ELEN0062 - Introduction to machine learning
Project 2 - Bias and variance analysis

October 20th, 2021

This second project will help you to better understand the 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, and of the first part (if any). The project must be carried out by groups of *two* to *three students* and submitted as a `tar.gz` file on Montefiore's submission platform (`http://submit.montefiore.ulg.ac.be`) before *November 21, 23:59 GMT+2*.

1 Analytical derivations

1.1 Bayes model and residual error in classification

Let us consider the same setting as in the first project, i.e. a binary classification task with two real input variables where examples are sampled from two circular Gaussian distributions with identical covariance matrices $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, centered at $[1.5, 1.5]^T$ for the negative class and $[-1.5, -1.5]^T$ for the positive class. Negative examples are three times more likely than positive ones.

(a) Derive an analytical formulation of the Bayes model $h_b(x_1, x_2)$ corresponding to the zero-one error loss. Justify your answer.

(b) How does the Bayes model change if the ratio between both classes is modified? Justify your answer.

(c) Derive an analytical formulation of the residual error, i.e. the generalization error of the Bayes model $E_{x_1,x_2,y}\{\mathbb{1}(y \neq h_b(x_1, x_2))\}$.
Note: Go as far as possible in your analytical development.

(d) Estimate this quantity with $\sigma = 1.6$, from your analytical formulation. Verify your estimation empirically.

1.2 Bias and variance in regression

Let us consider a regression problem where each example (x, y) is generated as follows:

- The input x is drawn uniformly in $[0, 1]$.
- The output y is given by $y = ax + \varepsilon$, where $a \in \mathbb{R}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise variable (independent of x).

We are given a learning sample $LS = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ of N pairs to train a model and use the square error as the loss function.

We rely on a very simple algorithm, which will (incorrectly) assume that the output y does not depend on the input x, hence estimating it with a constant. More formally, we will have

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

where μ is a constant that will be estimated to minimize the square error on the learning sample.

(a) Compute the Bayes model along with its residual error.

(b) Compute the mean squared bias and variance of this learning algorithm, as a function of the learning sample size N.

(c) Explain the impact of σ and a on bias and variance.

2 Empirical analysis

Let us consider a regression problem (see Figure 1) where each sample (x, y) is generated as follows:

- The input x is drawn uniformly in $[-10, 10]$
- The output y is given by

  $$y = \sin 2x + x \cos(x - 1) + \varepsilon$$

  where $\varepsilon \sim \mathcal{N}(0, 0.5^2)$ is a noise variable.

(a) Let us assume that you can generate an infinite number of samples (either (x, y) pairs or output values y for a given input x_0). Describe a protocol to estimate the residual error, the squared bias and the variance at a given point x_0, and for a given supervised learning algorithm.
Note: The data generator can only be used as a black box and it should not provide any description of the underlying generating function.

(b) Describe a similar protocol to estimate the mean values of the residual error, the squared bias and the variance of the learning algorithm.

(c) Let us now assume that you only have access to a finite number of samples, and that you can no longer use the generator to produce more samples, are your protocols still appropriate? Discuss.

Let us consider a function `make_data` that can generate a set of N samples (x, y) according to the above description of the data.

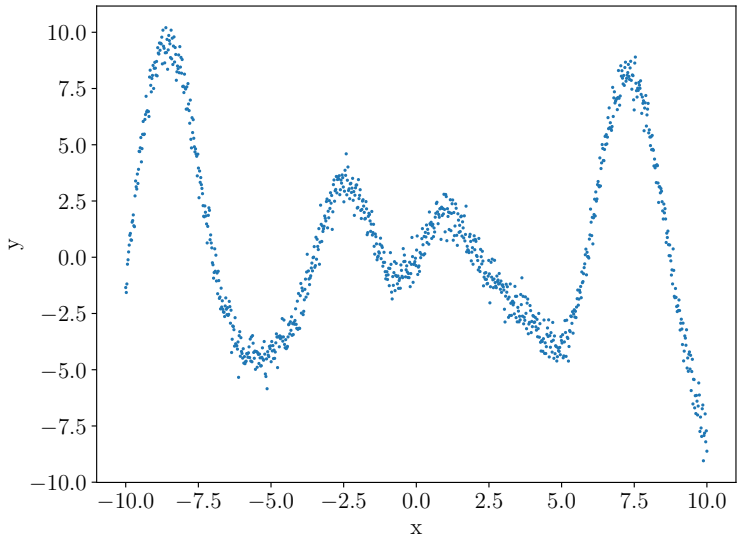

Figure 1: Illustration of the relation between x and y.

(d) Use your protocol (Question 2(a)) to estimate and plot the residual error, the squared bias, the variance and the expected error as a function of x for two regression methods of your choice (one linear and one non-linear). Comment your results.

(e) Use your protocol (Question 2(b)) to estimate the *mean* values of the squared error, the residual error, the squared bias and the variance for the same regression methods as a function of

- The size of the learning set,
- The model complexity,
- The standard deviation of the noise ε

Comment your results and support your observations with the appropriate plots.