6.4 Kernel Classification Rules

Slightly different from the k-NN rules that assign a positive weight to the k nearest neighbors of x, positive weights can be assigned to the points within a certain fixed distance from x. For instance, such a *moving window* classifier with the width of window h is defined by

$$f_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n I(\|x - x_i\| \le h) I(y_i = 1) > \sum_{i=1}^n I(\|x - x_i\| \le h) I(y_i = 0) \\ 0 & \text{otherwise.} \end{cases}$$

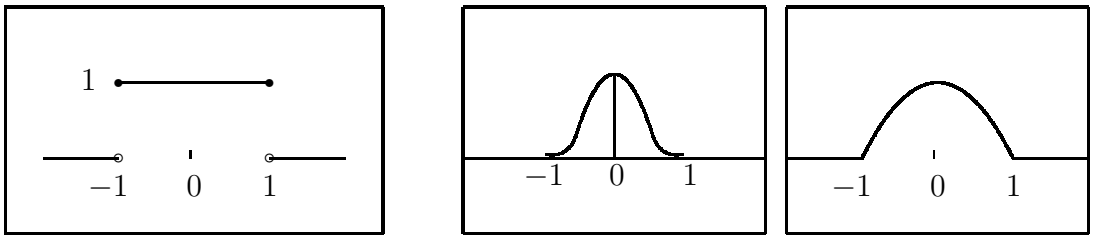
Furthermore, it is sensible to have weights decrease in the distance of x_i from x in a smooth fashion. This consideration leads to smooth weight functions called a *kernel function*, $K : \mathbb{R}^d \to \mathbb{R}$, which are usually non-negative and monotonically decreasing along rays starting from the origin. A few examples of such kernel functions K are

(a) **Uniform kernel**: $K(x) = I(\|x\| \le 1)$

(b) **Gaussian kernel**: $K(x) = \exp(-\|x\|^2)$

(c) **Epanechnikov kernel**: $K(x) = (1 - \|x\|^2) I(\|x\| \le 1)$.

See Figure 9 below for the kernel functions.



(a) Uniform (b) Gaussian (c) Epanechnikov

Figure 9: Examples of kernel functions

With such a kernel function K for weighting data points, the kernel classification rule is defined as

$$f_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n K(\frac{x-x_i}{h}) I(y_i = 1) > \sum_{i=1}^n K(\frac{x-x_i}{h}) I(y_i = 0) \\ 0 & \text{otherwise.} \end{cases} \tag{6.13}$$

Here h is a positive parameter called the bandwidth, which controls how local or global the rule is. For large values of h, the kernel gives positive weights to a larger number of points around x, and so the decision rule becomes more stable and global, whereas for small values of h, only the points very near to x contribute to the decision rule. Akin to the bias-variance trade-off in regression, one faces the same issue in choosing h for the kernel classification rule data-adaptively.

If $K(x) \ge 0$, the values of the kernel can be normalized to obtain proper weights given by

$$W_{ni}(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^n K(\frac{x-x_j}{h})}$$

so that $W_{ni}(x) \geq 0$ and $\sum W_{ni}(x) = 1$. From Stone's theorem for universal consistency of the decision rules based on locally weighted average estimators, some regularity conditions of kernels can be deduced for consistency of the kernel classification rules.

Let $B_r(x)$ denote the closed ball of radius $r > 0$ centered at $x \in \mathcal{X}$, that is,

$$B_r(x) := \{y \in \mathcal{X} : \|y - x\| \leq r\}.$$

Definition 9 (Kernel regularity conditions)**.** A kernel $K(\cdot)$ is called regular if

(i) $K(x) \geq 0$ for all x.

(ii) There exists an $r > 0$ and a constant $\beta > 0$ such that

$$K(x) \geq \beta I(x \in B_r(0)).$$

(iii) For the r in (ii),

$$\int_{\mathcal{X}} \sup_{y \in B_r(x)} K(y)dx < \infty.$$

Condition (ii) of the above definition basically means that there exists at least a small neighborhood around the origin where the kernel is strictly positive.

