ELEN0062 - Introduction to Machine Learning
Project 2 - Bias and variance analysis

October 25th, 2023

The goal of this second assignment is to help you better understand the important notions of bias and variance. The first part is purely theoretical, while the second part requires to perform some experiments with Scikit-learn. You should hand in a *brief* report giving your developments, observations and conclusions along with the scripts you have implemented to answer the questions of the second part. The project must be carried out by groups of at most *three students* and submitted on Gradescope[1] before *November 22, 23:59 GMT+2*. There will be two projects to submit to: one for your python scripts and one for your report.

1 Analytical derivations

Let us consider a regression problem with a single input variable and let us denote by

$$LS = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

the learning sample of N pairs, with $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. We will consider a simplified setting where the values of the input variable x_i of all training examples are **fixed**, while the output values y_i are given by:

$$y_i = wx_i + \epsilon_i,$$

where the ϵ_i's are independent and identically drawn from a normal distribution $\mathcal{N}(0, \sigma^2)$. The ϵ_i's are thus the only source of randomness in the learning sample.

We consider in this question the ridge regression algorithm which computes its predictions as follows:

$$\hat{f}_{LS}(x) = \hat{w}_{LS}x,$$

with $\hat{w}_{LS}$ defined by:

$$\hat{w}_{LS} = \arg\min_{w'} \{\sum_{i=1}^{N} (y_i - w'x_i)^2 + \lambda w'^2\},$$

with $\lambda \geq 0$ the regularization hyper-parameter.

(1.1) Give an analytical expression for the following quantities:

(a) $\hat{w}_{LS}$

(b) the bayes model, $h_B(x_0)$

(c) the residual error, noise(x_0)

(d) the squared bias, bias$^2(x_0)$

(e) the variance, variance(x_0)

at a specific test point x_0. You can express these quantities as a function of λ, σ, $s_{xx} = \sum_i x_i^2$, $s_{xy} = \sum_i x_i y_i$, w, and x_0.

[1] `https://www.gradescope.com`, Entry code: JK485B.

(1.2) From the resulting expressions, discuss the impact of the following parameters on both bias and variance:

(a) the regularization level, λ

(b) the noise, σ

(c) the learning sample size, N

Suggestion: to study the impact of N, you can assume that the x_i's have been drawn from a normal distribution $\mathcal{N}(0,1)$ and then show that in this case $s_{xx} \approx N$ when N is large.

(1.3) Derive an analytical expression of the optimal value of λ^*, i.e. the value of λ that leads to the smallest expected generalization error. Discuss the resulting expression.

2 Empirical analysis

In this section, we assume that we have access to a (large) dataset (or pool) of N_S pairs $P = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{N_S}, y_{N_S})\}$ from which we would like to design an experiment to estimate the bias and variance of several learning algorithms. We assume that N_S is large with respect to the size N of learning samples for which we want to estimate bias and variance. In the questions below, the different terms are to be understood as their mean over the input space ($E_{\mathbf{x}}$), not at a particular point $\mathbf{x}_0$ as in the previous section.

Data. For the experiments below, we propose to use the *Superconduct* dataset available in OpenML[2]. This dataset contains $N_S = 21263$ samples, described by 81 inputs. The goal is to predict the critical temperature of superconductors.

Note: in the following questions, we do not tell you precisely how to set all parameter values or ranges. It is your responsibility to choose them wisely to illustrate the expected behaviors. Some of the experiments may take time depending on your computer. Be patient.

(2.1) Explain why estimating the residual error term is very difficult in this setting.

(2.2) Describe a protocol to nevertheless estimate variance, the expected error, as well as the sum of the bias and the residual error from a pool P. Since the residual error is constant, this protocol is sufficient to assess how method hyper-parameters affect biases and variances.

(2.3) Implement and use this protocol on the given dataset to estimate the expected error, variance, and the sum of bias and residual error, for ridge regression, kNN, and regression trees. For all three methods, plot the evolution of the three quantities as a function of its main complexity parameter (respectively, λ[3], k, and maximum depth) on bias and variance. You can fix the learning sample size N to 500 for this experiment. Briefly discuss the different curves with respect to the theory.

(2.4) For the same three methods, show the impact of the learning sample size on bias and variance. In the case of kNN and ridge regression, choose one particular value of k and λ respectively. In the case of regression trees, compare fully grown trees with trees of fixed depth.

(2.5) One generic method to reduce variance is bagging (for "bootstrap aggregating"), which consists in growing several models from bootstrap samples drawn from the original LS and then to average their predictions. Apply this bagging[4] idea on all three methods, i.e., ridge regression (with fixed λ), kNN (with $k = 1$) and regression trees (fully grown), and evaluate its impact on bias and variance. Discuss the interest of bagging when combined with all three methods.

[2] A python script is provided on the project website to retrieve it.
[3] It is denoted λ in the lecture slides. In scikit-learn, this parameter is denoted α.
[4] You can use its implementation in scikit-learn.