Ensimag – Grenoble INP – UGA
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues
Alexandre Wendling
Year 2024-2025
pedro.rodrigues@inria.fr
alexandre.wendling@univ-grenoble-alpes.fr

TD 2: Some questions from previous exams

▶ Exercise 1 (credits to EPFL CS-433)

Assume we are doing linear regression with mean-squared loss and L2-regularization on four one-dimensional data points. Our prediction model can be written as $f(x) = ax + b$ and the optimization problem can be written as

$$a^\star, b^\star = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{4} \Big(y_i - f(x_i)\Big)^2 + \lambda a^2$$

Assume that our data points (x_i, y_i) are $\{(-2, 1), (-1, 3), (0, 2), (3, 4)\}$. What is the optimal value for the bias, $b^\star$?

(A) Depends on the value of λ.

(B) 3

(C) 2.5

(D) None of the above answers.

▶ Exercise 2 (credits to Berkeley CS-189)

In the following statements, the word “bias” is referring to the bias-variance decomposition. Which one of them is true?

(A) A model trained with N training points is likely to have lower variance than a model trained with $2N$ training points.

(B) If my model is underfitting, it is more likely to have high bias than high variance.

(C) Increasing the number of parameters (weights) in a model usually improves the test set accuracy.

(D) None of the above.

▶ Exercise 3 (credits to EPFL CS-433)

Consider a regression model where data (x, y) is generated by input $x \in \mathbf{R}$ uniformly sampled between $[0, 1]$ and $y = x + \varepsilon$, where ε is random noise with mean 0 and variance 1. Two models are carried out for regression: model $\mathcal{A}$ is a trained quadratic function $g_\mathcal{A}(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$ and model $\mathcal{B}$ is a constant function $g_\mathcal{B}(x) = \frac{1}{2}$. Compared to model $\mathcal{B}$, model $\mathcal{A}$ has

(A) Higher bias, higher variance.

(B) Lower bias, higher variance.

(C) Higher bias, lower variance.

(D) Lower bias, lower variance.

▶ Exercise 4

Consider the following `python` script:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
np.random.seed(0)
# number of variables
pt = 201
# number of predictors
p = pt - 1
# sample size
n = 30 * p
# generate data
D = np.random.randn(n, pt)
df = pd.DataFrame(data=D)
df = df.rename(columns={0:'Y'})
# do multiple linear regression
df['intercept'] = 1
model = sm.OLS(df['Y'], df.drop(columns='Y'))
results = model.fit()
print(results.summary())
```

(a) What does the script do? Run it on your computer.

(b) What is the true distribution of the random variable Y given the first 200 columns of matrix D, which we shall call $X_1, \ldots, X_{200}$?

(c) Write an equation defining the model estimated by 'model.fit()'. What is the difference between this model and the one defined above?

(d) Print 'results.params'. What is going on?

▶ Exercise 5

In this exercise, you will perform multiple linear regression on simulated data under different conditions. To ensure reproducibility on your results, set the seed with `numpy.random.seed(0)` at the beginning of your script.

(a) Simulate a dataset of size $N = 1000$ of the following generating model:

$$\begin{aligned} X_{1,i} &= \varepsilon_{1,i} \\ X_{2,i} &= 3X_{1,i} + \varepsilon_{2,i} \\ Y_i &= X_{2,i} + X_{1,i} + 2 + \varepsilon_{3,i} \end{aligned}$$

where $i \in \{1, \ldots, N\}$ and the ε_{ij} are independent $\mathcal{N}(0,1)$ random variables. For a given i, what is the distribution of $(X_{1,i}, X_{2,i})$? Plot the clouds of points of the simulated values of $(X_{1,i}, X_{2,i})_{i=1,\ldots,n}$. What is its shape? Can you write an analytical formula for it?

(b) Let us consider the following two regression models:

$$\begin{aligned} \text{Model A:} \quad Y_i &= \alpha_1 X_{1,i} + \alpha_0 + \tilde{\varepsilon}_{A,i} \\ \text{Model B:} \quad Y_i &= \beta_2 X_{2,i} + \beta_0 + \tilde{\varepsilon}_{B,i} \end{aligned}$$

where $\tilde{\varepsilon}_{A,i} \sim \mathcal{N}(0, \sigma_A^2)$ and $\tilde{\varepsilon}_{B,i} \sim \mathcal{N}(0, \sigma_B^2)$. What should be the values of $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\sigma}_A^2, \hat{\beta}_0, \hat{\beta}_2, \hat{\sigma}_B^2$ when $N \to \infty$? Consider $N = 1000$ and check whether the estimates of the parameters are close to the true values that you've calculated. Now do `np.random.seed(3)` and simulate again a dataset $X_{1,i}, X_{2,i}, Y_i$ for $n = 10$. Estimate the parameters. What happens?

(c) Let us now consider the full model

$$Y_i = \gamma_2 X_{2,i} + \gamma_1 X_{1,i} + \gamma_0 + \varepsilon_i$$

where $i \in \{1, \ldots, n\}$ and the ε_i are independent $\mathcal{N}(0, \sigma^2)$ random variables. For the previously simulated data with $n = 10$, estimate $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\sigma}^2$ and compare them with the parameters obtained in item (b). What can you say about the effects of X_1 and X_2 on Y? And about their correlation?