ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

HS 2022

Prof. R. Wattenhofer
F. Grötschla

Computational Thinking
Sample Solutions to Exercise 9

0 Matrices and Differentiation (for Q3)

Claim 9.1. *Let $\boldsymbol{u} \in \mathbb{R}^{n\times 1}$ be a vector. Then $\boldsymbol{u}^T\boldsymbol{u} = \sum_{i=1}^n u_i^2$.*

Proof.

$$\begin{bmatrix} u_1 & u_2 & u_3 & \dots & u_n \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} = \sum_{i=1}^n u_i^2$$

□

So in particular

$$\sum_{(\boldsymbol{x},y)\in D} (y - \boldsymbol{w}^T\boldsymbol{x})^2 = \sum_{i=1}^n (y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})$$

and

$$\sum_{i=0}^{d-1} w_i^2 = \boldsymbol{w}^T\boldsymbol{w}$$

Claim 9.2. *Let $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, where $\boldsymbol{y}$ is $m \times 1$, $\boldsymbol{x}$ is $n \times 1$ and $\mathbf{A}$ is $m \times n$, and $\mathbf{A}$ does not depend on $\boldsymbol{x}$. Then*

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \mathbf{A}$$

Proof. Since the i^{th} element of $\boldsymbol{y}$ is given by

$$y_i = \sum_{j=1}^n A_{ik}x_k$$

it follows that

$$\frac{\partial y_i}{\partial x_j} = A_{ij}$$

for all $i = 1, \dots, m$ and $j = 1, \dots, n$. Therfore

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \mathbf{A}$$

□

Claim 9.3. *Let $f = \boldsymbol{y}^T \mathbf{A}\boldsymbol{x}$, where $\boldsymbol{y}$ is $m \times 1$, $\boldsymbol{x}$ is $n \times 1$ and $\mathbf{A}$ is $m \times n$, and $\mathbf{A}$ does not depend on $\boldsymbol{x}$ or $\boldsymbol{y}$. Then*

$$\frac{\partial f}{\partial \boldsymbol{x}} = \boldsymbol{y}^T \mathbf{A} \text{ and } \frac{\partial f}{\partial \boldsymbol{y}} = \boldsymbol{x}^T \mathbf{A}^T = (\mathbf{A}\boldsymbol{x})^T$$

Proof. Define $\boldsymbol{z}^T = \boldsymbol{y}^T \mathbf{A}$, then $f = \boldsymbol{z}^T \boldsymbol{x}$, so by Claim 9.2 we have

$$\frac{\partial f}{\partial \boldsymbol{x}} = \boldsymbol{z}^T = \boldsymbol{y}^T \mathbf{A}$$

Since $f = f^T = \boldsymbol{x}^T \mathbf{A}^T \boldsymbol{y}$, we similarly have by Claim 9.2 that

$$\frac{\partial f}{\partial \boldsymbol{y}} = \boldsymbol{x}^T \mathbf{A}^T$$

□

1 Linear Regression

Here is a dataset D with 3 samples. You want to fit a linear model of the form $\hat{f}(x) = w_0 + w_1 x$.

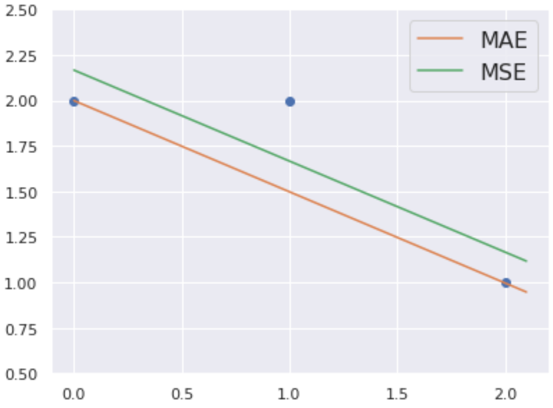

Figure 1: A dataset with 3 samples.

a) Which weights minimize the squared error loss? What is the total absolute error? What is the total squared error?

We can calculate these using the formulae from the lecture (Lemma 5.4 or Theorem 5.6). Using Lemma 5.4, we have

$$\begin{aligned}
\bar{x} &= \frac{0+1+2}{3} = 1 \\
\bar{y} &= \frac{2+2+1}{3} = \frac{5}{3} \\
w_1^* &= \frac{\sum_{(x,y)\in D}(y-\bar{y})(x-\bar{x})}{\sum_{(x,y)\in D}(x-\bar{x})^2} \\
&= \frac{(1/3)(-1)+0+(-2/3)(1)}{1+0+1} \\
&= -\frac{1}{2} \\
w_0^* &= \bar{y} - w_1^*\bar{x} \\
&= \frac{5}{3} + \frac{1}{2} \\
&= \frac{13}{6}
\end{aligned}$$

The errors are $(-\frac{1}{6}, \frac{2}{6}, -\frac{1}{6})$, so the total absolute error is $\frac{4}{6}$ and the total squared error is $\frac{1}{6}$

b) Which weights give a lower absolute error loss? What is the total absolute error? What is the total squared error?

In general there is no closed-form solution for minimizing the MAE. However we can immediately see that we can improve the absolute error loss by decreasing w_0. Notice that by decreasing it, we are reducing the error of 2 data points, whilst increasing the error of only 1 data point. In fact the line going through the points $(0,2)$ and $(2,1)$ is optimal. This line has weights $\boldsymbol{w} = (2, -\frac{1}{2})$, total absolute error $\frac{1}{2}$ and total squared error $\frac{1}{4}$.

