Exercises for Foundations of Data Science

University of Zurich^UZH

Prof. Dan Olteanu,
Dr. Ahmet Kara, Dr. Nils Vortmeier,
Haozhe Zhang

DaST
Data • (Systems+Theory)

Fall 2020/2021 | Sheet 3 | 09.10.2020

- The solutions will be discussed on Friday 23.10.2020, 14:00-15:45 on Zoom.
- Videos with solutions will be posted on OLAT after the exercise session.

Exercise 3.1 [Hiking Might Be a Good Suggestion]
Alice and Max meet Saturdays to pursue their hobbies. On every Friday evening, they discuss what they could do the next day. Assume that the probability that Alice suggests to go hiking is θ_1 and the probability that Max agrees with Alice is θ_2. We model the event that Alice suggests to go hiking by the Boolean random variable H. That is, H takes value 1 if Alice suggests to go hiking, otherwise it takes value 0. Similarly, we model the event that Max agrees with Alice by the Boolean random variable A. Note that A depends on H. Hence, we have the following probability distributions:

H	0	1
$p(H \mid \theta_1)$	$1-\theta_1$	θ_1

H	0	0	1	1
A	0	1	0	1
$p(A \mid H, \theta_2)$	θ_2	$1-\theta_2$	$1-\theta_2$	θ_2

(a) Give the joint probability distribution $p(H, A \mid \theta_1, \theta_2)$ in table form. The table should have three rows, one for H, one for A, and one for $p(H, A \mid \theta_1, \theta_2)$.

(b) The following dataset $\mathbf{D}$ reflects Alice's suggestions and Max's reactions over 10 weeks:

H	1	0	0	1	1	0	1	1	0	0
A	1	1	0	1	1	0	0	0	1	0

What are the maximum likelihood estimators (MLE) $\widehat{\theta_1}$ and $\widehat{\theta_2}$ for θ_1 and θ_2? Justify your answer by giving intermediate steps. What is the numerical value $p(\mathbf{D} \mid \widehat{\theta_1}, \widehat{\theta_2})$?

(c) We now model Alice's suggestions and Max's reactions using the four parameters $\mathbf{\Theta} = \{\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}\}$, where for $h, a \in \{0, 1\}$, θ_{ha} gives the joint probability $p(h, a \mid \mathbf{\Theta})$. What are the MLEs $\widehat{\mathbf{\Theta}} = \{\widehat{\theta}_{00}, \widehat{\theta}_{01}, \widehat{\theta}_{10}, \widehat{\theta}_{11}\}$ for the four parameters in $\mathbf{\Theta}$? What is the numerical value $p(\mathbf{D} \mid \widehat{\mathbf{\Theta}})$?

(d) We denote the 2- and 4-parameter models introduced above by M_2 and M_4, respectively. Assume that we do not know, which of these models is better in predicting future suggestions of Alice and Max. One way to decide this is to use leave-one-out cross-validation. The leave-one-out cross-validated log likelihood of a model M with parameters $\mathbf{\Theta}$ is defined as:

$$L(M) = \sum_{i=1}^{N} \log p(h_i, a_i \mid \widehat{\mathbf{\Theta}}(\mathbf{D}_{-i})),$$

where $\widehat{\mathbf{\Theta}}(\mathbf{D}_{-i})$ denotes the MLEs for the parameters in $\mathbf{\Theta}$ computed on $\mathbf{D}$ excluding the i-th observation. The method leave-one-out cross-validation picks the model with higher $L(M)$. Explain why it holds $L(M_2) \geq L(M_4)$. You do not need to completely compute $L(M_2)$ and $L(M_4)$ to answer this question.

Exercise 3.2 [Practice Makes Perfect]
Alice likes exercising basketball penalty shots in her garden. Max is not so much into basketball but he is fine with watching Alice's basketball sessions. These sessions are very simple. Alice performs penalty shots until she strikes and stops the training session immediately after her first strike. Let θ be the probability that Alice strikes in a penalty shot. Max watched Alice's sessions over N days. He observed that on each day $i \in \{1, \ldots, N\}$, the number of misses before the first strike is m_i.

