The Mathematics of Machine Learning
Homework Set 3

Due 9 March 2023 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*

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

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 β_0, β there exists a β_0' such that

   $$\beta_0 + x_i^\top \beta = \beta_0' + (x_i - \bar{x})^\top \beta \qquad \textit{for all } i = 1, \ldots, N. \tag{1}$$

(b) Prove the second part of the Theorem, that, after centering, the estimated intercept is always $\bar{y}$.
Hint: optimize β_0 for fixed β and interpret how the optimal value varies with β.

Another result, claimed in lecture 3, was about the bias-variance decomposition for regression. Let $\hat{f}$ be any estimator, depending on the training data $T = (x_1, y_1), \ldots, (x_N, y_N)$, 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 $f_{\mathrm{B}} = \arg\min_f \mathrm{EPE}(f)$ be the Bayes-optimal predictor, which we know equals $f_{\mathrm{B}}(x) = \mathbb{E}[Y \mid X = x]$.

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:*

$$
\begin{aligned}
\mathop{\mathbb{E}}_{T}[\mathrm{EPE}(\hat{f})] = & \mathop{\mathbb{E}}_{X}[\mathrm{Var}(Y|X)] && \textit{(Bayes optimal EPE)} \\
& + \mathop{\mathbb{E}}_{X}\left[\left(\bar{f}(X) - f_{\mathrm{B}}(X)\right)^2\right] && \textit{(bias squared)} \\
& + \mathop{\mathbb{E}}_{T,X}\left[\left(\hat{f}(X) - \bar{f}(X)\right)^2\right] && \textit{(variance).}
\end{aligned}
$$

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:

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