3.1 Kernel estimators and the curse of dimensionality

$$\eta(x) = \mathbb{E}(Y|X = x)$$
$$g_*(x) = \operatorname{sign}(\eta(x)).$$

Let $\widehat{\eta}(x)$ be an estimator of η. How good is the classifier $\widehat{g} = \operatorname{sign}(\widehat{\eta})$?

$$\begin{aligned}\mathcal{E}(\widehat{g}) &= P(Y \neq \widehat{g}(x)) - P(Y \neq g_*(x)) \\ &= \int_{x:\widehat{g}(x)\neq g_*(x)} |\eta(x)|\, d\Pi(x) \\ &\leq \int_{\mathbb{S}} |\widehat{\eta}(x) - \eta(x)|\, d\Pi(x).\end{aligned}$$

Assume that $X \in \mathbb{R}^d$ and Π is absolutely continuous with respect to the Lebesgue measure $p(\cdot)$. Further assume that p is Lipschitz continuous with Lipschitz constant L, i.e.

$$|p(x) - p(y)| \leq L|x - y|_2.$$

Let $X_1, \ldots, X_n$ be iid "copies" of X, i.e. drawn from the same distribution as X.

3.1.1 Kernel Estimators

Let $K : \mathbb{R}^d \to \mathbb{R}$ have the following properties:

1. $\int_{\mathbb{R}^d} K(x)\, dx = 1$
2. $\int_{\mathbb{R}^d} x_j K(x)\, dx_j = 0$ for all $j = 1 \ldots, d$
3. $\int_{\mathbb{R}^d} ||x||_2^2 K(x)\, dx < \infty$

Then K is called a **kernel**.

Example 1. $K(x) = I\{||x||_\infty \leq 1/2\}$ an indicator on the d-cube centered at the origin.

For any kernel $K(x)$ we can define

$$K_h(x) = \frac{1}{h^d} K(x/h).$$

Consider the convolution

$$(p * K_h)(x) = \int_{\mathbb{R}^d} p(x-y) K_h(y)\, dy.$$

By property 1 in the definition of kernel,

$$\begin{aligned}
|(p * K_h)(x) - p(x)| &= \left| \int_{\mathbb{R}^d} (p(x-y) - p(x)) K_h(y)\, dy \right| \\
&\leq \int |p(x-y) - p(x)| \frac{1}{h^d}\, dy \\
&\leq L \int ||y||_2 |K_h(y)|\, dy \\
&\leq Lh \int_{\mathbb{R}^d} K(y/h) ||y/h||^2\, d(y/h) \\
&= hLC(K).
\end{aligned}$$

Since convolution is symmetric, we have

$$(p * K_h)(x) = \int K_h(x-y) p(y)\, dy = \mathbb{E}(K_h(x - X)).$$

We can therefore define a kernel density estimator to be

$$\begin{aligned}
\widehat{p}_n(x) &= \frac{1}{n} \sum_{j=1}^{n} K_h(x - X_j) \\
&= \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)
\end{aligned}$$

since $\mathbb{E}(\widehat{p}_n(x)) = (p * K_h)(x)$.

Note the estimator is flexible: we may choose h to balance the estimator's bias and variance. Smaller h corresponds to smaller bias and larger variance (notice the h^d in the denominator of the density estimator).

We have

$$\eta(x) = \mathbb{E}(Y|X = x) = \int y\, d\Pi(y|X = x).$$

If (X, Y) have joint density $p(x, y)$, then

$$\eta(x) = \int y \frac{p(x,y)}{\int p(x,y)\, dy} dx$$

where $\frac{p(x,y)}{\int p(x,y)\,dy} = p(y|x)$. We can use kernel estimation to estimate this conditional probability density.

Suppose the marginal $p(x)$ is known. Consider

$$\widehat{\eta}_h(x) = \frac{1}{nh^d}\sum_{j=1}^{n} Y_j \frac{K(\frac{x-X_j}{h})}{p(x)}.$$

Then

$$\begin{aligned}
\mathbb{E}\widehat{\eta}_h(x) &= \mathbb{E}Y_1 \frac{K(\frac{x-X_1}{h})}{p(x)}\frac{1}{h^d} \\
&= \mathbb{E}\left[\mathbb{E}[Y_1 \frac{K(\frac{x-X_1}{h})}{p(x)}\frac{1}{h^d}|X_1]\right] \\
&= \mathbb{E}\frac{K(\frac{x-X_1}{h})}{p(x)}\frac{1}{h^d}\eta(X_1) \\
&= \int K\left(\frac{x-y}{h}\right)\frac{1}{h^d}\frac{1}{p(x)}p(y)\,dy...
\end{aligned}$$