Notes on statistical learning

AvanDavad

November 6, 2020

1 Linear models and least squares

On page 12 we have that the residual sum of squares:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{1}$$

How can we differentiate with respect to β?

$$\mathrm{RSS}(\beta) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \tag{2}$$

We have the following rules for differenciating w.r.t a vector:

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \frac{d}{d\mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \mathbf{y} \tag{3}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{x} \tag{4}$$

Using these rules we can differenciate RSS:

$$\frac{d}{d\beta}\mathrm{RSS}(\beta) = 0 - (\mathbf{y}^T\mathbf{X})^T - \mathbf{X}^T\mathbf{y} + (\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)\beta \tag{5}$$

$$\frac{d}{d\beta}\mathrm{RSS}(\beta) = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta \tag{6}$$

Setting this to zero we get the normal equations:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta \tag{7}$$

2 Statistical decision theory

On page 19 we have that

$$\beta = [E(XX^T)]^{-1}E(XY) \tag{8}$$

but how exactly do we get this equation? In general, we have the expected prediction error:

$$\mathrm{EPE}(f) = E(Y - f(X))^2 \tag{9}$$

And we have that the prediction function is linear:

$$f(X) = X^T\beta \tag{10}$$

We seek a β for minimizing the expected prediction error. X and Y are random variables, X being a vector, Y being a scalar.

$$\begin{aligned} \frac{d}{d\beta}\mathrm{EPE} = \frac{d}{d\beta}\mathrm{E}((Y - X^T\beta)^2) = \mathrm{E}\left(\frac{d}{d\beta}(Y - X^T\beta)^2\right) \\ = \mathrm{E}\left(2(Y - X^T\beta)\cdot(-X)\right) = -2\mathrm{E}(YX) + 2\mathrm{E}(X(X^T\beta)) \\ = -2\mathrm{E}(YX) + 2\mathrm{E}((XX^T)\beta) = -2\mathrm{E}(YX) + 2(\mathrm{E}(XX^T))\beta \end{aligned} \tag{11}$$

We used the fact that the expected value is linear, and that β is not random, so we could factor out from the expected value. Setting this to zero we have that:

$$\mathrm{E}(YX) = \mathrm{E}(XX^T)\beta \tag{12}$$

which yields

$$\hat{\beta} = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \tag{13}$$

2.1 application. Simple linear fit.

Let's see an application for this equation. Let $X = \begin{bmatrix} x \\ 1 \end{bmatrix}$, $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$.

Now $f(x) = a \cdot x + b$

$$XY = \begin{bmatrix} x \cdot y \\ y \end{bmatrix} \tag{14}$$

$$XX^T = \begin{bmatrix} x^2 & x \\ x & 1 \end{bmatrix} \tag{15}$$

If we have N datapoints $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$, we can approximate the expectation values.

$$\mathrm{E}(XY) \approx \frac{1}{N}\begin{bmatrix} \sum_i x_i \cdot y_i \\ \Sigma_i y_i \end{bmatrix} \tag{16}$$

$$\mathrm{E}(XX^T) \approx \frac{1}{N} \begin{bmatrix} \sum_i x_i^2 & \Sigma_i x_i \\ \Sigma_i x_i & N \end{bmatrix} \tag{17}$$

Let's denote the followings:

$$\alpha_X = \sum_i x_i \tag{18}$$

$$\alpha_Y = \sum_i y_i \tag{19}$$

$$\alpha_{XY} = \sum_i x_i y_i \tag{20}$$

$$\alpha_{X^2} = \sum_i x_i^2 \tag{21}$$

With these notations:

$$\mathrm{E}(XY) \approx \frac{1}{N} \begin{bmatrix} \alpha_{XY} \\ \alpha_Y \end{bmatrix} \tag{22}$$

$$\mathrm{E}(XX^T) \approx \frac{1}{N} \begin{bmatrix} \alpha_{X^2} & \alpha_X \\ \alpha_X & N \end{bmatrix} \tag{23}$$

Inverting $\mathrm{E}(XX^T)$:

$$[\mathrm{E}(XX^T)]^{-1} \approx \frac{N}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N & -\alpha_X \\ -\alpha_X & \alpha_{X^2} \end{bmatrix} \tag{24}$$

Plug these in to the equation:

$$\hat{\beta} = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \tag{25}$$

$$\hat{\beta} \approx \frac{N}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N & -\alpha_X \\ -\alpha_X & \alpha_{X^2} \end{bmatrix} \frac{1}{N} \begin{bmatrix} \alpha_{XY} \\ \alpha_Y \end{bmatrix} \tag{26}$$

$$\hat{\beta} \approx \frac{1}{N\alpha_{X^2} - \alpha_X^2} \begin{bmatrix} N\alpha_{XY} - \alpha_X \alpha_Y \\ \alpha_{X^2}\alpha_Y - \alpha_X \alpha_{XY} \end{bmatrix} \tag{27}$$

From here we can get $\hat{a}$ and $\hat{b}$, since $\hat{\beta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$

3 E$|Y-c|$ and the median

On page 20, it asks the question "What happens if we replace the L_2 loss function with the L_1 : E$|Y-f(X)|$?" Let's investigate this question.

