Cognitive Algorithms - Exercise Sheet 3

Linear Regression

Department of Machine Learning - TU Berlin

Disclaimer

For each exercise, but particularly for exercises involving calculations:

Show your work or you will not receive (full) credit!

Furthermore: Exercises marked with an asterisk * are not required and don't contribute towards the credit you receive, but you are welcome to do them because they are fun.

Task 1 - Ordinary Least Squares (OLS) Example [3 points]

Consider a data set with three data points,

$$x_1 = 0, x_2 = 1, x_3 = 2$$

with respective labels

$$y_1 = 0, y_2 = 1, y_3 = 0.$$

1. We want to fit a simple linear model $f(x) = w \cdot x$ to the data using ordinary least squares (OLS). Recall the OLS solution is obtained as

$$w = \arg\min_w \sum_{i=1}^{n} (y_i - f(x_i))^2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i x_i}$$

$$W = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i x_i} = \frac{0+1+0}{1+4} = \frac{1}{5}$$

   where $n = 3$ is the number of data points, $X = [x_1, x_2, x_3]$ and $\mathbf{y} = [y_1, y_2, y_3]$. Compute w. **[1 point]**

2. Now we want to fit a polynomial model $g(x) = w_1 \cdot x + w_2 \cdot x^2 = \mathbf{w}^\top \cdot \boldsymbol{\phi}(x)$ where we have defined a mapping $\boldsymbol{\phi} : \mathbb{R} \mapsto \mathbb{R}^2$ with

$$\boldsymbol{\phi}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

   and a weight vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. Recall the OLS solution is obtained as

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - g(x_i))^2 = (XX^\top)^{-1} X \mathbf{y}^\top$$

   where n and y are defined as above and

$$X = [\boldsymbol{\phi}(x_1), \boldsymbol{\phi}(x_2), \boldsymbol{\phi}(x_3)] = \begin{bmatrix} x_1 & x_2 & x_3 \\ (x_1)^2 & (x_2)^2 & (x_3)^2 \end{bmatrix}.$$

$$w = (XX^\top)^{-1} X y^\top$$

$$= \left(\begin{bmatrix} x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{bmatrix} \begin{bmatrix} x_1 & x_1^2 \\ x_2 & x_2^2 \\ x_3 & x_3^2 \end{bmatrix} \right)^{-1} \begin{bmatrix} x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2+x_2^2+x_3^2 & x_1^3+x_2^3+x_3^3 \\ x_1^3+x_2^3+x_3^3 & x_1^4+x_2^4+x_3^4 \end{bmatrix}^{-1} \begin{bmatrix} x_1y_1+x_2y_2+x_3y_3 \\ x_1^2y_1+x_2^2y_2+x_3^2y_3 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 9 \\ 9 & 17 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{1}{5 \times 17 - 81} \begin{bmatrix} 17 & -9 \\ -9 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} 8 \\ -4 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

   Compute $\mathbf{w}$ and the corresponding function $g(x)$. **[1 point]**

$$g(x) = \begin{bmatrix} 2 & -1 \end{bmatrix} \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$= 2x - x^2 = x(2-x)$$

Hint: The inverse of a 2×2 matrix can be calculated by the following formular

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}$$

3. Draw a 2D plot with the data points and the functions $f(x)$ and $g(x)$. **[1 point]**

Task 2 - Variance of OLS Estimation [2 points]

The following pseudocode computes the variance of the OLS estimator $\hat{w}$ of a simple regression:

Algorithm 1: Variance of the OLS Estimator

Input: n (number of data points); σ_ϵ^2 (noise variance); σ_x^2 (data variance); w (true slope)
Output: variance of $\hat{w}$

Create empty list $l_{\hat{w}} \leftarrow [\,]$
Generate n Gaussian data points $X = [x_1, \ldots, x_n]$, $x_i \sim \mathcal{N}(0, \sigma_x^2)$
for $r = 1, \ldots, 10^3$ **do**
    generate n Gaussian noise terms $E = [\epsilon_1, \ldots, \epsilon_n]$, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$
    compute $\mathbf{y} = w \cdot X + E$
    compute OLS estimate $\hat{w}_r = (XX^\top)^{-1} X \mathbf{y}^\top$
    append $\hat{w}_r$ to $l_{\hat{w}}$
return $\mathrm{Var}(l_{\hat{w}})$

Which of the input parameters influences the variance of $\hat{w}$ in which way? Complete the following statements.

1. If the number of data points n increases, the variance of $\hat{w}$ will **[0.5 points]**

   (a) decrease (b) increase (c) remain the same.

2. If the noise variance σ_ϵ^2 increases, the variance of $\hat{w}$ will **[0.5 points]**

   (a) decrease (b) increase (c) remain the same.

