Notes for *Elements of Statistical Learning*

Sebastian Claici

December 8, 2024

Contents

2 Overview of Supervised Learning

2.1 Statistical Decision Theory

The goal is to minimize the expected prediction error:

$$\begin{aligned}\mathrm{EPE}(f) &= \mathrm{E}\left(Y - f(X)\right)^2 \\ &= \int [y - f(x)]^2 p(x, y)\,\mathrm{d}x\,\mathrm{d}y \end{aligned} \tag{1}$$

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

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

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

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

for some $\varepsilon > 0$.

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

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

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

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

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:

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

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

$$\int [y - x^T\beta]^2 p(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

We can differentiate this wrt β[1]

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

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

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

2.2 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 $f(x)$[2]:

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

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

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

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

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

and thus

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

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

[2] See, for example, the wikipedia page.

which gives

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

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

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

to give the expression in the book.

Let's write out $\mathrm{EPE}(x_0)$. 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:

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

In this particular case, because Y depends on X stochastically, $\mathrm{E}_T = \mathrm{E}_X\mathrm{E}_{Y|X}$

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

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

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

$$\begin{aligned}\mathrm{E}_{y_0|x_0}\mathrm{E}_T[y_0 - \hat{y}_0]^2 =& \mathrm{E}_{y_0|x_0}[y_0 - x_0^T\beta]^2 + (x_0^T\beta - \mathrm{E}_T[\hat{y}_0])^2 + \mathrm{E}_T[\hat{y}_0 - \mathrm{E}_T[\hat{y}_0]]^2 + \\ & \text{cross terms}\end{aligned}$$

For the cross terms, we notice the following:

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

and

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

where we have used (2).

This gives

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

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

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

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

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

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

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

which is the value in the book.

To derive (2.28), we assume large N and that $X^TX \to N\text{Cov}(X)$, hence:

$$\begin{aligned}\text{E}_{x_0}\text{EPE}(x_0) &= \sigma^2 + \text{E}_{x_0}\left[x_0^T(X^TX)^{-1}x_0\right]\sigma^2 \\ &\sim \sigma^2 + \text{E}_{x_0}\left[x_0^T\text{Cov}(X)^{-1}x_0\right]\sigma^2/N \qquad (4)\end{aligned}$$

Now $x_0^T\text{Cov}(X)^{-1}x_0$ is a scalar, and can be written as $\text{trace}[x_0^T\text{Cov}(X)^{-1}x_0]$. Exploiting the cyclic properties of the trace operator and its linearity (so we can move the expectation inside):

$$\begin{aligned}\text{E}_{x_0}\text{trace}[x_0^T\text{Cov}(X)^{-1}x_0] &= \text{E}_{x_0}\text{trace}[\text{Cov}(X)^{-1}x_0x_0^T] \\ &= \text{trace}[\text{Cov}(X)^{-1}\text{E}_{x_0}(x_0x_0^T)] \\ &= \text{trace}[\text{Cov}(X)^{-1}\text{Cov}(x_0)]\end{aligned}$$

Since $\text{Cov}(X) = \text{Cov}(x_0)$, and since each training set point is p-dimensional:

$$\text{trace}[\text{Cov}(X)^{-1}\text{Cov}(x_0)] = \text{trace}[I_p] = p$$

which when replaced in (4) gives equation (2.28) in the book.

2.3 Exercises

2.1. (Classifying to max)

The wording here is a bit confusing, but I am assuming that $\hat{y}$ is constant, and we want to show

$$\arg\min_k \|\hat{y} - t_k\| = \arg\max_k \hat{y}_k.$$

To prove this, note that $\|t_k\| = 1, \forall k$, and the length of $\hat{y}$ is constant with $0 \leq \hat{y}_k \leq 1$. For this setup, we can square the cost function without affecting the optimal value. Then

$$\min_k \|\hat{y} - t_k\|^2 = \min_k \langle \hat{y} - t_k, \hat{y} - t_k \rangle = \|t_k\|^2 + \|\hat{y}\|^2 + 2 \min_k (-\langle \hat{y}, t_k \rangle)$$

