Some distributions for Bayesian analysis

Dror Baron

This supplement adds some information about some possible distributions that can be used for data in Bayesian analysis.

Bayesian classification: Suppose that we are given a posterior distribution for the data X given that the valued of the class G is $k \in \{1, \ldots, K\}$. We can denote this posterior distribution by

$$f_k(X) = f(X|G = k).$$

If we don't know anything about the distribution of the classes, then the posterior distribution for the class given the data X, i.e., $\Pr(G = k|X)$, is equal to $f_k(X)$, as we will see below. With all classes seeming equiprobable, either because we truly believe that they are equiprobable or due to our ignorance or even refusal to offer an opinion for the prior probabilities of different classes, the straightforward approach to classification is to select the $k \in \{1, \ldots, K\}$ that maximizes $f_k(X)$. This may be familiar to some of you as maximum likelihood estimation.

The seminal contribution by Bayes was to appreciate that when beliefs or priors about different classes having different probabilities are indeed available, the probability that the data was generated by class k must incorporate both our prior belief and the observed data X. To analyze $\Pr(G = k|X)$, we first define $\pi_k, k \in \{1, \ldots, K\}$ as our prior for class k. Because $\{\pi_1, \pi_2, \ldots, \pi_K\}$ is a probability mass function (PMF), each such value must be non-negative, i.e., $\pi_k \geq 0$, and they sum to one, i.e., $\sum_{k=1}^{K} \pi_k = 1$. Bayes' analysis of $\Pr(G = k|X)$ proceeds as follows,

$$\begin{aligned}
\Pr(G = k|X) &= \frac{f(G = k, X)}{f(X)} && (1)\\
&= \frac{f(G = k, X)}{\sum_{k'=1}^{K} f(G = k', X)} && (2)\\
&= \frac{\pi_k f(X|G = k)}{\sum_{k'=1}^{K} \pi_{k'} f(X|G = k')} && (3)\\
&= \frac{\pi_k f_k(X)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(X)}. && (4)
\end{aligned}$$

where (1) is due to the joint density $f(G = k, X)$ (meaning that both $G = k$ and X occur) being the product of $f(X)$ and the posterior density $\Pr(G = k|X)$, (2) relies on $f(X)$ being partitioned into a sum of K joint densities corresponding to the K classes, (3) resembles (1)

and expresses the joint density $f(G = k, X)$ as a product of π_k and $f(X|G = k)$, and (4) utilizes our notation, $f_k(X) = f(X|G = k)$.

Possible distributions: If we know $f_k(X)$, then Bayes' rule (4) offers a convenient closed form expression for computing the probabilities for the different classes. However, in practice we often do not know the precise form of $f_k(X)$, and it is estimated from the data. One possible distribution for $f_k(X)$ is a multivariate Gaussian density.[1] We could keep things simple and assume that the covariance matrices for all classes are identical, in which case we will soon see that linear discriminant analysis offers decision regions between classes whose borders are straight lines. In contrast, if we allow the K covariance matrices to differ, then the borders become second order quadratic curves.

A more complicated distribution is the mixture Gaussian, where $f_k(X)$ is a sum of several Gaussian densities. In many cases, decision boundaries between classes whose densities are mixture Gaussian are quite nonlinear.

Finally, a naive Bayes approach assumes that the multivariate pdf is the product of the p individual distributions governing the p coordinates of X. To help appreciate such a "naive" distribution, consider a simple example involving two random variables (RVs): (*i*) X_1 is 1 when the height of a student exceeds 1.70 meters (approximately 5'8"), else 0, and (*ii*) X_2 is 1 when the weight of a student exceeds 70 Kg (approximately 154 pounds), else 0. Real life experience suggests that these variables are correlated. In words, a taller student is more likely to be heavier, and vice versa. Possible probabilities for this pair of RVs could be $\Pr(X_1 = 0, X_2 = 0) = 0.3$, $\Pr(X_1 = 0, X_2 = 1) = 0.2$, $\Pr(X_1 = 1, X_2 = 0) = 0.2$, and $\Pr(X_1 = 1, X_2 = 1) = 0.3$. We can see that $\Pr(X_1 = 1) = \Pr(X_1 = 1, X_2 = 0) + \Pr(X_1 = 1, X_2 = 1) = 0.2 + 0.3 = 0.5$. Similarly, $\Pr(X_2 = 1) = \Pr(X_1 = 0) = \Pr(X_2 = 0) = 0.5$. However, conditioned on X_2 being 0, Bayes' rule informs us that $\Pr(X_1 = 1|X_2 = 0) = \Pr(X_1 = 1, X_2 = 0)/\Pr(X_2 = 0) = 0.2/0.5 = 0.4$. That is,

$$\Pr(X_1 = 1|X_2 = 0) = 0.4 \neq 0.5 = \Pr(X_1 = 1),$$

meaning that X_1 and X_2 are dependent. In contrast, if the RVs are independent, it means that

$$\Pr(X_1 = \alpha|X_2 = \beta) = \Pr(X_1 = \alpha) \quad \text{and} \quad \Pr(X_2 = \beta|X_1 = \alpha) = \Pr(X_2 = \beta)$$

for all pairs $(\alpha, \beta) \in \{0, 1\} \times \{0, 1\}$. The naive Bayes model requires this latter condition involving independence. It might be quite a strong assumption to assume that RVs are *precisely* statistically independent. That said, many p-tuples of RVs can be *approximated* as independent in practice.

Another point worth mentioning is that statistical independence is a relatively severe requirement, although (again), in many situations the RVs can be approximated as such. In contrast, linear uncorrelatedness is a weaker condition. That is, the set of distributions that are statistically independent is a subset of the set of uncorrelated distributions. Another way to express this point is that there might be statistically dependent RVs that are nonetheless linearly uncorrelated.

[1] If X contains a single random variable, meaning that the dimension p of the data is one, then $f_k(X)$ would be a univariate Gaussian pdf. The multivariate case implicitly assumes that $p > 1$.