Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Exercise Sheet #1: *Linear Regression*

Deadline: Thursday, November 16, 2023, 15:00

Before solving the exercises, read the instructions on the course website.

- For each theoretical problem, submit a single `pdf` file that contains your answer to the respective problem. This file may be a scan of your (legible) handwriting.
- For each practical problem, submit a single `zip` file that contains
  - the completed jupyter notebook (`.ipynb`) file,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows all your results.
- For the bonus question, submit a single `zip file` that contains
  - a `pdf` file that includes your answers to the theoretical part,
  - the completed jupyter notebook (`.ipynb`) file for the practical component,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows your results.
- **Every team member** has to submit a signed Code of Conduct.
- **IMPORTANT** You must make the team on CMS *before* you upload the solutions. If you upload the solutions first and create the team after it, the solution will not show for the new team member!

Problem 1 (T, 2 Points). **Warmup.**
Suppose we want to apply an appropriate statistical learning model to a given problem. Briefly explain how the following parameters influence our choice,

- labelled vs. unlabelled input data,
- numerical vs. categorical variables,
- interpretability vs. prediction task,
- fixed vs. flexible number of model parameters.

Problem 2 (T, 8 Points). **Error Measures.**
In the regression setting, we most commonly use RSS as an error measure. Consider instead the following loss function L,

$$L(\beta, r) = \frac{1}{n} \sum_{i}^{N} r_i \big(y_i - \beta_0 - x_i \beta\big)^2 \tag{2.1}$$

for a target y and single predictor $X \in \mathbb{R}^n$.

1. $[1pts]$ What is the effect of the parameters r_i?
2. $[3pts]$ Derive the minimizer $\hat{\beta}$ of Eq.(2.1) when we keep r fixed.
3. $[3pts]$ Consider the following choices for r. Explain the effect of each choice, as well as what purpose it could serve.
   - We set each r_i to an integer value $r_i \in \mathbb{Z}$ with $r_i > 1$.
   - Assuming that we know the noise variance σ_i of each data sample x_i, we set $r_i = \frac{1}{\sigma_i^2}$.

- Assume that we have two different datasets X_1 and X_2. From X_1, we estimate the probability density function over X as p, and from X_2 we estimate the density as q. We set $r_i = \frac{p(x_i)}{q(x_i)}$.

4. $[1pts]$ List three main assumptions that we rely upon in ordinary linear regression. Is there an assumption that we can address by using L instead of RSS?

Problem 3 (T, 4 Points). **Geometry.**
This exercise will take us through a geometric interpretation of linear regression using the following small example,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}, y = \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix}.$$

1. $[1pts]$ State the linear regression solution $\hat{\beta}$ for this system.
2. $[1pts]$ Now consider the space $S \in \mathbb{R}^3$ spanned by the columns $X^{(i)}$ of $\mathbf{X}$,

$$S = \text{span}(X^{(1)}, X^{(2)}) = \{a_1 X^{(1)} + a_2 X^{(2)} \mid a_i \in \mathbb{R}\}.$$

   Show that the matrix

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

   maps the vector y onto this space.
3. $[1pts]$ Show that the vector $y - Py$ is perpendicular to the space S, $y - Py \perp S$. *Hint: This is equivalent to showing $\langle y - Py, \mathbf{X}a\rangle = 0$ for all $a \in \mathbb{R}^2$, where $\langle \cdot, \cdot \rangle$ denotes the dot product.*
4. $[1pts]$ Based on the previous part (3.3), how are P and $\hat{\beta}$ related?

Problem 4 (T, 6 Points). **Bias and Variance.**
Consider the bias and variance of a linear model f.

1. $[1pts]$ Explain in concise terms the meaning of bias and variance in the context of linear regression. What is the relationship between them?
2. $[2pts]$ Consider the following equation,

$$\text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 = \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right].$$

   Explain the meaning of each term and show that the above holds.
3. $[2pts]$ How is $\mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right]$ related to the expected test MSE, $\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right]$? Consider the difference of these quantities and explain its meaning.
4. $[1pts]$ State whether the following statement is true or false and explain why.