3.1 discrete case

We can get rid of the conditional $X=x$, and just ask the question: What c will minimize E$|Y-c|$? Denote this function with g, so $g(c)=$ E$|Y-c|$. Let's look at two examples.

Example 1. The random variable Y takes 4 possible values with probabilities $\frac{1}{7}$, $\frac{1}{7}$, $\frac{3}{7}$, $\frac{2}{7}$. The figure below shows the probability mass funciton.

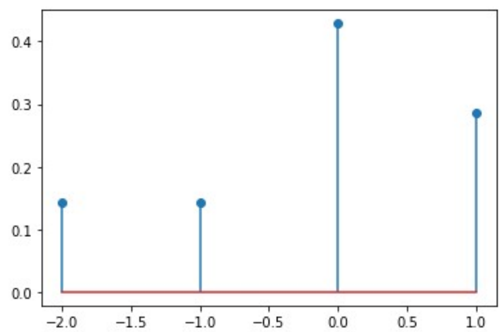

Figure 1: probability mass function of the first example random variable.

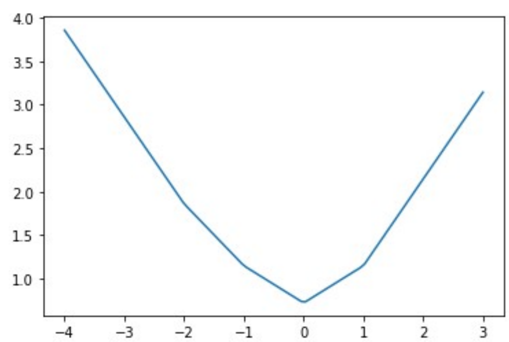

Figure 2: $g(c)$ function. The horizontal axis is c.

Example 2. The random variable Y takes 4 possible values with probabilities 0.1, 0.4, 0.3, 0.2. The figure below shows the probability mass funciton.

We can see that $g(c)$ is a piecewise linear function, and it has minimum, which is a point or a line segment. Let's say we have Y discrete random variable

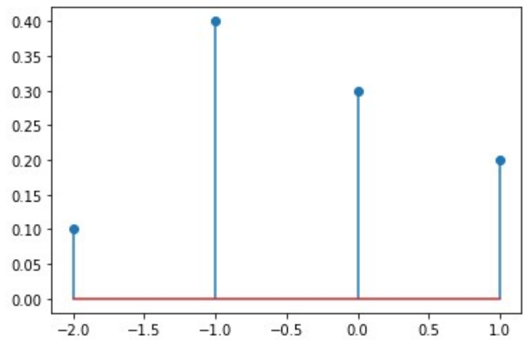

Figure 3: probability mass function of the second example random variable.

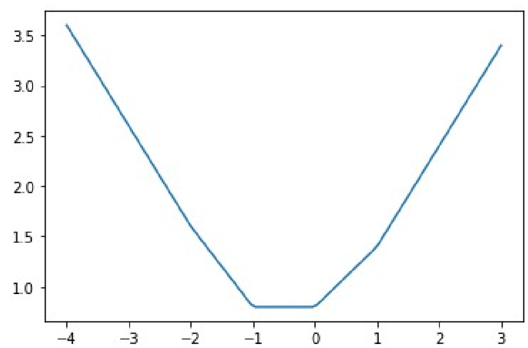

Figure 4: $g(c)$ function. The horizontal axis is c.

that takes values from $S = \{x_1, x_2, \dots x_n\}$. The values are ordered: $x_1 < x_2 < \dots < x_n$. Y takes these values with corresponding probabilities p_1, p_2, ..., p_n.

