Elements of Machine Learning, WS 2022/2023
Jilles Vreeken and Aleksandar Bojchevski
Exercise Sheet #3: *Generalization*

Deadline: Thursday, December 15, 2022, 16:00

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.

Problem 1 (T, 14 Points). **Cross-Validation.**

1. [$4pts$] Explain the impact of the value for k in k-fold cross validation. Where does k-fold CV fit in between the validation set approach and LOOCV and what is the advantage of using it?
2. [$2pts$] For the hat matrix H, defined as

$$H = X(X^T X)^{-1} X^T ,$$

the diagonal element $h_i = H_{(ii)}$ is called the leverage. What is the meaning of the leverage h_i for sample i? Consequently, what effect on model estimation does removing a sample with high leverage from the dataset have?

3. [$8pts$] Prove that for linear and polynomial least squares regression, the LOOCV estimate for the test MSE can be calculated as

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2 .$$

Hint: First, properly understand and define all variables that are relevant for the situation where you set sample i aside.

Solution.

1.
   (i) **Impact of k**: For $k = 2$, k-fold cross validation is essentialy the validation set approach (depending on how the data is split). For $k = n$, it is the same as LOOCV. k-fold CV can hence be seen as a “compromise“ between the validation set approach and LOOVC. (2 Points)

   (ii) **Advantages of k-fold**: Compared to the validation set approach, it is more stable. (1 Point) Compared to LOOCV, it has a computational advantage (1 Point).
   As we will show in part (3), for linear and polynomial least squares regression, the LOOCV estimate can be calculated by fitting a single model.

2\.

The leverage h_{ii} shows how the value of x_i compares to that of other samples, i.e., how unusual its value is w.r.t. X. If we remove a sample x_i, we expect a larger influence on model estimation and prediction if the leverage is large. (2 points).

Note that from how it is defined, the leverage can be seen as the weighted distance between point x_i and the sample mean. Also, the leverage is not to be confused with outliers, which have an unusual value w.r.t. the target y rather than the predictors X.

3\.

First, we define some terms.
Given: complete data set $\{X, y\}$
data set with ith sample left out $\{X^{-i}, y^{-i}\}$
H^{-i} hat matrix for X^{-i}
$\hat{y} = Hy$ and $\hat{y}^{-i} = H^{-i}y^{-i}$

We now add $\{x_i, \hat{y}^{-i}\}$ to $\{X^{-i}, y^{-i}\}$, with $\hat{y}_i^{-i} = \hat{f}^{-i}(x_i) = x_i^T \hat{\beta}^{-i}$, where $\hat{\beta}^{-i}$ is the estimated parameter from leave-one-out regression. We now denote the augmented dataset $\{X, \tilde{y}\}$, where X is the complete data set and $\tilde{y}$ is the true response y except the ith location, for which we have $\tilde{y}_i = \hat{f}^{-i}(x_i)$. (Clear definitions: 1 Point)
$\{X, \tilde{y}\}$ includes $\hat{f}^{-i}(x_i)$ as target. Hence adds zero additional loss for $\hat{f}^{-i}$. As all remaining datapoints are the same as in $\{X^{-1}, y^{-1}\}$ the same $\hat{\beta}$ minimizes the overall loss. Hence:

$$\hat{\beta}^{-i} = (X^{-i^T} X^{-i})^{-1} X^{-i^T} y^{-1} = (X^T X)^{-1} X^T \tilde{y}$$

$$\begin{aligned}
\hat{y}_i^{-i} = \hat{f}^{-i}(x_i) = x_i^T \hat{\beta}^{-i} &= x_i^T (X^T X)^{-1} X^T \tilde{y} \\
&= h_i^T \tilde{y} = \sum_{j \neq i} H_{ij} y_j + H_{ii} \hat{f}^{-i}(x_i) && \text{1 Point} \\
&= \sum_j H_{ij} y_j - H_{ii} y_i + H_{ii} \hat{f}^{-i}(x_i) && \text{1 Point} \\
&= \hat{y}_i - H_{ii} y_i + H_{ii}^{-i} \hat{f}^{-i}(x_i) && \text{1 Point} \\
&\iff \hat{y}_i^{-i} - H_{ii} \hat{y}_i^{-i} = \hat{y}_i - H_{ii} y_i && \text{1 Point} \\
&\iff \hat{y}_i^{-i} = \frac{\hat{y}_i - H_{ii} y_i}{1 - H_{ii}} && \text{1 Point} \\
&\iff y_i - \hat{y}_i^{-i} = y_i - \frac{\hat{y}_i - H_{ii} y_i}{1 - H_{ii}} && \text{0.5 Points} \\
&\iff y_i - \hat{y}_i^{-i} = \frac{y_i - \hat{y}_i}{1 - H_{ii}} && \text{0.5 Points}
\end{aligned}$$

Thus, we showed that the ith residual in LOOCV can be calculated given the above formula and we can thus calculate the LOOCV estimate to obtain the test error. (1 Point).

