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.

Solution.

1. [0.5*pts*] Given labelled data, i.e. pairs $(\boldsymbol{x}_i, y_i)$ consisting of a feature vector $\boldsymbol{x}_i$ and its associated label y_i, **supervised** learning describes the setting where we learn a mapping from features to outcomes. In contrast, **unsupervised** learning describes the setting where we are only given a set of unlabelled feature vectors, with the goal of learning useful structure or patterns in the data.
2. [0.5*pts*] Variables can be quantitative (i.e. numerical values - *continuous, e.g. income, or discrete, e.g., number of children*) or qualitative (i.e. values in one of K different categories - *nominal, e.g., sex, or ordinal, e.g., health diagnosis*). If the outcome of a supervised learning task is quantitative, the learning task is called **regression**, if the outcome is qualitative, it is called **classification**.
3. [0.5*pts*] In supervised learning there we can define two distinct tasks (goals) of learning models from data. We can either aim to **predict** outcomes for unseen inputs (new data points) or aim to **infer** the relationship between certain features and the response (e.g. understanding the way that Y is affected as features X change, i.e. the data generation process). *The learning task affects the model choice: inference requires more interpretable models than prediction.*

4. [0.5*pts*] **Parametric** methods make the strong assumption that the underlying model of the method lies in a well-defined class which is parameterised by a set of parameters whose size does not depend on the number of training points; here, picking the correct model amounts to simply specifying the correct set of parameters. As an example, linear regression makes a strong assumption that the outcome is a linear function of a (given number of) features, after which we only need to specify the linear coefficients of each feature.

   In contrast, the underlying models of **Non-parametric** methods have a flexible number of parameters which depends on the complexity of the data; note that this does not refer to the dimensions of each observation, but rather to the arrangement of the observations in space. An extreme case of a non-parametric method is the k-NN classifier, which essentially stores the entire training set.

   Importantly, parametric methods tends to outperform non-parametric ones given the same observations, whenever the assumptions of the former are accurate. Otherwise, and given enough data, the flexibility of the non-parameteric methods allows them to outperform parametric ones whose assumptions are not entirely correct.

   The relative performance of these two classes of methods can also be seen as yet another instance of the bias-variance tradeoff: parametric methods typically contain a smaller class of models than those of non-parametric ones, and therefore their variance is typically lower. However, if the true model lies far from the best parametric model—as is the case when its assumptions are wrong—the bias of the corresponding parametric method will eventually dominate the variance of a non-parameteric one, when given enough data.

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. [1*pts*] What is the effect of the parameters r_i?
2. [3*pts*] Derive the minimizer $\hat{\beta}$ of Eq.(2.1) when we keep r fixed.
3. [3*pts*] 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. [1*pts*] 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?

Solution.

1. The r_i can be seen as sample-dependent weights. *The higher r_i, the larger the influence of sample x_i on the parameter $\hat{\beta}$ minimizing L. This can allow us to discount samples that we know are noisy or overrepresented in our dataset, for example.*

2. Using some linear algebra, we get $\beta_0 = \bar{y}_w - \hat{\beta}\bar{x}_w$ and

$$\hat{\beta} = \frac{\sum_n^N w_n(x_n - \bar{x}_w)(y_n - \bar{y}_w)}{\sum_n^N w_n(x_n - \bar{x}_w)^2} ,$$

where $\bar{y}_w, \bar{x}_w$ are the weighted sample means,

$$\bar{y}_w = \frac{1}{\sum_n^N w_n} \sum_n^N w_n y_n, \bar{x}_w = \frac{1}{\sum_n^N w_n} \sum_n^N w_n x_n .$$