Let's calculate the equation of the piecewise linear function. Denote the interval I_k such that $x \in I_k$ if and only if k values from S are smaller than x. So $I_0 = (-\infty, x_1]$, $I_1 = [x_1, x_2]$, ..., $I_n = [x_n, \inf)$.

$$g(c) = \mathrm{E}|Y - c| = \Sigma_{i=1}^{n} p_i \cdot |x_i - c| \tag{28}$$

If $c \in I_k$, then

$$g(c) = \mathrm{E}|Y - c| = \Sigma_{i=1}^{k} p_i \cdot (c - x_i) + \Sigma_{i=k+1}^{n} p_i \cdot (x_i - c) \tag{29}$$

$$g(c) = c \cdot (\Sigma_{i=1}^{k} p_i - \Sigma_{i=k+1}^{n} p_i) + (\Sigma_{i=k+1}^{n} p_i x_i - \Sigma_{i=1}^{k} p_i x_i) \tag{30}$$

First we can show that this function is continuous. On one hand ($c \in I_k = [x_k, x_k + 1]$):

$$g_1 = g(x_k) = x_k \cdot (\Sigma_{i=1}^{k} p_i - \Sigma_{i=k+1}^{n} p_i) + (\Sigma_{i=k+1}^{n} p_i x_i - \Sigma_{i=1}^{k} p_i x_i) \tag{31}$$

On the other hand we have ($c \in I_{k-1} = [x_{k-1}, x_k]$):

$$g_2 = g(x_k) = x_k \cdot (\Sigma_{i=1}^{k-1} p_i - \Sigma_{i=k}^{n} p_i) + (\Sigma_{i=k}^{n} p_i x_i - \Sigma_{i=1}^{k-1} p_i x_i) \tag{32}$$

$$g_1 - g_2 = x_k \cdot (p_k + p_k) - p_k x_k - p_k x_k = 0 \tag{33}$$

Now that we showed that this function is continuous, let's find it's minimum. Since it is piecewise linear, its derivative is piecewise constant. Denote the derivative of g on the interval I_k with $g'(I_k)$.

$$\begin{aligned} g'(I_k) &= -1 \\ g'(I_1) &= -1 + 2p_1 \\ g'(I_2) &= -1 + 2p_1 + 2p_2 \\ &\dots \\ g'(I_n) &= -1 + 2p_1 + ... + 2p_n = 1 \end{aligned} \tag{34}$$

So the derivative is increasing from -1 to $+1$. We can distinguish two possibilities. First, assume that the derivative is never zero. In this case, we have a k where $g'(I_{k-1}) < 0$ but $g'(I_k) > 0$, so the minimum is at x_k, the median. The second case is where there is an interval where the derivative is zero. In this case the whole interval is minimum, again, the median.

3.2 continuous case

Let's have the following function:

$$f(x) = \int_a^x g(x,t)dt \tag{35}$$

I state without proof that the derivative of this function is as follows:

$$f'(x) = \int_a^x \frac{\partial g(x,t)}{\partial x} dt + g(x,x) \tag{36}$$

Now we have that

$$g(c) = \mathrm{E}(|Y - c||X = x) = \int_{-\infty}^{c} (c-y) f_{Y|X}(y|x)dy + \int_c^{\infty} (y-c) f_{Y|X}(y|x)dy \tag{37}$$

$$g'(c) = \int_{-\infty}^{c} f_{Y|X}(y|x)dy + \int_{\infty}^{c} f_{Y|X}(y|x)dy \tag{38}$$

Setting this to zero, we get that

$$\int_{-\infty}^{c} f_{Y|X}(y|x)dy = \int_c^{\infty} f_{Y|X}(y|x)dy \tag{39}$$

$$P(Y < c \mid X = x) = P(Y > c \mid X = x) \tag{40}$$

Again, this means the minimum is at the median.

4 Local methods in high dimensions

4.1 deriving the prediction formula

On page 24 we see an example of a linear data with noise. At first I was confused how it gets $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^N l_i(x_0)\epsilon_i$, where $l_i(x_0)$ is the ith element of $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$

In this example we make N experiments, storing the X_i values in the rows of $\mathbf{X}$, and we have also ϵ_i (elements of $\vec{\epsilon}$) and $Y_i = X_i^T\beta + \epsilon_i$ for some fixed β. For approximating β, we use the result:

$$\beta = [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \tag{41}$$

in this case it will be an approximation, since we have noise (ϵ).