Problem 2 (T, 6 Points). **The Bootstrap.**
We will now derive the probability that a given observation is part of a bootstrap sample of size n. Suppose that we obtain a bootstrap sample from a set of n observations.

1. [2*pts*] What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

2. [2*pts*] Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$.

3. [2*pts*] Comment on the behavior of the above probabilities with increasing sample size n.

Solution.

1. We know that the bootstrap takes random samples *uniformly* from the observed dataset. Therefore, the probability, of the first bootstrap observation being exactly the jth out of n original observations is precisely $1/n$. The complement is therefore $1/n$.

2. The jth observation x_j is not in a bootstrap sample of size n, if none of its elements $b_1, \cdots, b_n$ are are equal to the jth observation. Since all bootstrap observations are i.i.d. distributed, we can use the first part of this exercise to obtain

$$P(x_j \notin \text{ bootstrap sample}) = \prod_{i=1}^{n} P(x_j \neq b_i) = (1 - 1/n)^n \,.$$

3. It is well known that $(1 + x/n)^n \overset{n\to\infty}{\longrightarrow} e^x$, so that by plugging in $x = -1$ we obtain $(1 - 1/n)^n \overset{n\to\infty}{\longrightarrow} e^{-1} \approx 0.368$. That is, on average approximately a third of all observations are not contained in any given bootstrap sample.

Problem 3 (T, 10 Points). **Correlated Variables.**
It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the Lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}, x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimated intercept in a least squares, ridge regression, or lasso model is zero, $\hat{\beta}_0 = 0$.

1. [2*pts*] Write out the ridge regression optimization problem explicitly in this setting.
2. [2*pts*] What does the solution space for ordinary least squares regression look like here?
3. [4*pts*] Show that in this example, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.
4. [2*pts*] Explain how this example connects to the statement that ridge regression tends to give similar coefficient values to correlated variables.

Solution.

1. We can write this ridge regression problem as

$$(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22}))^2 + \lambda(\beta_1^2 + \beta_2^2) \,.$$

2. Let us first try to use the standard solution for OLS regression ($\lambda = 0$), given by

$$\widehat{\beta} = (X^\top X)^{-1} X^\top y \,.$$

We have

$$X^\top X = \begin{pmatrix} x_{11}^2 + x_{21}^2 & x_{11}x_{12} + x_{21}x_{22} \\ x_{12}x_{11} + x_{22}x_{21} & x_{12}^2 + x_{22}^2 \end{pmatrix} = \begin{pmatrix} x_{11}^2 + x_{22}^2 & x_{11}^2 + x_{22}^2 \\ x_{11}^2 + x_{22}^2 & x_{11}^2 + x_{22}^2 \end{pmatrix}$$

by using the equalities $x_{11} = x_{12}$ and $x_{21} = x_{22}$. The matrix $X^\top X$ is therefore rank 1 and not invertible. We can, however, make use of these equalities to note that regressing y only on $(x_{11}, x_{12})^t$ will result in the same minimum loss as regressing on X. From the lecture we know that the simple linear regression is solved by

$$\widehat{\beta}_1 = \frac{x_{11}y_1 + x_{21}y_2}{x_{11}^2 + x_{21}^2} ,$$

so that $(\widehat{\beta}_1, 0)$ is a valid solution to the OLS regression.

There are more solutions, however. We know that $x_{11} = x_{21}$ and $x_{12} = x_{22}$. Thus, for any $\tau \in \mathbb{R}$, we have

$$\begin{aligned}\beta_1^* x_{11} + \beta_2^* x_{12} &= (\beta_1^* + \tau)x_{11} + (\beta_2^* - \tau)x_{12} \\ \beta_1^* x_{21} + \beta_2^* x_{22} &= (\beta_1^* + \tau)x_{21} + (\beta_2^* - \tau)x_{22}\end{aligned}$$

so that every $\beta = (\beta_1^* + \tau, \beta_2^* - \tau)$ solves the OLS regression.

3. We know that the solution to the ridge regression problem with $\lambda > 0$ is given by

$$\widehat{\beta} = (X^\top X + \lambda I)X^\top y ,$$

which we can simply evaluate as follows

$$(X^\top X + \lambda)^{-1} = \frac{1}{(a+\lambda)^2 + a^2}\begin{pmatrix} a+\lambda & -a \\ -a & a+\lambda \end{pmatrix}$$

where $a = x_{11}^2 + x_{22}^2$. We further have

$$X^\top y = \begin{pmatrix} x_{11}y_1 + x_{21}y_2 \\ x_{12}y_1 + x_{22}y_2 \end{pmatrix} = \begin{pmatrix} x_{11}y_1 + x_{22}y_2 \\ x_{11}y_1 + x_{22}y_2 \end{pmatrix}$$

by using our trusty equations above. By multiplying these two quantities, we therefore obtain

$$\widehat{\beta} = \frac{\lambda}{(a+\lambda)^2 + a^2}\begin{pmatrix} x_{11}y_1 + x_{22}y_2 \\ x_{11}y_1 + x_{22}y_2 \end{pmatrix} ,$$

showing that $\widehat{\beta}_1 = \widehat{\beta}_2$.

