Classification using Bayes Rule

Bayes rule allows us to say something very strong about a particular (and very important) problem in statistical inference: classification.

The problem is stated as follows. We are given (random) data points $X_1 = \boldsymbol{x}_1, X_2 = \boldsymbol{x}_2, \ldots$ that we want to assign to one of K different *classes*. Underlying this is a probabilistic model, where the class (denoted by Y_n) controls which distribution the data point X_n comes from:

$$X_n \sim f_X(\boldsymbol{x}|Y = k), \quad \text{if } X_n \text{ comes from class } k.$$

The goal is to come up a *classification rule* (or **classifier** for short) that maps data points X to their respective classes Y as accurately as possible.

More formally, a **classifier** is a function $h : \mathbb{R}^D \to \{1, \ldots, K\}$; this function takes a data point $X \in \mathbb{R}^D$ (in machine learning, this is sometime referred to as a "feature vector") and returns a discrete class label. This means that the function h is *piecewise constant*: an equivalent way of specifying h is through a **partition** of $\mathbb{R}^D$ into K regions $\Gamma_1(h), \ldots, \Gamma_K(h)$, where $\Gamma_k(h)$ is the set of point that h maps to k:

$$\Gamma_k(h) = \{\boldsymbol{x} \in \mathbb{R}^D \ : \ h(\boldsymbol{x}) = k\}.$$

[PICTURE]

If we have a known probabilistic model (that is accurate), there is a clear-cut best answer for choosing the best classification rule. Here, we are treating the (data point, class label) pair (X, Y) as coupled random variables — we observe X, and want to infer Y. If we know the conditional densities

$$X|Y = k \sim f_X(\boldsymbol{x}|Y = k) \tag{1}$$

for all $k = 1, \ldots, K$, and the marginal probabilities

$$p_Y(k) = \mathrm{P}\left(Y = k\right) \tag{2}$$

again for all $k = 1, \ldots, K$, then we have completely specified the joint distribution of the continuous-valued random vector X and the discrete-valued random scalar Y. For a given classification rule h, the probability of h making an error is

$$R(h) = \mathrm{P}\left(h(X) \neq Y\right).$$

We might also write this as the probability that a sample arrives in the wrong partition specified by h:

$$R(h) = \mathrm{P}\left(X \notin \Gamma_Y(h)\right).$$

The quantity $R(h)$ is sometimes referred to as the **risk** associated with the rule h (hence the R for notation).

Given the probabilistic model specified by (1) and (2), the best classifier is simply to choose the class that has maximum conditional probability. This is codified in the following theorem.

Theorem: Define the classifier

$$h^*(\boldsymbol{x}) = \arg \max_{k \in \{1,\ldots,K\}} \mathrm{P}\left(Y = k|X = \boldsymbol{x}\right). \tag{3}$$

Then every other classifier h has

$$R(h) \geq R(h^*). \tag{4}$$

The classifier given by (3) is call the **Bayes classifier**, and its probability of error in (4) is called the **Bayes risk**.

Proof: The optimality of h^* in (3) follows from carefully writing down the risk for an arbitrary classifier h, applying Bayes rule, and then showing that h^* optimizes the resulting expression. We start with an expression for $1 - R(h)$, which we will show is as *large* as possible when $h = h^*$:

$$\begin{aligned}
1 - R(h) &= \mathrm{P}\left(h(X) = Y\right) \\
&= \sum_{k=1}^{K} \mathrm{P}\left(Y = k\right) \cdot \mathrm{P}\left(h(X) = k | Y = k\right) \\
&= \sum_{k=1}^{K} \mathrm{P}\left(Y = k\right) \int_{\Gamma_k(h)} f_X(\boldsymbol{x}|Y = k) \ \mathrm{d}\boldsymbol{x} \\
&= \sum_{k=1}^{K} \int_{\Gamma_k(h)} \mathrm{P}\left(Y = k\right) f_X(\boldsymbol{x}|Y = k) \ \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

By Bayes rule,

$$\begin{aligned}
\mathrm{P}\left(Y = k | X = \boldsymbol{x}\right) &= \frac{\mathrm{P}\left(Y = k\right) f_X(\boldsymbol{x}|Y = k)}{f_X(\boldsymbol{x})} \\
&= \frac{\mathrm{P}\left(Y = k\right) f_X(\boldsymbol{x}|Y = k)}{\sum_{\ell=1}^{K} \mathrm{P}\left(Y = \ell\right) f_X(\boldsymbol{x}|Y = \ell)}.
\end{aligned}$$