Since $\min_k(-\langle \hat{y}, t_k \rangle) = \max_k \langle \hat{y}, t_k \rangle$ and $\langle \hat{y}, t_k \rangle = \hat{y}_k$ we have

$$\arg\min_k \|\hat{y} - t_k\|^2 = \arg\max_k \langle \hat{y}, t_k \rangle = \arg\max_k \hat{y}_k.$$

2.2. (Bayes decision boundary)

Computing the Bayes decision boundary only makes sense after the means of the two classes have been sampled. If we integrate out that step, the decision boundary is the diagonal through the origin that separates $[0, 1]^T$ and $[1, 0]^T$.

The procedure for sampling each data point looks something like this (for the blue means):

$$\begin{aligned}
(\mu_1, \ldots, \mu_{10}) &\sim \mathcal{N}([1, 0]^T, I_n) \\
\text{for i} = 1..100 & \\
k &\sim \text{Cat}(10, [1/10, \ldots, 1/10]) \\
x &\sim \mathcal{N}(\mu_k, I_n/5) \\
\text{end for} &
\end{aligned}$$

The decision boundary is where

$$\text{P(blue)} \sum_{i=1}^{10} \exp\left\{-5\|x - \mu_k\|^2/2\right\} = \text{P(orange)} \sum_{i=1}^{10} \exp\left\{-5\|x - \nu_k\|^2/2\right\}$$

which can be computed easily.

2.3. (Median distance in high dimensions)

Let x be a point sampled from the unit hypersphere. The probability that $\|x\| \leq r$ is equal to the ratio between the volume of the radius r hypersphere and the volume of the unit sphere. That is, the cdf of $\|x\|$ in d-dimensions is

$$F(r) = r^d.$$

Now let's say we sample n points $x_1, x_2, \ldots, x_n$. The probability that the closest of these is distance r away can be written as

$$F(\min_i \|x\|_i \leq r) = 1 - (1 - F(r))^n = 1 - (1 - r^d)^n \tag{5}$$

since each of the x_i is independent, and the event that the closest is distance r away is the same as the event that all of them are at least distance r away.

The median is defined as the distance r for which $F(\min_i \|x\|_i \leq r) = 1/2$. If we substitute in (5), we have

$$\begin{aligned} 1 - (1 - r^d)^n &= \frac{1}{2} \\ 1 - r^d &= \left(\frac{1}{2}\right)^{1/n} \\ r &= \left(1 - \left(\frac{1}{2}\right)^{1/n}\right)^{1/d} \end{aligned}$$

To compute the distance to the mean is a bit more of a challenge. Recall that the pdf of a random variable is the derivative of the cdf, and so, from (5) we have

$$p(\min_i \|x\|_i \leq r) = nd(1 - r^d)^{n-1} r^{d-1}$$

The mean distance is found by integrating this over $r \in [0, 1]$:

$$d_{\text{mean}} = nd \int_0^1 (1 - r^d)^{n-1} r^{d-1} \, \mathrm{d}r$$

Integrating this explicitly is impossible, but we can relate it to the Beta function:

$$\mathrm{B}(z_1, z_2) = \int_0^1 t^{z_1 - 1} (1 - t)^{z_2 - 1} \, \mathrm{d}t$$

which can be integrated numerically and is available in most programming languages (e.g. scipy.special.beta).

2.4. (Average projection distance)

The standard normal distribution is invariant to rotations, and so there is no difference between projecting on $a = x_0/\|x_0\|$ and projecting on any of the axes. Since $\mathbf{x} \sim \mathcal{N}(0, I_p)$ is isotropic, each of the components of $\mathbf{x}$ are distributed as $\mathcal{N}(0, 1)$. Or, if you prefer a slightly different point of view, the joint distribution of $x_i \sim \mathcal{N}(0, 1)$ sampled independently is a 0-mean, identity covariance Gaussian in p dimensions.