4. We clearly see that for the perfectly correlated variables in this exercise, ridge regression will give *exactly* the same values to $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

By looking at the solution above, we see also that when we have only approximate equalities $x_{12} \approx x_{11}$ and $x_{21} \approx x_{22}$ then $(X^\top X + \lambda)^{-1}$ will have approximately the same form as above and $X^\top y$ will also have approximately equal entries, so that the solution would also have $\widehat{\beta}_1 \approx \widehat{\beta}_2$.

Problem 4 (P, 20 Points). **Programming Exercise.**
Go through *5.3 Lab: Cross-Validation and the Bootstrap* (ISLR p.212–218) as well as *6.5.2 Lab: Ridge Regression and the Lasso* (ISLR p. 274–278). The objective of this programming exercise is to predict the logarithm of the prostate specific antigen (PSA) level using several predictors. You can find the data *prostate_train.csv, prostate_test.csv* in the CMS.

1. [1*pts*] Load the data. Make sure to split the training, respectively test, dataset into one dataset containing the target column **lpsa**, and another one containing the remaining features.

2. [4*pts*] Use `sklearn.linear_model.Ridge` to fit a ridge regression model on the training data and predict **lpsa** from all other features. Repeat this for different regularization parameters λ and plot the resulting coefficient values in relation to λ (cf. Figure 6.4, p. 238, ISLR). What can you observe?

3. [4*pts*] Use `sklearn.model_selection.KFold` to perform 10-fold cross-validation on the training data and determine the optimal value for λ in the ridge regression model. Report the average train and test MSE for this λ, i.e., the average of the train and test MSEs on the holdout datasets.

4. [4*pts*] Now use `sklearn.linear_model.Lasso` to fit Lasso models for **lpsa**, and plot the Lasso coefficients in relation to the regularization parameter λ (cf. Figure 6.6, p. 242, ISLR). What can you say in comparison to the plot in part 2 of this exercise? Make at least two observations.

5. [5*pts*] Perform 10-fold cross-validation on the training data to determine the optimal value for λ in Lasso. Again report average train and test MSEs for this λ. How many and which features are used? Compare this to the coefficients determined for ridge regression in part 3 of this exercise.

6. [2*pts*] Compare the performance, in terms of MSE, of the best models generated in parts 3. and 5. Which model would you choose and why? What alternative model could have been used?

Problem 5 (Bonus). **Coefficient Uncertainty.**
This bonus problem has both theoretical and practical parts.

1. **Theoretical**. Consider the regularized regression problem

$$\underset{\beta}{\text{minimize}} \quad (y - X\beta)^T(y - X\beta) + \lambda|\beta|^q \ , \tag{5.1}$$

which becomes ridge regression for $q = 2$, respectively Lasso for $q = 1$. While minimizing the above objective gives us a *point estimate* of β, you have seen in the lecture how a validation approach such as bootstrapping can show us a *distribution* over possible values of β (e.g. Fig. 5.10 in ISLR).

However, bootstrapping can only be done after model estimation. In this exercise, we will investigate what happens if we include a distribution over β directly in our model.

(a) We consider a linear model with independent, Gaussian distributed samples of Y, that is, $p(y_i \mid x_i, \beta_i) \sim \mathcal{N}(y_i \mid x_n^T\beta, \sigma^2)$ for $i = 0, .., n$. Write out $p(Y \mid X, \beta)$.

(b) We now also place a distribution over β. For now, we assume a Gaussian distribution, where $p(\beta) \sim \mathcal{N}(0, 1/\lambda)$ for some parameter λ. Consider

$$\hat{\beta} = \max_{\beta} \ \log\Big(p(\beta) \ p(Y \mid X, \beta)\Big) \ . \tag{5.2}$$

Simplify the above expression. Where is the maximum attained?

(c) Compare the expression in (b) with ridge regression, i.e. Eq. (5.1) where $q = 2$. Can you draw a connection between accounting for the coefficient distribution and regularization?

(d) How would you need to modify (b) to arrive at ordinary least squares regression, Eq. (5.1) with $\lambda = 0$? Explain what this means for the distribution over β.

(e) How would you need to modify (b) to arrive at the Lasso regression, Eq. (5.1) with $q = 1$? Explain what this means for the distribution over β.

2. **Practical**. We will now investigate this problem further on synthetic data.

(a) Write a Python function computing the expression you found for Eq. (5.2).

(b) Solve the expression in Eq. (5.2) for β. *Useful function:* `scipy.optimize`.

(c) Now, we sample different values for β from the Gaussian distribution we assumed. Create a histogram over these values, and also indicate the estimate $\hat{\beta}$ that you found in part (b).

(d) Repeat steps (b) and (c) for increasing numbers of datapoints, $n = 10, n = 100$ and $n = 1000$. Can you observe a trend?

Solution. To be discussed in the tutorial.