Calculate first $\mathrm{E}(YX)$:

$$\mathrm{E}(YX) = \mathrm{E}((X^T\beta + \epsilon)X) = \mathrm{E}(XX^T)\beta + \mathrm{E}(\epsilon X) \tag{42}$$

Substitute this into the approximation of β:

$$\begin{aligned}\hat{\beta} &= [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(YX) \\ &= [\mathrm{E}(XX^T)]^{-1}(\mathrm{E}(XX^T)\beta + \mathrm{E}(\epsilon X)) \\ &= \beta + [\mathrm{E}(XX^T)]^{-1}\mathrm{E}(\epsilon X)\end{aligned} \tag{43}$$

We do not know of course the exact expectation values, but we have N data samples (training data). So how could we approximate the expectation values? Use the averages:

$$\mathrm{E}(\epsilon X)_i \approx \frac{1}{N}\sum_{k=1}^N \mathbf{X}_{ki}\epsilon_k \to \mathrm{E}(\epsilon X) \approx \frac{1}{N}\mathbf{X}^T\vec{\epsilon} \tag{44}$$

similarly,

$$[\mathrm{E}(XX^T)]^{-1} \approx N \cdot (\mathbf{X}^T\mathbf{X})^{-1} \tag{45}$$

putting these all together, we have:

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\epsilon} = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0 \tag{46}$$

and this is the formula that was to be explained.

4.2 equation (2.47) on page 37

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}\left((Y - \hat{f}_k(x_0))^2 | X = x_0\right) \\ &= \mathrm{E}\left((f(x_0) + \epsilon_0 - \hat{f}_k(x_0))^2\right)\end{aligned} \tag{47}$$

The data points are fixed: $x_1, x_2, \ldots, x_N$. Denote the closest data point to x_0 as $x_{(1)}$, the second closest $x_{(2)}$, etc. With this notation, the nearest neighbor estimate for $f(x_0)$:

$$\hat{f}_k(x_0) = \frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)}) + \epsilon_l) \tag{48}$$

In this equation, $f(x_{(l)})$ is fixed, and epsilons are iid random variables.

$$\begin{aligned}\mathrm{EPE}_k(x_0) &= \mathrm{E}\left(\left(f(x_0) + \epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)}) + \epsilon_l)\right)^2\right)\\ &= \mathrm{E}\left(\left[\left(f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right) + \left(\epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right)\right]^2\right)\\ = \mathrm{E}\left([\hat{F} + \hat{E}]^2\right) &= \mathrm{E}\left(\hat{F}^2 + 2\cdot\hat{F}\hat{E} + \hat{E}^2\right) = \hat{F}^2 + 2\hat{F}\cdot\mathrm{E}(\hat{E}) + \mathrm{E}(\hat{E}^2)\end{aligned} \tag{49}$$

where $\hat{F}$ is nonrandom, and $\hat{E}$ is random:

$$\begin{aligned}\hat{F} &\equiv f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)}),\\ \hat{E} &\equiv \epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}\epsilon_l\end{aligned} \tag{50}$$

Let's calculate the expectation of $\hat{E}$:

$$\mathrm{E}(\hat{E}) = \mathrm{E}\left(\epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right) = \mathrm{E}(\epsilon_0) - \frac{1}{k}\sum_{l=1}^{k}\mathrm{E}(\epsilon_l) = 0 - \frac{1}{k}\sum_{l=1}^{k} 0 = 0 \tag{51}$$

The expectation of $\hat{E}^2$:

$$\mathrm{E}(\hat{E}^2) = \mathrm{E}\left(\epsilon_0 - \frac{1}{k}\sum_{l=1}^{k}\epsilon_l\right)^2 = \mathrm{E}\left(\epsilon_0^2 + \sum_{l=1}^{k}\frac{\epsilon_l^2}{k^2} + CrossProducts\right) \tag{52}$$

The expectation of the cross products are zero, since epsilons are independent, so $\mathrm{E}(\epsilon_i\epsilon_j) = \mathrm{E}\epsilon_i \cdot \mathrm{E}\epsilon_j = 0 \cdot 0 = 0$

$$\begin{aligned}\mathrm{E}(\hat{E}^2) = \mathrm{E}\left(\epsilon_0^2 + \sum_{l=1}^{k}\frac{\epsilon_l^2}{k^2}\right) &= \mathrm{E}\epsilon_0^2 + \sum_{l=1}^{k}\frac{\mathrm{E}\epsilon_l^2}{k^2}\\ &= \sigma^2 + \sum_{l=1}^{k}\frac{\sigma^2}{k^2} = \sigma^2 + \frac{\sigma^2}{k}\end{aligned} \tag{53}$$

We used the fact that the error has zero mean, so the variance is $\sigma^2 = \mathrm{Var}(\epsilon) = \mathrm{E}(\epsilon^2) - (\mathrm{E}\epsilon)^2 = \mathrm{E}(\epsilon^2)$. So the final form is:

$$\begin{aligned} \mathrm{E}_k(x_0) &= \hat{F}^2 + 2\hat{F} \cdot \mathrm{E}(\hat{E}) + \mathrm{E}(\hat{E}^2) = \hat{F}^2 + \mathrm{E}(\hat{E}^2) \\ &= \left(f(x_0) - \frac{1}{k} \sum_{l=1}^{k} f(x_{(l)}) \right)^2 + \sigma^2 + \frac{\sigma^2}{k} \end{aligned} \tag{54}$$