*Note that if we have a single weight w_0, then the optimum must be when w_0 is equal to the median. Similarly with 2 weights, once we fix one of the weights (for example the slope w_1), then we know that the optimal line with this slope must go through the "middle/median" point.

2 Polynomial Regression

(e) $\hat{f} = w_0 + w_1 x + w_2 x^2 + w_3 x^3$. This model contains all the others so must give a bias at least as low as the others. The function f is of higher degree than quadratic and could in fact be cubic, so we expect $\hat{f} = w_0 + w_1 x + w_2 x^2 + w_3 x^3$ to have strictly lower bias than the other models.

Which model will give you the lowest variance?
(a) $\hat{f} = 3$ has 0 variance

3 Ridge Regression

a) Similarly as in Theorem 5.6, we can rewrite the Ridge loss in matrix form as

$$\frac{1}{n}(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) + \lambda \boldsymbol{w}^T \boldsymbol{w}$$

Again, we can differentiate with respect to $\boldsymbol{w}$ to find the optimal weights. Using the product rule and Claims 9.2 and 9.3 we get

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}} &= -\frac{1}{n}\left(\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})\right)^T - \frac{1}{n}(\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^T\mathbf{X} + 2\lambda \boldsymbol{w}^T \overset{!}{=} \mathbf{0}^T \\
\iff & -\frac{1}{n}\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) - \frac{1}{n}\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) + 2\lambda \boldsymbol{w} = \mathbf{0} \\
\iff & -\frac{2}{n}\mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) + 2\lambda \boldsymbol{w} = \mathbf{0} \\
\iff & \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{w}) - \lambda n \boldsymbol{w} = 0 \\
\iff & \mathbf{X}^T\boldsymbol{y} = \left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{w} + \lambda n \boldsymbol{w} \\
\iff & \mathbf{X}^T\boldsymbol{y} = \left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{w} + \lambda n \mathbf{I} \boldsymbol{w} \\
\iff & \mathbf{X}^T\boldsymbol{y} = \left(\mathbf{X}^T\mathbf{X} + \lambda n \mathbf{I}\right)\boldsymbol{w} \\
\iff & \boldsymbol{w} = \left(\mathbf{X}^T\mathbf{X} + \lambda n \mathbf{I}\right)^{-1}\mathbf{X}^T\boldsymbol{y}.
\end{aligned}$$

b) As $\lambda \to \infty$ the weights $\boldsymbol{w}^*_{ridge} \to 0$

c) As $\lambda \to 0$ the weights $\boldsymbol{w}^*_{ridge} \to \boldsymbol{w}^*$, the OLS solution

4 Rescaling

Suppose we have a dataset D with 1000 samples and 100 features $\{x_1, x_2, \ldots, x_{100}\}$. Now, we rescale one of these feature by multiplying with 10 (say that feature is x_1).

a) Show that the OLS weights remain unchanged for $i > 1$, and that $w_1^{*\prime} = \frac{1}{10}w_1^*$

We prove a more general claim: When $\mathbf{X}$ is right-multiplied by an invertible matrix $\mathbf{A}$, then the new OLS weights $\boldsymbol{w}^*_A$ are equal to $\mathbf{A}^{-1}\boldsymbol{w}^*$.

First note that this claim would indeed solve the question. Let $\mathbf{A} = \text{Diag}(10, 1, 1, \dots, 1)$. Then $\mathbf{XA}$ is exactly $\mathbf{X}$ with x_1 rescaled, and $\mathbf{A}^{-1} = \text{Diag}(1/10, 1, 1, \dots, 1)$ so $\mathbf{A}^{-1}\boldsymbol{w}^*$ is exactly $\boldsymbol{w}^*$ with w_1 rescaled.

Now to prove the claim:

$$\begin{aligned}\boldsymbol{w}_A^* &= \left((\mathbf{XA})^T(\mathbf{XA})\right)^{-1}(\mathbf{XA})^T\boldsymbol{y}\\ &= \left(\mathbf{A}^T\mathbf{X}^T\mathbf{XA}\right)^{-1}(\mathbf{A}^T\mathbf{X}^T)\boldsymbol{y}\\ &= \mathbf{A}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T\mathbf{X}^T\boldsymbol{y}\\ &= \mathbf{A}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}\\ &= \mathbf{A}^{-1}\boldsymbol{w}^*\end{aligned}$$

As in the lecture, we assume that $\mathbf{X}^T\mathbf{X}$ is invertible.

b) Conclude that the OLS predictions do not change

Let $\boldsymbol{x}$ be the point where we want to make a prediction. Then the scaled vector will be $\mathbf{A}^T\boldsymbol{x}$ and the prediction is

$$(\boldsymbol{w}_A^*)^T(\mathbf{A}^T\boldsymbol{x}) = (\mathbf{A}^{-1}\boldsymbol{w}^*)^T(\mathbf{A}^T\boldsymbol{x}) = (\boldsymbol{w}^*)^T(\mathbf{A}^{-1})^T\mathbf{A}^T\boldsymbol{x} = (\boldsymbol{w}^*)^T\boldsymbol{x}$$

c) What about with Lasso and Ridge regression? Do the weights change? Do the predictions change?

With Lasso and Ridge regression, scaling a feature allows the corresponding weight to be lower so that it does not count as much towards the cost relative to the other weights. This makes the weight more likely to be used so we expect the weights and the predictions to change.

This is why it is important to normalize the features before using Lasso or Ridge regression.