Note that the denominator is a function of $\boldsymbol{x}$ that is independent of k. Using this and the fact that the regions $\Gamma_k(h)$ are disjoint, we can

continue the string of equalities:

$$1 - R(h) = \int_{\mathbb{R}^D} \left(\sum_{k=1}^{K} 1_{\Gamma_k(h)}(\boldsymbol{x})\, f_X(\boldsymbol{x})\, \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right) \right) \, \mathrm{d}\boldsymbol{x},$$

where $1_{\mathcal{A}}(\boldsymbol{x})$ is the indicator function

$$1_{\mathcal{A}}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in \mathcal{A}, \\ 0, & \boldsymbol{x} \notin \mathcal{A}. \end{cases}$$

The way we choose h^* in (3) chooses the regions so that the function inside the integral above is as large as possible; it is clear[1] that

$$\sum_{k=1}^{K} 1_{\Gamma_k(h)}(\boldsymbol{x}) f_X(\boldsymbol{x})\, \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right) \;\leq\; \sum_{k=1}^{K} 1_{\Gamma_k(h^*)}(\boldsymbol{x}) f_X(\boldsymbol{x})\, \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right),$$

for all $\boldsymbol{x} \in \mathbb{R}^D$. Thus

$$\begin{aligned} 1 - R(h) &\leq \int_{\mathbb{R}^D} \left(\sum_{k=1}^{K} 1_{\Gamma_k(h)}(\boldsymbol{x})\, f_X(\boldsymbol{x})\, \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right) \right) \, \mathrm{d}\boldsymbol{x} \\ &= 1 - R(h^*), \end{aligned}$$

and so $R(h^*) \leq R(h)$.

[1] For a fixed $\boldsymbol{x}$, only one term in each of the sums is non-zero, and by definition h^* chooses the term that will be the largest.

Examples:

1. $D = 1$, $K = 2$, $\mathrm{P}(Y = 1) = \mathrm{P}(Y = 2) = 1/2$, and $X|Y = k \sim \mathrm{Normal}(k, 1)$.

2. $D = 2$, $K = 2$, $\mathrm{P}(Y = 1) = \mathrm{P}(Y = 2) = 1/2$, and $X|Y = k \sim \mathrm{Normal}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I})$, where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are given vectors in $\mathbb{R}^D$.

Nearest-neighbor classification

The Bayes classifier is great ... if we have perfect knowledge of the probability model. But this is only the case in a few special applications[2] — in fact, it's not an exaggeration to say that the entire field of machine learning sprouted up around the fact that it is entirely unrealistic that we know these distributions at all, and that everything should be learned from examples.

There is a super-easy data-driven rule for classification that requires no probability model, is very intuitive, and whose (asymptotic) performance comes within a factor of 2 of the optimal Bayes risk. It is called the *nearest neighbor rule*.

Instead of a probability model, we are given examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$. Our analysis will still assume that these are random, have some joint distribution, each pair $(\boldsymbol{x}_n, y_n)$ is independent from the other pairs. Our classification algorithm, however, will operate without knowledge of this joint distribution.

Given an unlabeled data point $\boldsymbol{x}$, the nearest-neighbor rule simply assigns the same label as the closest example point to $\boldsymbol{x}$. If $\mathrm{NI}(\boldsymbol{x})$ is the index of the closest point to $\boldsymbol{x}$,

$$\mathrm{NI}(\boldsymbol{x}) = \arg \min_{n=1,\ldots,N} \|\boldsymbol{x} - \boldsymbol{x}_n\|_2,$$

then the nearest-neighbor classifier is

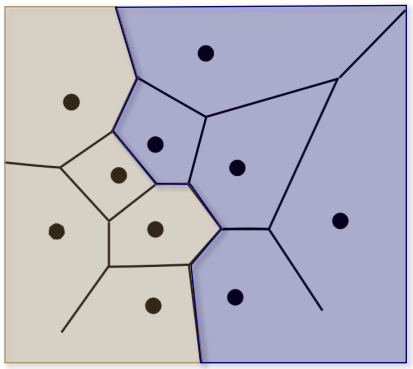
$$h_{\mathrm{NN}}(\boldsymbol{x}) = y_{\mathrm{NI}(\boldsymbol{x})}.$$

[2] Detectors for digital communications is one area where we are pretty confident that the class conditional densities are Gaussian, for example.

