - \mathcal{R}(C^{\text{Bayes}})$. Thus this non-negative quantity is important for assessing the performance of different classification techniques. A classifier is said to be consistent if the excess risk converges to zero as the size of the training data set tends to infinity.

See also

- Naive Bayes classifier

References

1. Devroye, L.; Gyorfi, L. & Lugosi, G. (1996) *A probabilistic theory of pattern recognition*. Springer. ISBN 0-3879-4618-7.