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

October 2020

The goal of this second assignment is to help you to better understand the important notions of bias and variance. Make sure that your experiments are reproducible (*e.g.*, by fixing manually random seeds). We ask you to write a brief report (*pdf format*) giving your observations and conclusions. Answers are expected to be concise. You will need to write several scripts to answer some of the questions below, add all of them.

The assignment must be carried out by group[1] of *two students* and submitted as a tar.gz or zip file on Montefiore's submission plateform (`http://submit.montefiore.ulg.ac.be`) before November 17, 23:59 GMT+2. Note that attention will be paid to how you present your results. Careful thoughts in particular - but not limited to - should be given when it comes to plots.

1 Analytical derivations

1.1 Bayes model and residual error in classification

Let us consider a binary classification problem with an output $y \in \{-1, +1\}$ and two real input variables x_0 and x_1. Each sample $\boldsymbol{x}^i = (x_0^i, x_1^i)$ is generated by first selecting its class y^i at random (with an equal probability for each class), and then drawing their values from a multivariate Gaussian distribution

$$\begin{pmatrix} x_0^i \\ x_1^i \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^i \\ \rho^i & 1 \end{bmatrix}\right)$$

where ρ^i is class-dependent, *i.e.*, $\rho^i = \rho^+ > 0$ if $y^i = +1$ and $\rho^i = \rho^- = -\rho^+$ if $y^i = -1$.

(a) Derive an analytical formulation of the Bayes model $h_b(x_0, x_1)$ corresponding to the zero-one error loss. Justify your answer.

(b) Derive an analytical formulation of the residual error, *i.e.*, the generalization error of the Bayes model:

$$E_{x_0,x_1,y}\{1(y \neq h_b(x_0, x_1)\}.$$

Then, estimate its value using the analytical formulation (*i.e.*, *not* empirically) if $\rho^+ = 0.75$. Justify. Verify your estimation empirically.

1.2 Bias and variance of ridge regression

Let us consider a regression problem $y = f(\boldsymbol{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and let $LS = \{(\boldsymbol{x}_i, y_i | i = 1, \ldots, N\}$ denote the learning sample (of fixed size N), with $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$. Assuming for simplicity that we know that $f(0) = 0$ (so that no intercept is needed), we want to approximate this function with a linear model defined as $\hat{y}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{w}$, with $\boldsymbol{w} \in \mathbb{R}^p$. Let us consider the following two way to train the vector $\boldsymbol{w}$:

- Ordinary least-square: $\boldsymbol{w}_{OLS} = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^N (y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2$
- Ridge regression: $\boldsymbol{w}_R = \underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^N (y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2 + \lambda\boldsymbol{w}^T\boldsymbol{w}$

[1] See instructions on `https://people.montefiore.uliege.be/asutera/iml.php`.

Let us denote by $\boldsymbol{X}$ the $n \times p$ data matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$.

(a) Assuming that $\boldsymbol{X}$ is orthogonal (*i.e.*, $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$), show that

$$\boldsymbol{w}_R = \frac{\boldsymbol{w}_{OLS}}{1+\lambda}.$$

(b) Even if the data matrix is not orthogonal, let us assume that we use $\hat{y}(\boldsymbol{x}) = \dfrac{\boldsymbol{x}^T\boldsymbol{w}_{OLS}}{1+\lambda}$ as our model, with $\boldsymbol{w}_{OLS}$ defined as above.

i. Show analytically the relationships between the bias and variance of $\boldsymbol{x}^T\boldsymbol{w}_{OLS}$ and $\dfrac{\boldsymbol{x}^T\boldsymbol{w}_{OLS}}{1+\lambda}$;

ii. Explain the impact of λ on bias and variance on the basis of these formulas.

2 Empirical analyses

Let us consider a regression problem $y = f(x) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $x \sim \mathcal{U}(0,2)$. Let us denote by $LS = \{(x_i, y_i) | i = 1, \ldots, N\}$ the learning sample (of fixed size N) and by $\mathcal{A}$ a supervised learning algorithm.

(a) Give the analytical expressions of the residual error, the squared bias, the variance and the expected error *at a given point* x_0.

(b) Assuming that you can generate samples for a given value x (as is the case when f and the noise distribution are known), describe an experimental protocol to estimate *at a given point* x_0:

(i) the value of the Bayes model and the residual error,

(ii) the squared bias of the learning algorithm,

(iii) the variance of the learning algorithm,

(iv) the expected error of the learning algorithm,

Let us now assume that $f(x) = -x^3 + 3x^2 - 2x + 1$ and $\sigma^2 = 0.1$. Let us consider as candidate models $\hat{y}_m(x) = \sum_{i=0}^{m} a_i x^i$ of increasing degrees $0 \leq m \leq 5$ (*i.e.*, $\hat{y}_1(x) = a_0 + a_1 x^1$, $\hat{y}_2(x) = a_0 + a_1 x^1 + a_2 x^2$, $\ldots$), fitted by ordinary least-square[2]).

Answer the following questions by using your protocol of question (b).

(c) Estimate and plot the value of the Bayes model and the residual error for $x \in [0,2]$. Compare these (experimental) results with the analytical values.

(d) Estimate and plot the following quantities for $x \in [0,2]$ and using learning samples of size $N = 30$, for $m \in \{0, \ldots, 5\}$:

(i) the squared bias of $\hat{y}_m$,

(ii) the variance of $\hat{y}_m$,

(iii) the expected error of $\hat{y}_m$.

Discuss the impact of model complexity m on these quantities and in particular for their values in $x = 0$, $x = 0.5$, $x = 1$ and $x = 1.75$.

Remark: For sake of simplicity and to make the estimations more stable, you can use directly $f(x)$ in your protocol to answer this question (instead of its empirical estimate as obtained in (c)).

(e) Adapt the protocol of question (b) to estimate the mean values of the previous quantities over the input space. Plot the mean quantities for increasing values of m. Explain what you observe.

(f) Instead of OLS, use ridge regression[3] to fit the models. Setting m to 5, observe empirically the effect of the regularisation level $\lambda \in [0.0, 2.0]$ on squared bias, variance, and the error. Explain what you observe.

[2] This method is implemented in Scikit-learn by the function `LinearRegression`.

[3] This method is implemented in Scikit-learn by the function `Ridge`.