(a) Using Max's observations, compute the MLE for θ.

Hint: The probability of exactly m_i misses before the first strike can be modeled using the geometric distribution:
$$p(m_i \mid \theta) = \theta \cdot (1-\theta)^{m_i}.$$

(b) Assume that on one day, Alice misses 8 times before her first strike, and on an other day, she misses 6 times before her first strike. Compute the MLE for θ based on this information.

Exercise 3.3 [Lifetime of Radioactive Material]
We observe the decay of some radioactive material and obtain the following probability distribution for the lifetime t of the material:
$$d(t \mid \tau) = \frac{1}{\tau} e^{\frac{-t}{\tau}}$$
where the parameter τ denotes the mean lifetime. Assume that we observe $N \in \mathbb{N}$ lifetimes $t_1, \ldots, t_N$. Give the maximum likelihood function for these observations. Give the MLE for τ.

Exercise 3.4 [The Huber Loss as a Regulariser]
In Exercise Sheet 2 we considered the Huber loss function $h_{\lambda,\mu} : \mathbb{R} \mapsto \mathbb{R}$, which for parameters $\lambda, \mu \in \mathbb{R}$ with $\lambda, \mu > 0$ is defined as:
$$h_{\lambda,\mu}(z) = \begin{cases} \lambda\big(|z| - \frac{\lambda}{4\mu}\big), & \text{if } |z| \geq \frac{\lambda}{2\mu} \\ \mu z^2, & \text{otherwise} \end{cases}$$

Given a vector $\mathbf{z} = [z_1, \ldots, z_D]^\mathsf{T} \in \mathbb{R}^D$, we extend $h_{\lambda,\mu}$ such that $h_{\lambda,\mu}(\mathbf{z}) = \sum_{i=1}^{D} h_{\lambda,\mu}(z_i)$.
In this exercise, we discuss the Huber loss as a regulariser. Let $\ell : \mathbb{R} \mapsto \mathbb{R}$ be an arbitrary loss function, and consider the following regularised loss functions:
$$\mathcal{H}(\mathbf{z}, \mathbf{D}) = h_{\lambda,\mu}(\mathbf{z}) + \frac{1}{N} \sum_{i=1}^{N} \ell(y_i - \mathbf{z}^\mathsf{T}\mathbf{x}_i)$$
$$\mathcal{S}(\mathbf{v}, \mathbf{w}, \mathbf{D}) = \lambda||\mathbf{v}||_1 + \mu||\mathbf{w}||_2^2 + \frac{1}{N} \sum_{i=1}^{N} \ell(y_i - (\mathbf{v} + \mathbf{v})^\mathsf{T}\mathbf{x}_i)$$
Suppose we let $\lambda \mapsto \infty$ in $\mathcal{H}(\mathbf{z}, \mathbf{D})$ and $\mathcal{S}(\mathbf{v}, \mathbf{w}, \mathbf{D})$. Which types of regularised regression do we obtain? What happens when $\mu \mapsto \infty$?

Exercise 3.5 [Centering and Ridge Regression]
Assume that $\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \mathbf{0}$, i.e., the data is centered. Let us denote the parameter for the leading constant term as b (for *bias*). So the linear model is $\widehat{y} = b + \mathbf{x}^\mathsf{T}\mathbf{w}$. Consider minimizing the ridge objective:
$$\mathcal{L}_{\text{ridge}}(\mathbf{w}, b) = (\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^\mathsf{T}\mathbf{w}$$

Here **1** is the vector of all ones and note that b^2 is not regularized. Show that if $\widehat{b}$ and $\widehat{w}$ are the resulting solutions obtained by minimising the above objective, then

