

# The Mathematics of Machine Learning

## Homework Set 3

Due 10 March 2022 before 13:00  
via Canvas

You are allowed to work on this homework in pairs. One person per pair submits the answers via Canvas. Make sure to put both names on the submission.

### 1 Theory Exercises

Let  $\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i$  be the mean of the feature vectors, and let  $\bar{y} = \frac{1}{N} \sum_{i=1}^n y_i$  be the mean of the response vectors in the training data. Centering the features is a pre-processing procedure, which replaces all feature vectors  $x_i$  by

$$x_i \mapsto x_i - \bar{x}.$$

In the context of getting rid of the intercept, the 4-th lecture claimed the following result:

**Theorem 1.** *Let  $\lambda \geq 0$ . Then, for any regression estimator of the form*

$$(\hat{\beta}_0, \hat{\beta}) = \arg \min_{(\beta_0, \beta)} \sum_{i=1}^N (y_i - \beta_0 - x_i^\top \beta)^2 + \lambda \text{pen}(\beta),$$

*centering the features only changes the intercept  $\hat{\beta}_0$ , but not  $\hat{\beta}$ . Moreover, after centering, the estimated intercept is always  $\hat{\beta}_0 = \bar{y}$ .*

1. This exercise is about proving Theorem 1.

- (a) Prove the first part of the Theorem, that centering only changes  $\hat{\beta}_0$ .  
*Hint 1: This part actually holds more generally, for a shift of the features  $x_i \mapsto x_i - a$  by any constant vector  $a$ .*  
*Hint 2: As an intermediate step, show that for any  $\beta_0, \beta$  there exists a  $\beta'_0$  such that*

$$\beta_0 + x_i^\top \beta = \beta'_0 + (x_i - \bar{x})^\top \beta \quad \text{for all } i = 1, \dots, N. \quad (1)$$

- (b) Prove the second part of the Theorem, that, after centering, the estimated intercept is always  $\bar{y}$ .

*Hint: optimize  $\beta_0$  for fixed  $\beta$  and interpret how the optimal value varies with  $\beta$ .*

Another result claimed in the lecture was about the bias-variance decomposition for regression. Let  $\hat{f}$  be any estimator, depending on the training data  $T = (x_1, y_1), \dots, (x_N, y_N)$ , and let  $\bar{f} = \mathbb{E}_T[\hat{f}]$  be the average of the estimated functions, i.e.  $\bar{f}(x) = \mathbb{E}_T[\hat{f}(x)]$  for all new inputs  $x$ , and let  $f_B = \arg \min_f \text{EPE}(f)$  be the Bayes-optimal predictor, which we know equals  $f_B(x) = \mathbb{E}[Y | X = x]$ .

**Theorem 2.** *Consider regression with the squared loss. Then the expected prediction error for any estimator  $\hat{f}$  can be decomposed into the following three parts:*

$$\begin{aligned} \mathbb{E}_T[\text{EPE}(\hat{f})] &= \mathbb{E}_X[\text{Var}(Y|X)] && \text{(Bayes optimal EPE)} \\ &+ \mathbb{E}_X [(\bar{f}(X) - f_B(X))^2] && \text{(bias squared)} \\ &+ \mathbb{E}_{T,X} [(\hat{f}(X) - \bar{f}(X))^2] && \text{(variance)}. \end{aligned}$$

2. Prove Theorem 2.

*Hint: One way to prove the result is by repeated use of the following identity, which holds for any random variables  $A, B, C$ :*

$$\mathbb{E}[(A-B)^2] = \mathbb{E}[(A-C+C-B)^2] = \mathbb{E}[(A-C)^2] + 2\mathbb{E}[(A-C)(C-B)] + \mathbb{E}[(C-B)^2].$$

3. Suppose that the responses  $y$  are generated by a linear model with some parameters  $\beta_0^*, \beta^*$  and with noise  $\epsilon$ :

$$y = \beta_0^* + x^\top \beta^* + \epsilon,$$

where the conditional mean of the noise is zero:  $\mathbb{E}[\epsilon | X = x] = 0$ .

- (a) Show that  $f_B(x) = \beta_0^* + x^\top \beta^*$ .

- (b) Show that the least-squares estimator  $\hat{f}$  is unbiased (has zero bias), i.e. that  $\bar{f} = f_B$ .

*Hint: Consider the formula for the least-squares parameters  $(\hat{\beta}_0, \hat{\beta})$  and take expectation over all responses  $y_1, \dots, y_N$  conditional on all feature vectors  $x_1, \dots, x_N$ .*