Task 1 - Ordinary Least Squares (OLS) Example

Consider a data set with three data points, $x_1 = 0, x_2 = 1, x_3 = 2$ with respective labels $y_1 = 0, y_2 = 1, y_2 = 0$.

1. We want to fit a simple linear model $f(x) = \omega \cdot x$ to the data using Ordinary Least Squares (OLS). Recall the OLS solution is obtained as

$$\omega = \underset{\omega}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - f(x_n))^2 = (XX^\top)^{-1} X y^\top \tag{1}$$

   where $N = 3$ is the number of data points, $X = [x_1, x_2, x_3]$ and $y = [y_1, y_2, y_3]$. Compute ω.

2. Now we want to fit a polynomial model $g(x) = w_1 \cdot x + w_2 \cdot x^2 = \mathbf{w}^T \cdot \phi(x)$ where we have defined a mapping $\phi : \mathbb{R} \ni x \mapsto \begin{bmatrix} x \\ x^2 \end{bmatrix} \in \mathbb{R}^2$ and a weight vector $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. Recall the OLS solution is obtained as

$$\mathbf{w} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{n=1}^{N} (y_n - g(x_n))^2 = (XX^\top)^{-1} X y^\top \tag{2}$$

   where N and y are defined as above and $X = [\phi(x_1), \phi(x_2), \phi(x_3)] = \begin{bmatrix} x_1 & x_2 & x_3 \\ (x_1)^2 & (x_2)^2 & (x_3)^2 \end{bmatrix}$. Compute $\mathbf{w}$ and the corresponding function $g(x)$.

   (You might need a calculator and the following formula: $\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \frac{1}{ac-b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}$.)

3. Draw a 2D plot with the data points and the functions $f(x)$ and $g(x)$.

Task 2 - Variance of OLS Estimation

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

Algorithm 1: Variance of the OLS Estimator

Require: Number of Data points N, Noise variance σ_ϵ^2, true slope ω

```
# Generate N data points X = [x_1, ..., x_N] from a Gaussian distribution, x_i ~ N(0, 1)
for Repetition r = 1, ..., 10^5 do
    # Generate N noise terms E = [ε_1, ..., ε_N] from a Gaussian distribution, ε_i ~ N(0, σ_ε^2)
    # Compute y = ω · X + E
    # Compute OLS estimate ω̂[r] = (XX^⊤)^{-1} X y^⊤
end for
Output: Variance of ω̂
```

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

1. If the number of data points N increases, the variance of $\hat{\omega}$ will
   (a) decrease (b) increase (c) remain the same.

2. If the noise variance σ_ϵ^2 increases, the variance of $\hat{\omega}$ will
   (a) decrease (b) increase (c) remain the same.

Task 3 - Bias-Variance Tradeoff

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 oberseve N data points and model the relationship as an mth 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 $\sum_{n=1}^{N}(y_n - \hat{f}(x_n))^2$ 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.
2. Annotate the plot with the two terms "Overfitting" and "Underfitting"
3. Draw two more curves (in the sketch or in a second sketch): The bias of $\hat{f}$ and the variance of $\hat{f}$. 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 $\hat{f}$ won't change much if the training set varies.
4. Suppose we chose m such that we are in the "Overfitting" region, but we use Ridge Regression with a regularisation parameter $\lambda > 0$, i.e. we chose $w_0, w_1, \ldots, w_m$ such that $\sum_{n=1}^{N}(y_n - \hat{f}(x_n))^2 + \lambda \sum_{i=1}^{m} w_i^2$ is minimized. Compared to OLS,
   (a) will the training error decrease or increase? (Or is it ambigious?)
   (b) will the test error decrease or increase? (Or is it ambigious?)
   (c) will the bias of $\hat{f}$ decrease or increase? (Or is it ambigious?)
   (d) will the variance of $\hat{f}$ decrease or increase? (Or is it ambigious?)

Task 4 - Invariance under transformations

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

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

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 $\hat{y}_{ols}$ is invariant under arbitrary transformations A, but $\hat{\mathbf{w}}_{ols}$ is not.
2. Show that $\hat{y}_{ridge}$ is invariant under orthogonal transformations A (i.e. $AA^\top = A^\top A = I$).

Solutions

Task 1 - Ordinary Least Squares (OLS) Example

1. $X = \begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$, $y = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$
   Plugging in: $\omega = (XX^\top)^{-1}Xy^\top = 1/5$, and $f(x) = 1/5 \cdot x$

2. $X = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 1 & 4 \end{bmatrix}$, $y = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$
   Plugging in: $\mathbf{w} = (XX^\top)^{-1}Xy^\top = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, and $g(x) = 2x - x^2 = -(x^2-1)^2 + 1$

Task 2 - Variance of OLS Estimation

1. (a) decrease

2. (b) increase

Task 3 - Bias-Variance Tradeoff

See, for example, 'Optional reading: Bias-Variance tradeoff, James' on the ISIS course.

1. The curves should look like e.g. Figure 2.9 (Right) (with 'Number of features m' on the x-axis)

2. 'Overfitting' is in the right part of Figure 2.9 (Right), 'Underfitting' in the left part.

3. The curves should look like e.g. Figure 2.12 (with 'Number of features m' on the x-axis)

4. (a) increase
   (b) decrease/ambigious
   (c) increase
   (d) decrease
   Information on Ridge Regression can be found for example in the two optional Bishop texts in the ISIS course ('Overfitting and regularisation', and 'Linear Regression').

Task 4 - Invariance under transformations

Denote with $\hat{\mathbf{w}}^A_{ols}$, $\hat{y}^A_{ols}$, $\hat{\mathbf{w}}^A_{ridge}$, $\hat{y}^A_{ridge}$ the respective estimators on the transformed data. We then have:

$$\begin{aligned}
\hat{\mathbf{w}}^A_{ols} &= (AX(AX)^\top)^{-1}(AX)y^\top = (A^\top)^{-1}(XX^\top)^{-1}A^{-1}(AX)y^\top \\
&= (A^\top)^{-1}(XX^\top)^{-1}Xy^\top = (A^\top)^{-1}\hat{\mathbf{w}}_{ols} \\
\hat{y}^A_{ols} &= (\hat{\mathbf{w}}^A_{ols})^\top AX = ((A^\top)^{-1}\hat{\mathbf{w}}_{ols})^\top AX = \hat{\mathbf{w}}^\top_{ols}X = \hat{y}_{ols}
\end{aligned}$$

For orthogonal matrices A we also have:

$$\begin{aligned}
\hat{\mathbf{w}}^A_{ridge} &= (AX(AX)^\top + \lambda I)^{-1}AXy^\top = (AX(AX)^\top + \lambda AA^\top)^{-1}AXy^\top \\
&= (A(XX^\top + \lambda I)A^\top)^{-1}AXy^\top = (A^\top)^{-1}\hat{\mathbf{w}}_{ridge} \\
\hat{y}^A_{ridge} &= (\hat{\mathbf{w}}^A_{ridge})^\top AX = ((A^\top)^{-1}\hat{\mathbf{w}}_{ridge})^\top AX = \hat{\mathbf{w}}^\top_{ridge}X = \hat{y}_{ridge}
\end{aligned}$$