Proposition 2. *If there exists a k^* such that $K \leq k^*$ and K has a compact support, then such a kernel is regular.*

Proof. Suppose that K has a compact support A. Consider a cover of A, $\{B_r(x) : x \in A\}$. Then by the virtue of compactness, there exists a finite subcover determined by $\{x_1, \ldots, x_N\} \subset A$ such that

$$A \subset \bigcup_{i=1}^{N} B_r(x_i).$$

Then, for $x \notin \bigcup_{i=1}^{N} B_{2r}(x_i)$, $\inf_{y \in A} \|x - y\| > r$. To see this, suppose $\inf_{y \in A} \|x - y\| \leq r$. Then, there exists $\{y_n\} \subset A$ such that $\lim_{n \to \infty} \|x - y_n\| \leq r$. Now, for each n,

$$\begin{aligned} \|x - x_i\| &\leq \|x - y_n\| + \|y_n - x_i\| \\ \Rightarrow \min_{1 \leq i \leq N} \|x - x_i\| &\leq \|x - y_n\| + \min_{1 \leq i \leq N} \|y_n - x_i\| \\ &\leq \|x - y_n\| + r \qquad \text{by the assumption.} \end{aligned}$$

Taking limits on both sides, we get $\min_{1 \leq i \leq N} \|x - x_i\| \leq 2\,r$, which means that there exists i_0 such that $x \in B_{2\,r}(x_{i_0})$. It contradicts $x \notin \cup_{i=1}^{N} B_{2r}(x_i)$.

Therefore

$$\begin{aligned} \int \sup_{y \in B_r(x)} K(y)dx &= \int_{\bigcup_{i=1}^{N} B_{2\,r}(x_i)} \sup_{y \in B_r(x)} K(y)dx \leq \int_{\bigcup_{i=1}^{N} B_{2\,r}(x_i)} k^* dx \\ &= k^* \text{ Vol} \left(\bigcup_{i=1}^{N} B_{2\,r}(x_i) \right) < \infty. \end{aligned}$$

□

The following proposition gives a little more general conditions for regularity of kernels functions covering the Gaussian kernel, for instance.

Proposition 3. *If $K(x) = L(\|x\|)$ for some bounded non-negative function $L(\cdot)$, L is decreasing on $[0, \infty)$ and $\int K(x)dx < \infty$, then K is regular.*

Under the regularity conditions, the following theorem states the universal consistency of kernel classification rules.

Theorem 20. *Assume that K is a regular kernel. If $h_n \to 0$ and $n\, h_n^d \to \infty$ as $n \to \infty$, then for all distributions*

$$ER(f_n) \to R^*$$

where f_n is the kernel classification rule defined in (6.13).

Remark 13. Note that the bandwidth h_n should decrease with n for universal consistency. However, it should not decrease too fast. $h_n \to 0$ guarantees the local nature of the decision while $h_n^d \to \infty$ ensure small variance.

In practice, when d is large, very large values of h are necessary to control statistical variation of kernel classification rules but then, the local nature of the decision rules would be lost. For illustration, let $X \sim \text{Uniform}([0,1]^d)$. Consider a hypercubical neighborhood of x to capture a fraction α of the data. Since the volume of a hypercube of edge length l, is l^d, we need a neighborhood of $\alpha^{1/d}$ edge length on average. When $\alpha = 1\%$ and $d = 10$, $\alpha^{1/d} \approx 0.63$. When $\alpha = 10\%$ and $d = 10$, $\alpha^{1/d} \approx 0.8$. Thus we can see that even for a modest α, neighborhoods have to be very large.

The consistency of kernel classification rules is theoretically reassuring. However, the theorem above does not provide any means to choose an optimum h for kernel classification rules just as the theorem of universal consistency of k-NN rules. In practice, an optimal value of h (or k) has to be chosen in a data-dependent manner.