

Notes for *Elements of Statistical Learning*

Sebastian Claici

November 25, 2024

Chapter 2: Overview of Supervised Learning

Statistical Decision Theory

The goal is to minimize the expected prediction error:

$$\begin{aligned} \text{EPE}(f) &= \mathbb{E} (Y - f(X))^2 \\ &= \int [y - f(x)]^2 p(x, y) \, dx \, dy \end{aligned} \tag{1}$$

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

$$\begin{aligned} \text{EPE}(f) &= E_X E_{Y|X=x} (Y - f(X))^2 \\ &= \int_X \int_Y [y - f(x)]^2 p(y|X=x) p(x) \, dy \, dx \\ &= \int_X p(x) \left(\int_Y [y - f(x)]^2 p(y|X=x) \, dy \right) \, dx \end{aligned}$$

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:

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

Differentiating wrt c and using the fact that

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

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 $\text{NN}(x) \rightarrow x$ as the number of training points $N \rightarrow \infty$.

To sketch this proof out, assume x_1, \dots, 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

$$[P(\|x - x_i\| \geq \epsilon)]^N$$

which goes to 0 as $N \rightarrow \infty$ 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:

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

If we plug this for f into (1), we get

$$\int [y - x^T \beta]^2 p(x, y) \, dx \, dy.$$

We can differentiate this wrt β ¹

$$\begin{aligned} \frac{\partial \text{EPE}}{\partial \beta} &= 2 \int x[y - x^T \beta] p(x, y) \, dx \, dy \\ &= 2 \left(\int xy \, p(x, y) \, dx \, dy - \int xx^T \beta \, p(x, y) \, dx \, dy \right) \end{aligned}$$

Since β is not a random variable, we can set this to 0 to arrive at the minimizer in (2.16) in the book:

$$\beta = [E(XX^T)]^{-1} E(XY)$$

¹See this [link](#) for a review of matrix calculus.