5 Solutions for the Exercises of chapter 2

5.1 Ex. 2.2

We have $X \in \mathbb{R}^p$ continuous and G discrete random variables. Assume we have K classes. Each class has its own distribution, let's say that class g has a pdf $f_g(x)$ ($x \in \mathbb{R}^p$). When generating points, we first choose a class with associated probabilities $p_1, p_2, \ldots, p_K$ ($\sum p_i = 1$). When we have chosen the class, we generate a point with the appropriate distribution.

The Bayes classifier classifies each point x to the most probable class. So let's calculate the probability of class g, given the point. It should be noted that when I write $\mathrm{P}(x)$, I mean "the probability that the chosen point is in the infinitesimal neighborhood of x". So I should write $\mathrm{P}(X \in b_{dx}(x))$, i.e., the probability that X is in the dx-volume ball around x. If the pdf was $f(x)$, this probability is $f(x)dx$. But instead, I'll write $\mathrm{P}(x) = f(x)$. Likewise, when I write $\mathrm{P}(g)$, I mean $\mathrm{P}(G = g)$.

$$\mathrm{P}(g|x) = \frac{\mathrm{P}(g \cap x)}{\mathrm{P}(x)} = \frac{\mathrm{P}(g \cap x)}{\mathrm{P}(x)} = \frac{\mathrm{P}(x|g)\mathrm{P}(g)}{\sum_{g'} \mathrm{P}(x|g')\mathrm{P}(g')} \tag{55}$$

The denominator is a normalizing constant, so the chosen class, for which $\mathrm{P}(g|x)$ is maximum:

$$\hat{g}(x) = max_g \mathrm{P}(x|g)\mathrm{P}(g) \tag{56}$$

5.2 Ex. 2.3

Given a unit ball in p-dimension. We sample N data points from it uniformly. Let X be the distance from the origin. The pdf must be proportional to x^{p-1}, and integrating it from 0 to 1 gives 1, thus the pdf:

$$f(x) = p \cdot x^{p-1} \tag{57}$$

The probability that a random sample is at least x distant from the origin is:

$$\mathrm{P}(X > x) = \int_x^1 f(x)dx = 1 - x^p \tag{58}$$

The probability that all N sample points are further from origin as x:

$$\mathrm{P}(X_1 > x \cap X_2 > x \cap \dots \cap X_N > x) = (1 - x^p)^N \tag{59}$$

We seek and x for that this probability is a half (that will give us the median):

$$\begin{aligned} (1 - x^p)^N &= \frac{1}{2} \\ 1 - x^p &= \left(\frac{1}{2}\right)^{1/N} \\ \left[1 - \left(\frac{1}{2}\right)^{1/N}\right]^{1/p} &= x \end{aligned} \tag{60}$$

5.3 Ex. 2.4

If we choose a as the first unit base vector ($a = [1, 0, 0, \dots, 0]^T$), then $a^T \cdot x_i$ is the first coordinate of x_i. It is by definition (standard) normally distributed. Since the distribution is spherically symmetric, we can choose any direction a, $a^T \cdot x_i$ remains standard normal.

I created an experiment on this. Created 1000 sample points in p dimension, and rotated them into the first 2 dimension, so that we can visualize the distances. On the first image below we can see that the points get further and further away from the origin as the dimension increases.

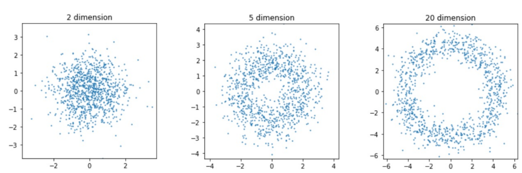

Figure 5: The sample points get further from the origin as we increase the dimension.

But this doesn't mean the points are close from each other. In the following experiment I took the random sample points, chose one of them and set it as the new origin. We can see that still the points are far from a random sample point as we increase the dimension.

