Machine Learning Topic 8

Kernel Methods

Aimee Barciauskas

29 March 2016

The moving window rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} \mathbb{I}_{Y_i=0, X_i \in S_{x,h}} \geq \sum_{i=1}^{n} \mathbb{I}_{Y_i=1, X_i \in S_{x,h}} \\ 1 & \text{otherwise} \end{cases}$$

The kernel-classification rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} \mathbb{I}_{Y_i=0} K(\frac{x-X_i}{h}) \geq \sum_{i=1}^{n} \mathbb{I}_{Y_i=1} K(\frac{x-X_i}{h}) \\ 1 & \text{otherwise} \end{cases}$$

Clearly, the kernel rule is a generalization of the moving window rule, since taking the special kernel $K(x) = \mathbb{I}_{x \in S_{0,1}}$ yields the moving window rule.

We state the universal consistency theorem for a large class of kernel functions, namely, for all regular kernels:

Definition 10.1. The kernel K is called regular if it is nonnegative, and there is a ball S_0, r of radius $r > 0$ centered at the origin, and constant $b > 0$ such that $K(x) \geq b\mathbb{I}_{S_0,r}$ and $\int sup_{y \in x+S_0} K(y)dx < \infty$

Theorem 10.1

Assume that K is a regular kernel. If $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$,

then for any distribution of (X, Y), and for every $\epsilon > 0$ there is an integer n_0 such that for $n > n_0$ for the error probability L_n of the kernel rule:

$P\{L_n - L^* > \epsilon\} \leq 2e^{-n\epsilon^2/32\rho^2}$

where the constant ρ depends on the kernel K and the dimension only. Thus, the kernel rule is strongly universally consistent.

Trivial example

Take $n = 1$ and $h = 1$, we have the classifier:

$$g_1(x) = \begin{cases} 0 & \text{if } Y = 0, |x - X_1| \text{ or if } Y = 1, |x - X_1| \geq 1 \\ 1 & \text{otherwise} \end{cases}$$

If $K > 0$ everywhere:

$\mathbb{E}L_1 = \mathbb{P}\{Y_1 = 0, Y = 1\} + \mathbb{P}\{Y_1 = 1, Y = 0\} = 2\mathbb{E}[\eta(x)]\mathbb{E}[1 - \eta(x)]$

which may be $\frac{1}{2}$ (if the expected value of $\eta(x)$ is $\frac{1}{2}$) even if L^* is 0 (which happens when $\eta \in \{0, 1\}$ everywhere). If $K \equiv 1$, we ignore the X_i's and take a majority vote:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} \mathbb{I}_{Y_i=0} \geq \sum_{i=1}^{n} \mathbb{I}_{Y_i=1} \\ 1 & \text{otherwise} \end{cases}$$

Let N_n be the number of Y_i's equal to zero. As N_n is binomial $(n, 1-p)$ with $p = \mathbb{E}\eta(X) = \mathbb{E}\{Y = 1\}$, we see that:

$$\mathbb{E}L_n = p\mathbb{P}\left\{N_n \geq \frac{n}{2}\right\} + (1-p)\mathbb{P}\left\{N_n \leq \frac{n}{2}\right\} \to min(p, 1-p)$$

It is interesting to note the following:

$\mathbb{E} = 2p(1-p)$

$= 2min(p(1-p))(1 - min(p, 1-p))$

$\leq 2p(1-p)$

$= 2 \lim_{n\to\infty} \mathbb{E}L_n$

The expected error with one observation is at most twice as bad as the expected error with an infinite sequence.