3. These choices have the following effects,

   - Setting $r_i = z$ to an integer z has the effect of replicating a sample x_i. That is, using L as a loss function with integer weights r will have the same effect as using a simple linear regression with RSS on a dataset where we duplicated each sample x_i in the dataset $r_i = z$ times, which can be used to increase the importance of some samples relative to others.
   - If we know the noise variance σ_i^2 at each datapoint, we assign noisier samples smaller weights $r_i = \frac{1}{\sigma_i^2}$, respectively less noisy samples higher weights. Hence, this model can address heteroskedastic cases where the noise variance is data-dependent.
   - This reweighting scheme can be used if we have two datasets that differ in distribution p, q. A model fitted on X_1 can be made to more closely match the distribution of X_2 in this way. *For example, consider the case where in the first dataset X_1, samples in some interval $[i_1, i_2]$ are overrepresented, while in the second dataset X_2, much fewer samples fall into $[i_1, i_2]$ (known as selection bias). By assigning the samples in $[i_1, i_2]$ a smaller weight, we can make a model trained on X_1 match the distribution of X_2 more closely.*

4. We assume linearity, uncorellated error terms, noncollinearity, homoskedasticity, and i.i.d.-ness of the data. As we have seen in the previous part, using the samle weights allows us to account for sample-dependent noise, thus addressing heteroskedastic cases.

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. [1*pts*] State the linear regression solution $\hat{\beta}$ for this system.
2. [1*pts*] 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. [1*pts*] 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. [1*pts*] Based on the previous part (3.3), how are P and $\hat{\beta}$ related?

Solution.

1. We have
$$X^T X = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, (X^T X)^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix},$$
so that
$$\widehat{\beta} = (X^\top X)^{-1} X^\top y = \frac{1}{3}\begin{bmatrix} 1 & -1 & -2 \\ 1 & 2 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} -4 \\ 5 \end{bmatrix}.$$

2. We can show this for any y by rewriting
$$Py = X(X^T X)^{-1} X^T y$$
as $Py = Xa$, with $a := (X^T X)^{-1} X^T y$, and Xa is a linear combination of the columns of X.

3. We have
$$(Xa)^T (y - X(X^T X)^{-1} X^T y) = a^T (X^T y - X^T y) = 0.$$

4. We have the following relationship,
$$X\hat{\beta} = X(X^T X)^{-1} X^T y = Py.$$
We conclude that minimizing the least squared errors corresponds to an orthogonal projection of y onto the space spanned by the columns of X.

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

1. [1*pts*] Explain in concise terms the meaning of bias and variance in the context of linear regression. What is the relationship between them?

2. [2*pts*] Consider the following equation,
$$\mathrm{Var}(\hat{f}(x_0)) + \Big[\mathrm{Bias}(\hat{f}(x_0))\Big]^2 = \mathbb{E}\left[(f(x_0) - \hat{f}(x_0))^2\right].$$
Explain the meaning of each term and show that the above holds.

3. [2*pts*] 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. [1*pts*] 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 $\hat{\beta}$."

Solution.

1. The **bias** of a model refers to the systematic error that we make due to our choice of the learning model and number of parameters, i.e. whether these are misspecified for a given dataset or represent it well. The **variance** refers to the error due to noise, i.e. whether our model is sensitive to changes in the data due to random noise. There is a trade-off between bias and variance because we aim to choose an appropriate model that has high *accuracy* the training data (lower bias) while also *generalizing* to unseen data (lower variance). In general, models with higher bias may be prone to *underfitting*, models with higher variance to *overfitting*.

2. **Definitions**: $\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right]$ refers to the **expected test mean squared error (MSE)**, i.e., the average test MSE that we would obtain, by first using a large number of different training sets and for each training set learning a model $\hat{f}(x)$ estimating the true relationship $y = f(x) + \epsilon$ with irreducible error ϵ. $\mathbb{E}\left[(y_0 - \hat{f}(x_0))^2\right]$ is then the average prediction error over all of these models for a fix data point (x_0, y_0). We **assume** $\mathbb{E}(\epsilon) = 0$, i.e. the irreducible error to have a zero mean. The overall expected test MSE can be computed by averaging over all models and all possible values of x_0 (i.e., all data points) in the test set. We also have $\mathrm{Var}(\hat{f}) = \mathbb{E}(\hat{f}^2) - \left[\mathbb{E}(\hat{f})\right]^2$ and $\mathrm{Bias}(\hat{f}) = \mathbb{E}(\hat{f} - f)$. (1 Point)
We show now that