5.4 Ex. 2.5

5.4.1 equation (2.27) on page 26

I won't use indices at the expectation sign, it always confuses me. So this is the expected prediction error:

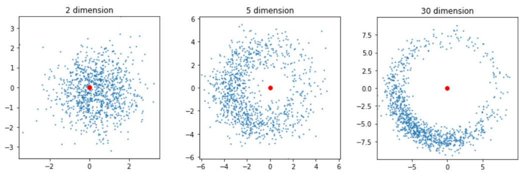

Figure 6: The points get further and further from each other (red dot is a randomly selected sample point) as we increase the dimension.

$$\mathrm{EPE}(x_0) = \mathrm{E}(y_0 - \hat{y}_0)^2 \tag{61}$$

Recall, that $y_0 = x_0^T\beta + \epsilon$ is a random variable, since $\epsilon \sim N(0, \sigma^2)$. This is the label (the ground truth) for x_0. The prediction that we make for x_0 is $\hat{y}_0 = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0$. x_0 is a p-vector, β is a p-vector, $\vec{\epsilon}$ is an n-vector, and $\mathbf{X}$ is a n by p matrix (each row is a training sample vector). $\hat{y}_0 = x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{Z}^T x_0$. Here I introduced the p by n matrix $\mathbf{Z}$:

$$\mathbf{Z} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{62}$$

In the expression of $\hat{y}_0$, $\vec{\epsilon}$ and $\mathbf{Z}$ are the only random variables. The elements of $\vec{\epsilon}$ are iid RVs. $\vec{\epsilon}$ and $\mathbf{Z}$ are independent. Let's calculate the expectation values of y_0 and $\hat{y}_0$:

$$\mathrm{E}(y_0) = \mathrm{E}(x_0^T\beta + \epsilon) = x_0^T\beta \tag{63}$$

$$\begin{aligned}
\mathrm{E}(\hat{y}_0) &= \mathrm{E}(x_0^T\beta + \vec{\epsilon}^T \cdot \mathbf{Z}^T x_0) \\
&= x_0^T\beta + \mathrm{E}(\vec{\epsilon}^T \cdot \mathbf{Z}^T x_0) \\
&= x_0^T\beta + \mathrm{E}(\vec{\epsilon}^T)\mathrm{E}(\mathbf{Z}^T x_0) \\
&= x_0^T\beta + \vec{0}^T \cdot \mathrm{E}(\mathbf{Z}^T x_0) \\
&= x_0^T\beta
\end{aligned} \tag{64}$$

Here I used that $\vec{\epsilon}$ and $\mathbf{Z}$ are independent, so the expectation value of their product is the product of their expectations. For simplicity, denote $\mu \equiv \mathrm{E}(y_0) = \mathrm{E}(\hat{y}_0) = x_0^T\beta$. The expected prediction error:

$$\begin{aligned}
\mathrm{EPE}(x_0) &= \mathrm{E}(y_0 - \mu + \mu - \hat{y}_0)^2 \\
&= \mathrm{E}(y_0 - \mu)^2 - 2 \cdot \mathrm{E}((y_0 - \mu)(\hat{y}_0 - \mu)) + \mathrm{E}(\hat{y}_0 - \mu)^2 \\
&= \mathrm{E}(y_0 - \mu)^2 + 0 + \mathrm{E}(\hat{y}_0 - \mu)^2 \\
&= \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0)
\end{aligned} \tag{65}$$

Note that y_0 and $\hat{y}_0$ are independent. The epsilon in y_0 is a scalar and is nothing to do with the vector epsilon in $\hat{y}_0$. This is why $\mathrm{E}((y_0 - \mu)(\hat{y}_0 - \mu))$ is zero. Now let's derive the variances:

$$\mathrm{Var}(y_0) = \mathrm{Var}(\mu + \epsilon) = \mathrm{Var}(\epsilon) = \sigma^2 \tag{66}$$

This was the easy part. It is much more difficult to get the variance for $\hat{y}_0$. We know that if a is a constant scalar, and X is a scalar random variable, then

$$\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X) \tag{67}$$

Okay, but what if a is a vector, just as X, and we take the inner product $a^T \cdot X$? What is the variance $\mathrm{Var}(a^T \cdot X)$?

$$\begin{aligned}
\mathrm{Var}(a^T \cdot X) &= \mathrm{Var}(a_1 \cdot X_1 + a_2 \cdot X_2 + \cdots + a_n \cdot X_n) \\
&= \mathrm{E}\left(\sum_i a_i \cdot (X_i - \mathrm{E}X_i)\right)^2 \\
&= \mathrm{E}\left(\sum_{i,j} a_i \cdot (X_i - \mathrm{E}X_i) \cdot a_j \cdot (X_j - \mathrm{E}X_j)\right) \\
&= \sum_{i,j} a_i \cdot a_j \cdot \mathrm{E}\left((X_i - \mathrm{E}X_i) \cdot (X_j - \mathrm{E}X_j)\right) \\
&= \sum_{i,j} a_i \cdot a_j \cdot \mathrm{Cov}(X_i, X_j) \\
&= a^T \cdot \mathrm{Cov}(X, X) \cdot a = a^T \cdot \Sigma \cdot a
\end{aligned} \tag{68}$$