3. If the data variance σ_x^2 increases, the variance of $\hat{w}$ will **[0.5 points]**

   (a) decrease (b) increase (c) remain the same.

4. If the true slope w increases, the variance of $\hat{w}$ will **[0.5 points]**

   (a) decrease (b) increase (c) remain the same.

Task 3 - Bias-Variance Tradeoff [4 points]

Suppose there is a true, but unknown, non-linear relationship between a one-dimensional input x and a one-dimensional output y,

$$y = f(x) + \epsilon$$

where ϵ is uncorrelated noise. Suppose we observe n data points and model the relationship as an m-th order polynomial, i.e.

$$\hat{f}(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_m x^m.$$

The number of training points is fixed, and the parameters $w_0, w_1, \ldots, w_m$ are estimated by ordinary least squares regression (OLS), i.e. chosen such that $\sum_{i=1}^{n} (y_n - \hat{f}(x_i))^2$ is minimized.

1. Draw a sketch showing two curves: training error vs. the number of features m and test error vs. the number of features m. **[0.5 points]**
2. Annotate the plot with the two terms "Overfitting" and "Underfitting" **[0.5 points]**
3. Draw two more curves in a second sketch: The bias of $\hat{f}$ and the variance of $\hat{f}$ against the number of features m. Recall: A low bias means that on average (over different training sets) we accurately estimate f. A low variance of the model means that the estimated $\hat{f}$ will not change much if the training set varies. **[1 point]**
4. Suppose we choose m such that we are in the "overfitting" region, but we use ridge regression with a (good) regularisation parameter $\lambda > 0$, i.e. we chose $w_0, w_1, \ldots, w_m$ such that

$$\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 + \lambda \sum_{i=1}^{m} w_i^2$$

is minimized. Compared to OLS,

   (a) will the training error decrease,increase or is it ambigious? **[0.5 points]**

   (b) will the test error decrease, increase or is it ambigious? **[0.5 points]**

   (c) will the bias of $\hat{f}$ decrease, increase or is it ambigious? **[0.5 points]**

   (d) will the variance of $\hat{f}$ decrease, increase or is it ambigious? **[0.5 points]**

Task 4 - Invariance under transformations [4 points]

In this task we want to analyse if the OLS estimator and the ridge regression estimator are invariant under certain transformations. Using the notation of the lecture $X \in \mathbb{R}^{d \times n}$ and $\mathbf{y} \in \mathbb{R}^{1 \times n}$, the estimators are given as

$$\begin{aligned}
\hat{\mathbf{w}}_{\text{OLS}} &= (XX^\top)^{-1}X\mathbf{y}^\top \\
\hat{\mathbf{y}}_{\text{OLS}} &= \hat{\mathbf{w}}_{\text{OLS}}^\top X \\
\hat{\mathbf{w}}_{\text{RR}} &= (XX^\top + \lambda I)^{-1}X\mathbf{y}^\top \\
\hat{\mathbf{y}}_{\text{RR}} &= \hat{\mathbf{w}}_{\text{RR}}^\top X
\end{aligned}$$

We analyse invariance with respect to linear transformations of the data, $X \mapsto AX$ where $A \in \mathbb{R}^{d \times d}$ is an invertible matrix. Invariance means that the estimator is the same on the original data than on the transformed data.

1. Show that $\hat{\mathbf{y}}_{\text{OLS}}$ is invariant under arbitrary transformations A,
but $\hat{\mathbf{w}}_{\text{OLS}}$ is not. **[2 points]**
2. Show that $\hat{\mathbf{y}}_{\text{RR}}$ is invariant under orthogonal transformations A. **[2 points]**

Task 5 - Cross Validation [4 points]

1. You are a reviewer for a international conference on machine learning and you read a paper that selected a small number of features out of a large number of features for a given classification problem. The paper argues as follows:

   (a) We uses all our available data to select a subset of "good" features that had fairly strong correlation with the class labels.

   (b) Our final model contained only those features. We evaluate the prediction error of the final model by 10-fold crossvaldiation on all the available data.

(c) We obtained a low cross-validation error. Thus, we have achieved high classification accurarcy with only few meaningful features. (This is novel and amazing.)

Would you accept or reject the paper? Why? **[2 points]**

2. Suppose you are testing a new algorithm on a data set consisting of 100 positive and 100 negative examples. You plan to use leave-one-out cross-validation (that is 200-fold cross-validation) and compare your algorithm to a baseline function, a simple majority classifier. You expect the majority classifier to achieve about 50% classification accuracy, but to your surprise, it scores zero every time. Why?

   Majority Classifier: Given a set of training data, the majority classifier always outputs the class that is in the majority in the training set, regardless of the input. **[2 points]**