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

where for the sake of brevity, use $f = f(x_0)$, $\hat{f} = \hat{f}(x_0)$, thus $y_0 = f + \epsilon$.

$$\begin{aligned}
\mathrm{Var}(\hat{f}) + \left[\mathrm{Bias}(\hat{f})\right]^2 &= \mathbb{E}(\hat{f}^2) - \left[\mathbb{E}(\hat{f})\right]^2 + \left[\mathbb{E}(\hat{f} - f)\right]^2 \\
&= \mathbb{E}(\hat{f}^2) - \left[\mathbb{E}(\hat{f})\right]^2 + \left[\mathbb{E}(\hat{f}) - \mathbb{E}(f)\right]^2 \\
&= \mathbb{E}(\hat{f}^2) - \left[\mathbb{E}(\hat{f})\right]^2 + \left[\mathbb{E}(\hat{f})\right]^2 - 2\mathbb{E}(\hat{f})\mathbb{E}(f) + [\mathbb{E}(f)]^2 && \text{binomial formula} \\
&= \mathbb{E}(\hat{f}^2) - 2f\mathbb{E}(\hat{f}) + f^2 && (\mathbb{E}(f) = f, \text{ since f is deterministic}) \\
&= \mathbb{E}\left[(f - \hat{f})^2\right] && \text{(1 Point)}
\end{aligned}$$

3. The difference of this quantity is the variance of the error terms. (0.5 point) That is,

$$\mathbb{E}\left[(y_0 - \hat{f})^2\right] = \mathbb{E}\left[(f - \hat{f})^2\right] + \mathrm{Var}(\epsilon) \qquad \text{(0.5 Point)}$$

To see this, we first simplify $\mathbb{E}\left[(y_0 - \hat{f})^2\right]$:

$$\begin{aligned}
\mathbb{E}\left[(y_0 - \hat{f})^2\right] &= \mathbb{E}\left[(f + \epsilon - \hat{f})^2\right] \\
&= \mathbb{E}\left[\left((f - \hat{f}) + \epsilon\right)^2\right] \\
&= \mathbb{E}\left[(f - \hat{f})^2 + 2(f - \hat{f})\epsilon + \epsilon^2\right] \\
&= \mathbb{E}\left[(f - \hat{f})^2\right] + 2\mathbb{E}\left[(f - \hat{f})\epsilon\right] + \mathbb{E}\left[\epsilon^2\right] && \text{(0.5 Point)}
\end{aligned}$$

The last term is $\mathrm{Var}(\epsilon)$ since $\mathrm{Var}(\epsilon) = \mathbb{E}\left[(\epsilon - \mathbb{E}[\epsilon])^2\right] = \mathbb{E}\left[\epsilon^2\right] + [\mathbb{E}(\epsilon)]^2$ and we assume $\mathbb{E}(\epsilon) = 0$.
The second term equals to zero because ϵ is independent of $(f - \hat{f})$, so
$2\mathbb{E}\left[(f - \hat{f})\epsilon\right] = 2\mathbb{E}\left[(f - \hat{f})\right]\mathbb{E}[\epsilon] = 0$ with again the assumption $\mathbb{E}(\epsilon) = 0$. (0.5 point)
Note that combining the results of 1. and 2. gives:

$$\mathbb{E}\left[(y_0 - \hat{f})^2\right] = \mathrm{Var}(\hat{f}) + \left[\mathrm{Bias}(\hat{f})\right]^2 + \mathrm{Var}(\epsilon)$$

The expected test mean square error (MSE), for a given value x_0, can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ.

4. This statement is false since the Gauss-Markov theorem states that the least squares estimates of the parameters β have the smallest variance among all linear **unbiased** estimates under certain

conditions. Thus, there may well exist a biased estimator with smaller mean squared error. Such an estimator would trade a little bias for a larger reduction in variance. Biased estimates are commonly used.

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?

Solution.

1. See notebook.

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 $\text{cov}(X) \in \mathbb{R}^{p \times p}$.

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?

Solution. To be discussed in tutorials on 27th and 28th November.