Here $\Sigma \equiv \mathrm{Cov}(X, X)$ is the covariance matrix, $\mathrm{Cov}(X, X)_{i,j} = \mathrm{Cov}(X_i, X_j)$. Let's state our finding again. a is a constant vector, X is a random vector (same dimension), then:

$$\mathrm{Var}(a^T \cdot X) = a^T \cdot \mathrm{Cov}(X, X) \cdot a \tag{69}$$

Furthermore, we can write $\mathrm{Cov}(X, X) = \mathrm{E}((X - \mathrm{E}X) \cdot (X - \mathrm{E}X)^T) = \mathrm{E}(X \cdot X^T) - (\mathrm{E}X)(\mathrm{E}X^T)$. Now we can apply this to derive the variance of $\hat{y}_0$:

$$\begin{aligned}
\mathrm{Var}(\hat{y}_0) &= \mathrm{Var}(\mu + x_0^T \cdot \mathbf{Z} \cdot \vec{\epsilon}) = \mathrm{Var}(x_0^T \cdot \mathbf{Z} \cdot \vec{\epsilon}) \\
&= x_0^T \cdot \mathrm{Cov}(\mathbf{Z} \cdot \vec{\epsilon}, \mathbf{Z} \cdot \vec{\epsilon}) \cdot x_0 \\
&= x_0^T \cdot \Big(\mathrm{E}(\mathbf{Z}\vec{\epsilon} \cdot \vec{\epsilon}^T \mathbf{Z}^T) - \mathrm{E}(\mathbf{Z}\vec{\epsilon}) \cdot \mathrm{E}(\vec{\epsilon}^T \mathbf{Z}^T)\Big) \cdot x_0 \\
&= x_0^T \cdot \mathrm{E}(\mathbf{Z}\vec{\epsilon} \cdot \vec{\epsilon}^T \mathbf{Z}^T) \cdot x_0
\end{aligned} \tag{70}$$

Here we used the fact that $\mathbf{Z}$ and $\vec{\epsilon}$ are independent, so $\mathrm{E}(\mathbf{Z}\vec{\epsilon}) = \mathrm{E}\mathbf{Z} \cdot \mathrm{E}\vec{\epsilon} = \mathrm{E}\mathbf{Z} \cdot \vec{0} = \vec{0}$, and $\vec{0} \cdot \vec{0}^T = \mathbf{0}$, zero matrix.

$$
\begin{aligned}
\left(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T\right)_{i,j} &= \sum_{k,l}(\mathbf{Z})_{i,k}\cdot(\vec{\epsilon}\vec{\epsilon}^T)_{k,l}\cdot(\mathbf{Z}^T)_{l,j} \\
\rightarrow \mathrm{E}\left(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T\right)_{i,j} &= \sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot\epsilon_k\cdot\epsilon_l\cdot Z_{j,l}\right) \\
=\sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot Z_{j,l}\right)\cdot\mathrm{E}\left(\epsilon_k\cdot\epsilon_l\right) &= \sum_{k,l}\mathrm{E}\left(Z_{i,k}\cdot Z_{j,l}\right)\cdot\sigma^2\delta_{k,l} \\
=\sigma^2\cdot\sum_k\mathrm{E}\left(Z_{i,k}\cdot Z_{j,k}\right) = \sigma^2\cdot\mathrm{E}\sum_k\left(Z_{i,k}\cdot Z_{j,k}\right) &= \sigma^2\cdot\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)_{i,j} \\
\rightarrow \mathrm{E}(\mathbf{Z}\vec{\epsilon}\cdot\vec{\epsilon}^T\mathbf{Z}^T) &= \sigma^2\cdot\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)
\end{aligned} \tag{71}
$$

Substituting this into (70):

$$\mathrm{Var}(\hat{y}_0) = x_0^T\cdot\sigma^2\mathrm{E}(\mathbf{Z}\mathbf{Z}^T)\cdot x_0 \tag{72}$$

$$\mathrm{E}(\mathbf{Z}\mathbf{Z}^T) = \mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) = \mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\right) \tag{73}$$

Putting it all together:

$$\mathrm{EPE}(x_0) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 + \sigma^2\cdot x_0^T\cdot\mathrm{E}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)\cdot x_0 \tag{74}$$

And this is what we wanted to derive.

