Notes for *Elements of Statistical Learning*

Sebastian Claici

November 25, 2024

1 Chapter 2: Overview of Supervised Learning

Statistical Decision Theory

The goal is to minimize the expected prediction error:

$$\begin{aligned}\mathrm{EPE}(f) &= \mathrm{E}\left(Y - f(X)\right)^2 \\ &= \int [y - f(x)]^2 p(x, y)\, \mathrm{d}x\, \mathrm{d}y\end{aligned}$$

If we break down the expectation as $E_{X,Y} = E_X E_{Y|X=x}$ we can rewrite this as

$$\begin{aligned}\mathrm{EPE}(f) &= \mathrm{E}_X \mathrm{E}_{Y|X=x}(Y - f(X))^2 \\ &= \int_X \int_Y [y - f(x)]^2 p(y|X = x) p(x)\, \mathrm{d}y\, \mathrm{d}x \\ &= \int_X p(x) \left(\int_Y [y - f(x)]^2 p(y|X = x)\, \mathrm{d}y \right) \mathrm{d}x\end{aligned}$$

We have moved the dependence on $p(x)$ outside the inner expectation. Since f is unconstrained, we can solve for the optimal f pointwise. That is:

$$\arg\min_f \mathrm{EPE}(f) = \arg\min_c \int_Y [y - c]^2 p(y|X = x)\, \mathrm{d}y$$

Differentiating wrt c and using the fact that

$$\int_Y y\ p(y|X = x)\, \mathrm{d}y = \mathrm{E}(Y|X = x)$$

gives us (2.13) in the book.

Nearest-neighbor methods try to model the regression function directly by averaging predictions around the query point x. To drive this point home, we can show that $\mathrm{NN}(x) \to x$ as the number of training points $N \to \infty$.

To sketch this proof out, assume $x_1, \ldots, x_N$ are drawn i.i.d from X. We want to bound $\min_i \|x - x_i\|$, but since this is a bit complicated, let's instead compute

$$P(\|x - x_i\| \geq \epsilon, \forall i).$$

for some $\epsilon > 0$.

Since the x_i are sampled independently, we can expand the probability as

$$P(\|x - x_i\| \geq \epsilon, \forall i) = \prod_{i=1}^{N} P(\|x - x_i\| \geq \epsilon).$$

As the x_i are also identically distributed, the product can be written as

$$\left(P(\|x - x_i\| \geq \epsilon)\right)^N$$

which goes to 0 as $N \to \infty$ as long as the probability is not exactly 1. This shows that with infinite samples the Nearest-neighbor of x is x and so nearest neighbors yields the Bayes optimal decision boundary even with a single neighbor.

However, we often do not have enough samples to use a model-free approach to regression. The second proposal is to assume the regression function is linear in its arguments:

$$f(x) \approx x^T \beta$$