Bayes classifier and Bayes error

Jyrki Kivinen

19 November 2013

After the lecture today, several students asked for clarification to the concepts of Bayes classifier and Bayes error. To put it simply, the *Bayes classifier* (or Bayes optimal classifier) is a classifier that minimizes a certain probabilistic error measure. The *Bayes error* is then the error of the Bayes classifier. As far as I know, the word "Bayes" appears simply because the Bayes formula is often very useful in analysing these concepts.

The notions of Bayes optimality and Bayes error generalize directly also to regression, but for simplicity we consider only classification here.

To keep things even more simple, assume the objects to be classified come from a finite set $\mathcal{X}$. If the set of classes is $\mathcal{Y}$, a classifier is then a function $f\colon \mathcal{X} \to \mathcal{Y}$.

We now fix some cost function $C\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The most common case is the 0-1-loss where $C(y, y) = 0$ for all y, and $C(y, y') = 1$ if $y = y'$, giving the number of misclassifications. (see, for example, pages 88–92 in lecture notes). If a classifier f is applied to a point x for which the real class is y, the cost $C(y, f(x))$ is incurred.

Suppose further we are given a probability distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. We can then define the *expected cost* $\mathrm{Cost}(f)$ of a classifier f as

$$\mathrm{Cost}(f) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P(x, y)C(y, f(x)).$$

A classifier f_* is called *Bayes optimal*, or *Bayes classifier*, if it minimises $\mathrm{Cost}(\cdot)$, that is,

$$f_* = \arg\min_f \mathrm{Cost}(f).$$

The minimum expected loss $\mathrm{Cost}(f_*)$ is called the *Bayes error*. (In general, the Bayes optimal classifier need not be unique, since there may be several classifiers that achieve the same minimal error.)

If we want to find the Bayes classifier, note that we can write

$$\text{Cost}(f) = \sum_{x \in \mathcal{X}} G_x(f(x))$$

where

$$G_x(y) = \sum_{y' \in \mathcal{Y}} P(x, y') C(y', y).$$

Then we can, separately for each $x \in \mathcal{X}$, pick $f(x) = \arg\min_{y \in \mathcal{Y}} G_x(y)$, and it is easy to see that this gives a Bayes optimal $f = f_*$. For example, for the 0-1-loss, we have

$$G_x(y) = \sum_{y' \neq y} P(x, y'),$$

so we get simply

$$f_*(x) = \arg\min_{y \in Y} \sum_{y' \neq y} P(x, y') = \arg\max_{y \in Y} P(x, y).$$

Often one is given the class conditional probabilities instead of the total probabilities $P(X, Y)$, so this becomes

$$f_*(x) = \arg\max_{y \in Y} P(x \mid y) P(y).$$

We have not said anything about where the probability distribution $P(X, Y)$ comes from. In theoretical analyses we may assume that it is some unknown "true" distribution whereby the data are generated. In practice it might be the result of some machine learning algorith, and by determining the Bayes optimal prediction we are converting probabilistic predictions to forced-choice so as to minimize the resulting number of mistakes, assuming our initial probabilities were (roughly) correct.