Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Classwork Solution Sheet #3: *Generalization*

Problem 1 (C, For Tutorials 04.12 and 05.12). Cross-Validation (Exercise 5.4.3 in ISLR)

1. Explain how k-fold cross-validation is implemented.

2. Discuss k-fold cross-validation in the context of the validation set approach and LOOCV. What are the advantages and disadvantages?

Solution.

1.

   1. Divide dataset **randomly** in k Groups of approximately equal size.
   2. For each fold i
      i. Fit model on folds $\{1, ..., k\} \setminus \{i\}$ (1 Point)

      ii. $error[i]$ = test model on $\{i\}$ (1 Point)

   3. return $\frac{1}{k}\sum_{i=1}^{k} error[i]$

2.

   i. More stable than validation set approach.
   ii. Faster than LOOCV. In general. As we will show in P2 LOOCV can be calculated by fitting one model for linear and polynomial least square regression.
   iii. k-fold is a "compromise" between the two approaches. For $k = 2$ essential validation set approach (depending on how the data is split). For $k = n$ k-fold is equal to LOOCV. (full points only when identifying this relationship between k-fold and validation set approach and LOOCV)

Problem 2 (C, For Tutorials 04.12 and 05.12). Subset selection (Exercise 6.8.1 in ISLR) We perform **best subset**, **forward stepwise** and **backward stepwise** selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, ..., p$ predictors.

1. Which of the three models, with k predictors, has the smallest training RSS? Justify your answer.

2. True or False:

   (a) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by forward stepwise selection.
   (b) The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by backward stepwise selection.
   (c) The predictors in the k-variable model identified by back- ward stepwise are a subset of the predictors in the (k + 1)- variable model identified by forward stepwise selection.
   (d) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.
   (e) The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

Solution.

1. Best subset selection selects for each k the best predictors whereas forward and backward selection do not reconsider predictors chosen in previous steps.
2. (a) True
   (b) True
   (c) False
   (d) False
   (e) False

Problem 3 (C, For Tutorials 11.12 and 12.12). Two stage linear regression
Consider a two-stage linear regression task on training data $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$. We construct a two-stage regressor with $W \in \mathbb{R}^{d\times p}$ where $p \leq d$ and $a \in \mathbb{R}^p$. The regression loss on training data is computed as

$$E(W, a) = \arg\min_{W,a} ||XWa - y||_2^2$$

1. Argue that $\hat{y} = X(X^TX)^{-1}X^Ty$ is optimal solution for our two stage regressor.
2. Show that $\arg\min_{W,a} ||XWa - y||_2^2 = ||XWa - \hat{y}||_2^2 + ||\hat{y} - y||_2^2$

Solution.

1. Observe that $Wa \in R^d$. Assume $p = d$ then $w = Wa$ can be an invertible linear map i.e. both injective and subjective. In cases where W is not an invertible map, the codomain of Wa is a proper subset of R^d. In the case of $p < d$, W is non-invertible. Then the optimal solution for $E(w) = \arg\min_w ||Xw - y||_2^2$ i.e. $\hat{y} = X(X^TX)^{-1}X^Ty$ is also the optimal solution for our two-stage linear problem.

2. $$\arg\min_{W,a} ||XWa - y||_2^2 = ||XWa - \hat{y} + \hat{y} - y||_2^2 \quad = ||XWa - \hat{y}||_2^2 + ||\hat{y} - y||_2^2 + 2(XWa - \hat{y})^T(\hat{y} - y)$$

   To complete our proof we need to show that $(XWa - \hat{y})^T(\hat{y} - y) = 0$

   $$\begin{aligned}
   &= (a^TW^TX^T - \hat{y}^T)(\hat{y} - y) \\
   &= a^TW^TX^T\hat{y} - a^TW^TX^Ty + \hat{y}^T\hat{y} - y^T\hat{y} \\
   &= a^TW^T\cancel{X^TX(X^TX)^{-1}}X^Ty - a^TW^TX^Ty - (X(X^TX)^{-1}X^Ty)^T(X(X^TX)^{-1}X^Ty) - y^TX(X^TX)^{-1}X^Ty \\
   &= a^TW^TX^Ty - a^TW^TX^Ty - y^TX\cancel{(X^TX)^{-1}X^TX}(X^TX)^{-1}X^Ty - y^TX(X^TX)^{-1}X^Ty \\
   &= y^TX(X^TX)^{-1}X^Ty - y^TX(X^TX)^{-1}X^Ty \\
   &= 0
   \end{aligned}$$