

# Notes for *Elements of Statistical Learning*

Sebastian Claici

November 29, 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

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

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

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

for some  $\varepsilon > 0$ .

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

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

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

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

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  $\beta$ <sup>1</sup>

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

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

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

---

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

## Bias-Variance Decomposition

We can express the mean-squared error in terms of a squared bias term and a variance term. In equation (2.25) in the book, these vary w.r.t. the training set  $T$ . To clarify the notation a bit,  $x_0$  is the point 0,  $\hat{y}_0$  is the model estimate (in this case the nearest neighbor estimate), and  $f(x_0)$  is the true value at 0, but the following derivation holds generally for any model approximation  $\hat{y}$  of a function  $f(x)$ <sup>2</sup>:

$$\begin{aligned}\text{MSE}(x_0) &= \mathbb{E}_T [f(x_0) - \hat{y}_0]^2 \\ &= \mathbb{E}_T [\hat{y}_0 - \mathbb{E}_T[\hat{y}_0] + \mathbb{E}_T[\hat{y}_0] - f(x_0)]^2 \\ &= \mathbb{E}_T [\hat{y}_0 - \mathbb{E}_T[\hat{y}_0]]^2 + (f(x_0) - \mathbb{E}_T[\hat{y}_0])^2 \\ &= \text{Var}_T(\hat{y}_0) + \text{Bias}^2(\hat{y}_0)\end{aligned}$$

It is a somewhat instructive exercise to figure out how to go from the second line to the third. Easiest if you recall that

$$\begin{aligned}\mathbb{E}_T[\mathbb{E}_T[y]] &= \mathbb{E}_T[y] \\ \mathbb{E}_T[f(x)] &= f(x)\end{aligned}$$

In the example in the book, the variance is consistently low, but the bias increases with dimension as the nearest point to 0 becomes increasingly distant.

We can discuss equations (2.27) and (2.28) in the book briefly. We have

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + \varepsilon) = \beta + (X^T X)^{-1} X^T \varepsilon$$

and thus

$$\hat{y}_0 = x_0^T \hat{\beta} = x_0^T (\beta + (X^T X)^{-1} X^T \varepsilon)$$

which gives

$$\hat{y}_0 = x_0^T \beta + \sum_{i=1}^N l_i(x_0) \varepsilon_i \tag{2}$$

since  $x_0^T (X^T X)^{-1} X^T \varepsilon$  is a scalar and

$$(x_0^T (X^T X)^{-1} X^T)^T = X (X^T X)^{-1} x_0$$

to give the expression in the book.

Let's write out  $\text{EPE}(x_0)$ . Note that because the true data was generated from a noisy process, we have to integrate out for  $y_0$  given a fixed  $x_0$ :

$$\text{EPE}(x_0) = \mathbb{E}_{y_0|x_0} \mathbb{E}_T [y_0 - \hat{y}_0]^2$$

---

<sup>2</sup>See, for example, the [wikipedia](#) page.

In this particular case, because  $Y$  depends on  $X$  stochastically,  $E_T = E_X E_{Y|X}$

We will write out  $y_0 - \hat{y}_0$  as:

$$y_0 - \hat{y}_0 = (y_0 - x_0^T \beta) + (x_0^T \beta - E_T[\hat{y}_0]) + (E_T[\hat{y}_0] - \hat{y}_0)$$

Let's square this, and keep in mind that  $E[\varepsilon] = 0$ , and  $\text{Var}[\varepsilon] = \sigma^2$ :

$$E_{y_0|x_0} E_T[y_0 - \hat{y}_0]^2 = E_{y_0|x_0} [y_0 - x_0^T \beta]^2 + (x_0^T \beta - E_T[\hat{y}_0])^2 + E_T[\hat{y}_0 - E_T[\hat{y}_0]]^2 + \text{cross terms}$$

For the cross terms, we notice the following:

$$\begin{aligned} E_{y_0|x_0} (y_0 - x_0^T \beta) &= 0 \\ E_T(E_T[\hat{y}_0] - \hat{y}_0) &= 0 \end{aligned}$$

and

$$\begin{aligned} (x_0^T \beta - E_T[\hat{y}_0]) &= E_T \left[ \sum_{i=1}^N l_i(x_0) \varepsilon_i \right] \\ &= E_X \left[ \sum_{i=1}^N l_i(x) E_{Y|X}(\varepsilon_i) \right] \\ &= 0 \end{aligned} \tag{3}$$

where we have used (2).

This gives

$$\begin{aligned} E_{y_0|x_0} E_T[y_0 - \hat{y}_0]^2 &= E_{y_0|x_0} [y_0 - x_0^T \beta]^2 + (x_0^T \beta - E_T[\hat{y}_0])^2 + E_T[\hat{y}_0 - E_T[\hat{y}_0]]^2 + \\ &= \text{Var}[y_0|x_0] + \text{Bias}^2(\hat{y}_0) + \text{Var}_T(\hat{y}_0) \end{aligned}$$

but the bias is 0 by (3), and  $\text{Var}[y_0|x_0] = \sigma^2$ , we have:

$$E_{y_0|x_0} E_T[y_0 - \hat{y}_0]^2 = \sigma^2 + \text{Var}_T(\hat{y}_0).$$

To finish the derivation, let's write out  $\text{Var}_T(\hat{y}_0)$ . We have just proved that  $E_T(\hat{y}_0) = x_0^T \beta$ , and so

$$\begin{aligned} \text{Var}_T(\hat{y}_0) &= E_T [x_0^T (X^T X)^{-1} X^T \varepsilon]^2 \\ &= E_T [x_0^T (X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1} x_0] \end{aligned}$$

Since  $\varepsilon \sim N(0, \sigma^2)$ ,  $\varepsilon \varepsilon^T = \sigma^2 I_N$ , and we can replace above:

$$\begin{aligned} \text{Var}_T(\hat{y}_0) &= E_T [x_0^T (X^T X)^{-1} X^T X (X^T X)^{-1} x_0] \sigma^2 \\ &= E_T [x_0^T (X^T X)^{-1} x_0] \sigma^2 \end{aligned}$$

which is the value in the book.

## Exercises