5.4.2 equation (2.28) on page 26

$$\mathrm{E}\left(x_0^T\mathrm{Cov}(X)^{-1}x_0\right) = \mathrm{E}\sum_{i,j}x_{0,i}\mathrm{Cov}(X)^{-1}_{i,j}x_{0,j} \tag{75}$$

Assuming that $x_0 \sim X$, i.e., x_0 (the test point) has the same distribution as X (the training data), and the expectation of it is the zero vector, $\mathrm{Cov}(x_0) = \mathrm{Cov}(X) = \mathrm{E}(x_0x_0^T)$.

$$
\begin{aligned}
\mathrm{E}\sum_{i,j}x_{0,i}\mathrm{Cov}(X)^{-1}_{i,j}x_{0,j} &= \mathrm{E}\sum_{i,j}\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(x_0)_{i,j} \\
=\mathrm{E}\sum_{i,j}\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(X)_{i,j} &= \mathrm{E}\sum_i\left(\sum_j\mathrm{Cov}(X)^{-1}_{i,j}\mathrm{Cov}(X)_{j,i}\right) \\
=\mathrm{E}\sum_i[\mathrm{Cov}(X)^{-1}\mathrm{Cov}(X)]_{i,i} &= \mathrm{E}\left(\mathrm{Trace}(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(X))\right) \\
&= \mathrm{E}\left(\mathrm{Trace}I_{pxp}\right) = p
\end{aligned} \tag{76}
$$

5.5 Ex. 2.6

Assume that we have n identical inputs $x_1 = x_2 = \cdots = x_n \equiv x$ with outputs $y_1, y_2, \ldots, y_n$. The least squares formula:

$$RSS(\theta) = \sum_{i=1}^{n} (y_i - f_\theta(x))^2 \tag{77}$$

The weighted least squares formula:

$$RSS_w(\theta) = n \cdot \left(\frac{\sum_{i=1}^{n} y_i}{n} - f_\theta(x) \right)^2 \tag{78}$$

I claim that the two expressions differ by a constant term that doesn't depend on θ, so both expressions lead to the same solution. This naturally extends to the case when we have groups of equal inputs.

Expanding RSS:

$$RSS(\theta) = \sum_{i=1}^{n} y_i^2 - 2f_\theta(x) \sum_{i=1}^{n} y_i + f_\theta^2(x) \tag{79}$$

Expanding RSS_w:

$$RSS_w(\theta) = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - 2f_\theta(x) \sum_{i=1}^{n} y_i + f_\theta^2(x) \tag{80}$$

So the difference of the 2 expressions is a constant that doesn't depend on θ. So when we derive wrt θ, we get the same formulae.

Whenever we have observations with identical values x, we can always refactor the RSS for the groups according to (77) → (78).

5.6 Ex. 2.7

Our estimator according to the problem statement:

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \tag{81}$$

a) For kNN, the weights are:

$$l_i(x_0; \mathcal{X}) = \frac{1}{k} \delta(x_i \in \text{kNN}(x_0)) \tag{82}$$

where $\delta(x_i \in \text{kNN}(x_0))$ is 1 if x_i is in the set of k-nearest neighbors of x_0, and 0 otherwise. So in this case we average the ys of the k-nearest neighbors of x_0.

For linear regression we have

$$\hat{f}(x_0) = x_0^T \beta \tag{83}$$

Where β comes from the following equation (see Section 2):

$$\beta = \mathrm{E}(XX^T)^{-1}\mathrm{E}(XY) \tag{84}$$

Now let's calculate this expression. We estimate the expectation values with averages.

$$\begin{aligned} \mathrm{E}(XX^T)^{-1} &\approx \left(\frac{1}{N}\sum_{j=1}^{N} x_j x_j^T\right)^{-1} \\ \mathrm{E}(XY) &\approx \frac{1}{N}\sum_{i=1}^{N} x_i y_i \end{aligned} \tag{85}$$

With these, we can formulate $\hat{f}(x_0)$ as follows:

$$\begin{aligned} \hat{f}(x_0) &= x_0^T \left(\frac{1}{N}\sum_{j=1}^{N} x_j x_j^T\right)^{-1} \cdot \frac{1}{N}\sum_{i=1}^{N} x_i y_i \\ &= \sum_{i=1}^{N} x_0^T \left(\sum_{j=1}^{N} x_j x_j^T\right)^{-1} x_i y_i \\ &\equiv \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \end{aligned} \tag{86}$$

From this we get the weights:

$$l_i(x_0; \mathcal{X}) = x_0^T \left(\sum_{j=1}^{N} x_j x_j^T\right)^{-1} x_i \tag{87}$$