We need to prove two related facts:

1. $\mathrm{E}[x_i^2] = 1, \forall i$
2. $\mathrm{E}[\|x_0\|^2] = p$

The first item follows from the definition of the standard deviation of a Gaussian, but let's derive from first principles as a reminder. We want to integrate

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} \, \mathrm{d}x$$

We can integrate this by parts with

$$\begin{aligned} u &= x \\ v &= -e^{-\frac{x^2}{2}} \\ \mathrm{d}u &= \mathrm{d}x \\ \mathrm{d}v &= x e^{-\frac{x^2}{2}} \, \mathrm{d}x \end{aligned}$$

and so

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} \, \mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \, \mathrm{d}v = \frac{1}{\sqrt{2\pi}} \left[uv\right]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} v \, \mathrm{d}u \\ &= \frac{1}{\sqrt{2\pi}} \left[-x e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, \mathrm{d}x \end{aligned}$$

The first summand is 0 since the exponential dominates the linear term and both limits are 0. The second summand is related to the Gaussian integral after the transformation $t = x/\sqrt{2}$, $\mathrm{d}x = \sqrt{2} \, \mathrm{d}t$ which gives

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2} e^{-t^2} \, \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \sqrt{2} \sqrt{\pi} = 1$$

This proves $\mathrm{E}[x_i^2] = 1$. To show that the distance from any sample point to the origin is p, we can write out:

$$\mathrm{E}\left[\|x_0\|^2\right] = \mathrm{E}\left[\sum_{i=1}^{p} x_i^2\right] = \sum_{i=1}^{p} \mathrm{E}\left[x_i^2\right] = p$$

2.5. (Bias-variance tradeoff for linear models)

See main text for the derivation of the two equations.

2.6. (Repeated values and weighted least squares)

I will introduce some notation to make life easier. Let N_u be the number of unique values of x, n_i be the number of repetitions of x_i, and let y_{ij}, $j \in \{1, \ldots, n_i\}$ represent the output values at those x_i. The RSS can be written as

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} \left(y_{ij} - f_\theta(x_i)\right)^2$$

We want to show that the minimizer here is equivalent to the minimizer of a weighted least squares problem:

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N_u} \left(\bar{y}_i - w_i f_\theta(x_i)\right)^2 .$$

The problem setup inclines us towards the following definition of $\bar{y}_i$:

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

Let's expand the quadratic of the original RSS, and simplify:

$$\begin{aligned}
\mathrm{RSS}(\theta) &= \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} \left(y_{ij}^2 + f_\theta(x_i)^2 - 2 y_{ij} f_\theta(x_i)\right) \\
&= \sum_{i=1}^{N_u} n_i f_\theta(x_i)^2 - 2 \sum_{i=1}^{N_u} n_i f_\theta(x_i) \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} + \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} y_{ij}^2
\end{aligned}$$

If we collect the first two terms, we have

$$\mathrm{RSS}(\theta) = n_i \sum_{i=1}^{N_u} \left(f_\theta(x_i)^2 - 2 f_\theta(x_i) \bar{y}_i\right) + \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} y_{ij}^2$$

and to complete the square, we add and subtract $\sum_{i=1}^{N_u} \hat{y}_i^2$:

$$\mathrm{RSS}(\theta) = n_i \sum_{i=1}^{N_u} \left(f_\theta(x_i) - \bar{y}_i\right)^2 + \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{N_u} \bar{y}_i^2$$

Since the last two terms in this equation do not contain θ, the overall minimizer is independent of those terms, and we conclude

$$\arg\min_\theta \mathrm{RSS}(\theta) = \arg\min_\theta \sum_{i=1}^{N_u} n_i (f_\theta(x_i) - \bar{y}_i)^2$$

This is a *weighted least squares* problem because each input sample x_i is fit to the average output value at that sample, and we weight residual errors by how often x_i appears. It is a reduced problem because $N_u < N$, the original number of samples.

2.7. (Bias-variance decompositions)