The example data point closest to $\boldsymbol{x}$ is of course referred to as the "nearest neighbor" and is denoted

$$\mathrm{NN}(\boldsymbol{x}) = \boldsymbol{x}_{\mathrm{NI}(\boldsymbol{x})}.$$

The regions defined by the nearest neighbor classifier are a *Voronoi partition*.



As we can see, this rule is completely data-driven, and does not even make an attempt at estimating the conditional probability model $f_X(\boldsymbol{x}|Y = k)$. Nevertheless, its asymptotic performance is within a constant factor of the Bayes risk.

Theorem. As the number of examples N goes to infinity, the nearest-neighbor classifier $h_{\mathrm{NN}}(\boldsymbol{x})$ has a probability of error that obeys

$$R(h_{\mathrm{NN}}) \leq 2R(h^*),$$

where h^* is the Bayes classifier.

Proving this in detail is a bit technical, but we can get a good idea of how the argument works by just looking at the $K = 2$ case and

applying some basic conditional probability. First, we write the NN-risk as

$$\mathrm{P}\left(h_{\mathrm{NN}}(X) \neq Y\right) = \int_{\boldsymbol{x}\in\mathbb{R}^D} \mathrm{P}\left(h_{\mathrm{NN}}(X) \neq Y | X = \boldsymbol{x}\right)\ f_X(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}.$$

We now look at the probability we make a mistake at a fixed point $X = \boldsymbol{x}$. We have[3]

$$\begin{aligned}\mathrm{P}\left(h_{\mathrm{NN}}(X) \neq Y | X = \boldsymbol{x}\right) = \mathrm{P}\left(Y = 1 | X = \mathrm{NN}(\boldsymbol{x})\right) \mathrm{P}\left(Y = 2 | X = \boldsymbol{x}\right) \\ + \mathrm{P}\left(Y = 2 | X = \mathrm{NN}(\boldsymbol{x})\right) \mathrm{P}\left(Y = 1 | X = \boldsymbol{x}\right)\end{aligned}$$

As $N \to \infty$, $\| \mathrm{NN}(\boldsymbol{x}) - \boldsymbol{x}\|_2 \to 0$, so[4]

$$\mathrm{P}\left(Y = k | X = \mathrm{NN}(\boldsymbol{x})\right) \to \mathrm{P}\left(Y = k | X = \boldsymbol{x}\right), \quad \text{for } k = 1, 2.$$

To ease the notation here, set

$$\eta_k(\boldsymbol{x}) = P(Y = k | X = \boldsymbol{x}).$$

Thus we have

$$\mathrm{P}\left(h_{\mathrm{NN}}(X) \neq Y | X = \boldsymbol{x}\right) \to 2\eta_1(\boldsymbol{x})\eta_2(\boldsymbol{x}).$$

This relation holds pointwise at all $\boldsymbol{x} \in \mathbb{R}^D$.

Now we integrate. Let Γ_1^* and Γ_2^* be the regions in $\mathbb{R}^D$ associated with where the Bayes classifier takes values 1 and 2 respectively:

$$\Gamma_k^* = \{\boldsymbol{x}\ :\ h^*(\boldsymbol{x}) = k\}, \quad k = 1, 2.$$

[3] Implicit here is the fact that the example data points $\{(X_n, Y_n)\}$ are independent from the unlabeled point (X, Y) we are analyzing.

[4] Obviously, this is where we are being mathematically imprecise. For the full argument, see Chapter 5 of *A Probabilistic Theory of Pattern Recognition* by Devroye, Györfi, and Lugosi (Springer, 1996).

The probability of error of the nearest-neighbor classifier is thus

$$\begin{aligned} R(h_{\mathrm{NN}}) &= \mathrm{P}\left(h_{\mathrm{NN}}(X) \neq Y\right) \\ &\to 2\int_{\boldsymbol{x}\in\Gamma_1^*} \eta_1(\boldsymbol{x})\eta_2(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} + 2\int_{\boldsymbol{x}\in\Gamma_2^*} \eta_1(\boldsymbol{x})\eta_2(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \\ &\leq 2\int_{\boldsymbol{x}\in\Gamma_1^*} \eta_2(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} + 2\int_{\boldsymbol{x}\in\Gamma_2^*} \eta_1(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \\ &= 2R(h^*) \end{aligned}$$

where the first inequality comes from the fact that both $\eta_1(\boldsymbol{x}), \eta_2(\boldsymbol{x}) \leq 1$, and the second by the definition of the Bayes classifier.