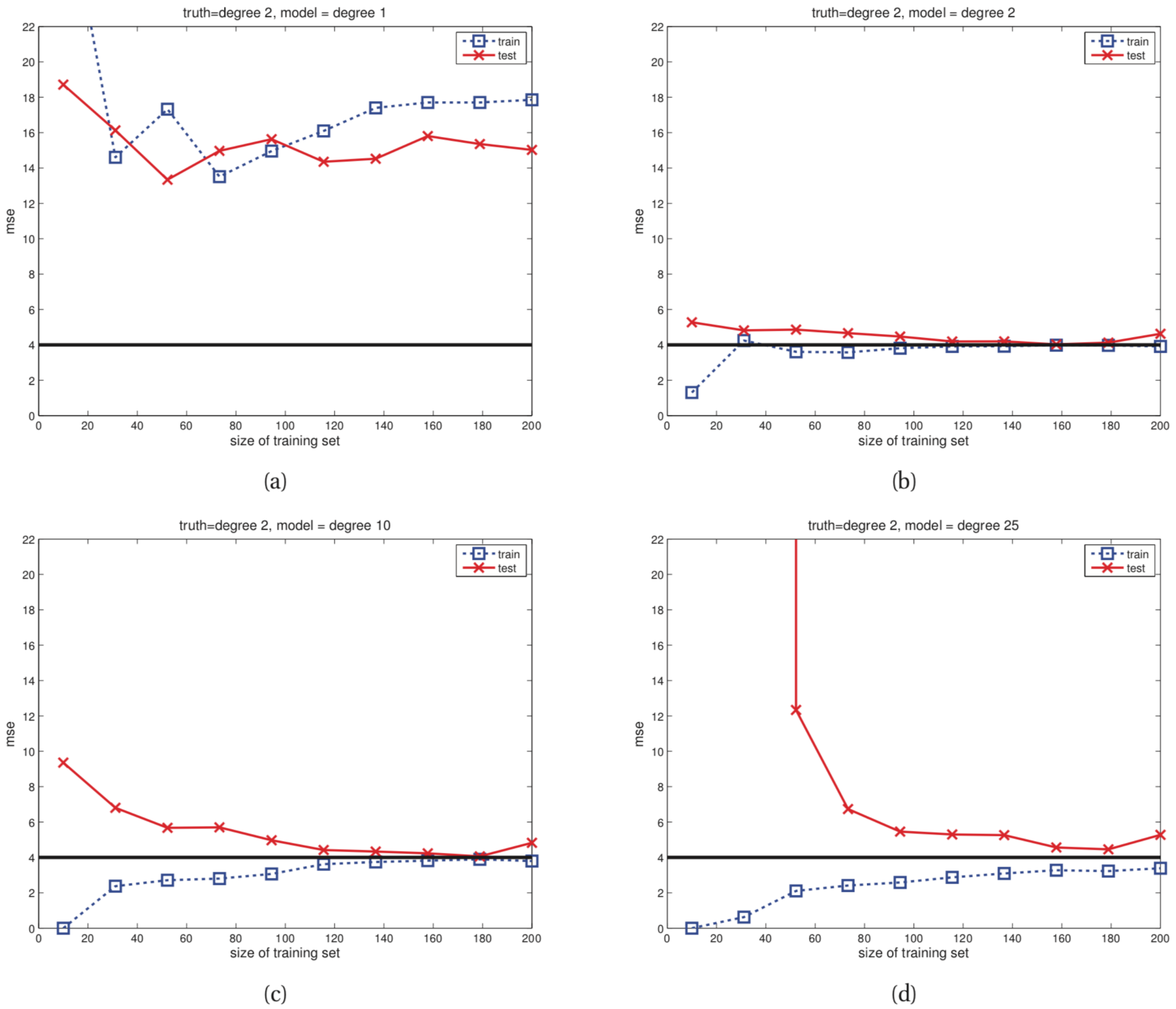
$$\widehat{b} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

What happens if we also center $\mathbf{y}$?

Exercise 3.6 [Training Error and Test Error]

Figure 1 shows the mean squared error (MSE) on some training and test datasets in dependency of the size of the training dataset. Explain the following observations. (1) The test error decreases as we get more training data. (2) For sufficiently complex models, the training error can increase as we get more training data, until we reach some plateau.



(a)

(b)

(c)

(d)

Figure 1: MSE on some training and test datasets in dependency of the size of the training dataset, for data generated from a degree 2 polynomial with Gaussian noise of variance $\sigma^2 = 4$. We fit polynomial models of varying degree to this data: (a) degree 1, (b) degree 2, (c) degree 10, (d) degree 25.