Biomedical Data Science & AI

Exercise sheet 4

SoSe2025

Prof. Dr. Holger Fröhlich, Mohamed Aborageh, Jannis Guski, Sebastian Schwick

Due date: May 6th

Questions

Exercise 1 – Random Variables and MLE (total: 7 points)

1. Consider a Bernoulli distributed random variable X (with $P[X = 1] = p$ and $P[X = 0] = 1 - p$) and a normal distributed random variable Y with mean μ and variance σ^2. Assuming that X and Y are independent, compute $E[XY]$ and $Var[XY]$. **(1 point)**

2. The probability density function of the continuous uniform distribution $Unif(0,\ \theta)$ with $\theta > 0$ is given by $f(x) = 1/\theta$ for $0 \leq x \leq \theta$ and $f(x) = 0$ otherwise.
   a. Write down and simplify the likelihood function $L(x_1,\ \ldots,\ x_n; \theta)$ for n independent samples of $Unif(0,\ \theta)$. **(1 point)**
   b. What is the maximum likelihood estimator for this setting? **(1 point)**
   c. What changes if you consider $f(x) = 1/\theta$ for $0 \leq x < \theta$ and $f(x) = 0$ otherwise? **(1 point)**

3. The probability density function of the Laplace distribution $Laplace(\theta,\ 1)$ with $\theta \in \mathbb{R}$ is given by $f(x) = 1/2\ e^{-|x-\theta|}$.
   a. Write down and simplify the log-likelihood function $l(x_1,\ \ldots,\ x_n;\ \theta)$ for n independent samples of $Laplace(\theta,\ 1)$. **(1 point)**
   b. What is the maximum likelihood estimator for this setting? Is it unique? **(1 point)**

Exercise 2 - (total: 9 points)

Multiple Linear Regression and violated model assumptions

Load the datasets toy_1.csv and toy_2.csv using pandas. These are small simulated datasets: toy_1 contains 10 features and 10 observations, while toy_2 contains 10 features and 100 observations. Both datasets also include their y values.

1. In principle, there are two possible failure scenarios when model assumptions are violated, (1) “silent failure”: the model makes estimates, but due to the violations they are misleading; (2) “error” the model is not able to work and cannot provide estimates. List the 3 important model assumptions stated in the lecture and annotate to which failure scenario either belongs to. Justify. **(1 point)**
2. Perform an EDA (explanatory data analysis) to identify and correct violations.
    a. Compute a Pearsons’ correlation matrix np.corrcoef for both datasets and read carefully the description of parameters “x” and “rowvar”. Visualize the correlations with a seaborn heatmap (set the argument “annot=True” to display correlation coefficients). Compare the correlation structure of both datasets. **(1 point)**
    b. From (a) and the shape of both datasets (number of observations, number of features) detect and name two violated assumptions that could be found in toy_1 but not toy_2 **(1 point)**
3. Fit a linear regression model using statsmodels linear regression on both datasets
    a. Standardize the data. Hint: Use standard scaler **(1 point)**
    b. Report and interpret the model summary. **(1 point)**
    c. How well does the model fit both datasets? Was it reasonable to use linear regression? Explain. **(1 point)**
    d. Correct for violated assumptions in toy_1 and refit the regression model. Justify. Note: There is no unique solution, different valid solutions are possible. **(1 point)**
    e. From (d), are there any significant (p-value <= 5%) feature-related coefficients? Which? Would you expect a change in significance in dependence of the solution in 3c? Explain. **(1 point)**
    f. toy_1 was simulated such that 3 of the 10 features are informative and the remainder 7 are just noise. Which features are most likely informative ones? **(1 point)**

Exercise 3 – Average Treatment Effect (total: 9 points)

In April 2025, the anti body Lecanemab became the first drug for the treatment of Alzheimer’s Disease (AD) to be approved in the EU. Lecanemab targets Amyloid plaques in the brain, which are known to cause cognitive deterioration as AD progresses.
However, there are two limitations that restrict who can benefit from the drug:

- The drug is only effective in early disease stages because it cannot revert the brain damage that Amyloid plaques cause over time. It makes no sense to prescribe it at later AD stages.

- The risk of brain bleeding as a side effect is particularly high for patients who have one or two ApoE4 alleles. Thus, the drug should only be prescribed to patients with no or at most one ApoE4 allele. Note that ApoE4 is also an important risk factor of conversion to AD in the first place.

Load the (purely fictional!) dataset "lecanemab.csv" that contains Clinical Dementia Rating (CDR) scale results for patients with different disease stages at a baseline and after 18 months. Your task is to estimate the Average Treatment Effect (ATE) that the administration of Lecanemab at baseline has on CDR after 18 months. Since the dataset was synthetically generated, we know that the true effect is –0.5.

1. Compute a (naive) ATE estimate by subtracting the mean in the untreated group from the mean in the treated group. Is your result different from the true effect? If yes, explain why that is the case. **(2 points)**
2. Inform yourself about propensity scores and Inverse Probability of Treatment Weighting (IPTW).
    a. Which features would you use as predictors in a propensity model and why? **(1 point)**
    b. Fit a logistic regression as a propensity model and derive IPTW. **(1 point)**
    c. Re-estimate the ATE with the use of IPTW. How does your result compare to the true effect and the naive ATE estimate? **(1 point)**
3. Inform yourself about the S-Learner.
    a. Fit a multiple linear regression using the same predictors chosen in 2a. **(1 point)**
    b. Re-estimate the ATE with the use of an S-Learner. How does your result compare to the true effect and the naive ATE estimate? **(1 point)**
4. Inform yourself about Augmented Inverse Probability of Treatment Weighting (AIPTW) on this page: https://matteocourthoud.github.io/post/aipw/
    a. Re-estimate ATE by plugging in your results from 3.3 and 3.4 into the AIPTW equation. **(1 point)**
    b. What happens if you replace your estimated propensity scores with the constant 0.5? What is the advantage of AIPTW and how does this show in your results? **(1 point)**