   "The Gauss-Markov theorem states that the least-squares estimates $\hat{\beta}$ have the smallest variance among all linear estimates. Since the least-squares estimates $\hat{\beta}$ are unbiased, this means that biased estimators will always have a larger variance than $\hat{\beta}$."

Problem 5 (P, 15 Points). **Penguins.**
In this exercise, we will explore the *palmerpenguins* dataset. Consult the provided Jupyter notebook `Practical_Problem_1.ipynb` for this problem and add your answers and code. **Please rename the file to include the matriculation numbers of all team members (e.g. `7010000_2567890_A1.ipynb`).**

1. [1*pts*] Load the data. Impute the missing values of all numerical features by replacing them with the mean value for the respective feature.
2. [3*pts*] Select only the samples belonging to the species Gentoo. Consider the variables `flipper_length_mm`, `body_mass_g`, `bill_length_mm`, `bill_depth_mm` and find the corellations between each pair. Which appear to be most highly corellated?
3. [3*pts*] Fit a linear model predicting `body_mass_g` from `bill_depth_mm` for the species Bentoo and show the linear parameters. Judge the goodness of fit using an appropriate measure. *Useful function: `sklearn.LinearRegression`.*
4. [4*pts*] Now consider the pair of variables `body_mass_g` and `bill_depth_mm` over *all* penguin species. Perform a hypothesis test on whether there is a statistically significant relationship between the predictors. What problem do you see? *Hint: Consider visualizing the relationship between the variables using a scatterplot. Useful function: `seaborn.scatterplot`.*
5. [2*pts*] Consider again the species Gentoo. Suppose we observe a new penguin with bill length of 17. Using the body mass of its four closest neighbors (in terms of the bill lengths), predict the body mass of the new penguin. *Useful function: `sklearn.neighbors.KNeighborsRegressor`.*
6. [2*pts*] For Gentoo, plot the RSS of a kNN regression predicting `body_mass_g` from `bill_depth_mm` for different choices of k ($k \in \{1, ..., 10\}$). Which k would you choose here and why?

Problem 6 (Bonus). **Projections.**
In this exercise, we consider a high-dimensional $X \in \mathbb{R}^{n \times p}$ with $p \gg n$. Before doing our linear regression analysis, we want to summarize this data in a smaller number of d dimensions. We start with $d = 1$.

1. Summarizing X in $d = 1$ directions can be thought of as projecting X to a line parameterized by some unit vector u such that we obtain scalars $(x_0 \cdot u)u$ for each sample x_0. What is the residual error of this projection?
2. To choose u such that it retains as much information about X as possible, we optimally want different points x_i, x_j to map to different projections, in other words, ensure a high variance of projected points. Find the unit vector $w^T w = 1$ that maximizes the variance σ_u^2 of the projected points $(x_i \cdot u)u$. *Hint: Use Lagrangean multipliers to include the unit constraint.*
3. Interpret how your result u relates to the covariance matrix $\mathrm{cov}(X) \in \mathbb{R}^{p \times p}$.

The above can be generalized to the case $d > 1$ by projecting to multiple orthogonal vectors $\{u_1, ..., u_d\}$.

1. Devise an iterative way of obtaining $\{u_2, ..., u_d\}$ starting from u_1 obtained from the previous part. Using the resulting u_i, describe (on a high level) how to implement a lower-dimensional regression.
2. Another way to obtain $\{u_1, ..., u_d\}$ is via the following factorization of X,
$$X = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T$$
where $\mathbf{\Sigma} \in \mathbb{R}^{n \times p}$ is a diagonal matrix containing the so-called singular values of X, and where the columns of $\mathbf{U} \in \mathbb{R}^{n \times n}$, and $\mathbf{V} \in \mathbb{R}^{p \times p}$ are orthogonal unit vectors. Write $X^T X$ in terms of the above decomposition. What are its eigenvalues and how are they related to $\mathbf{\Sigma}$?
3. Consider the linear regression using the projected features, in comparison to ordinary least squares using all features. Mention the advantages that you see. Do you also see a potential problem?