Machine Learning, Spring 2019 Chao Lan
EN 2105 EN 4084A
MWF 9-9:50am clan@uwyo.edu

Nearest Neighbor Classifiers

Recall the bayes decision rule is

$$p(y = k \mid x) \geq p(y = c \mid x), \forall k \neq c. \tag{1}$$

The nearest neighbor classifiers can be interpreted as adopting the bayes decision rule. But first we will introduce its heuristic.

Let S be a set of instances in $\mathbb{R}^p$, and $z \in \mathbb{R}^p$ be a query instance. The k nearest neighbors of z are the instances (in S) whose distances to z are smaller than other instances. A common choice of distance measure is Euclidean distance, i.e.,

$$d(x, z) = \sqrt{\sum\nolimits_{i=1}^{p} (x_{\cdot i} - z_{\cdot i})^2} = ||x - z||_2. \tag{2}$$

The k-nearest neighbor (kNN) classifier assigns z to class c if most of its k nearest neighbors are from that class. For example, if we pick up five neighbors of z and find that three are from class 1 and two are from class 2, then z will be assigned to class 1. This classification rule is called majority voting. The neighborhood size k is a hyperparameter.

[*Discussion*] What to do if there is a tie in voting?

[*Discussion*] How would k affect the model complexity of kNN?

kNN does not have a model to fit. It simply stores all training data to classify testing data. Thus it is sometimes called 'lazy classifier', 'memory-based classifier', or 'non-parameteric classifier'. It can suffer from high memory and computational costs.

A variant of kNN is fixed-radius nearest neighbor classifier. Instead of collecting votes from k nearest neighbors, it collects votes from the nearest neighbors whose distances to z are smaller than some radius ϵ. This guarantees the voters are sufficiently similar to z (so they are more likely to be generated from the same distribution as z and thus representative of $p(y \mid z)$).

Statistical Justification of kNN

Under some assumptions, the kNN classification rule is equivalent to the bayes decision rule (1).

Let us draw a sphere centered at x and containing exactly k neighbors. Let k_c be the number of neighbors from class c. Let n be the total number of training examples, and n_c be number of training examples from class c. Let $p(x)$ be the pdf of the population. Let V be the volume of the sphere. Let P be the probability mass of the sphere, which indicates how likely an instance will fall in the sphere. We can estimate P in two ways:

(1) $P = \frac{k}{n}$, because k instances (out of n) fall in the sphere

(2) $P = Vp(x)$, because if the sphere is small, it is reasonable to assume $p(x)$ is constant in it.

Combining both, we have $\frac{k}{n} = Vp(x)$ and

$$p(x) = \frac{k}{nV} \tag{3}$$

Similarly analysis applies to neighbors from a particular class c. That gives

$$p(x \mid y = c) = \frac{k_c}{n_c V} \tag{4}$$

[*Exercise*] Verify (4).

On the other hand, we have

$$p(y = c) = \frac{n_c}{n}. \tag{5}$$

Putting (4), (3) and (5) together, and applying the Bayes' theorem, we have

$$p(y = c \mid x) = \frac{p(x \mid y = c)p(y = c)}{p(x)} = \frac{\frac{k_c}{n_c V}\frac{n_c}{n}}{\frac{k}{nV}} = \frac{k_c}{k}. \tag{6}$$

Recall kNN assigns x to class c if $k_c \geq k_{c'}$ for any class c'. Based on (6), this is equivalent to

$$p(y = c \mid x) \geq p(y = c' \mid x). \tag{7}$$

This justifies the kNN classification rule is equivalent to the bayes decision rule, and thus can minimize the bayes error.