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

November 2018

The aim of this project is to help you better understand the important notions of bias and variance. The first part is purely theoretical, while the second part is more practical and requires to perform experiments. Each project should be executed by groups of two students. We expect each group to provide:

- A *brief* report (in PDF format and of **maximum 10 pages**) collecting the answers to the different questions. Your report should include all necessary plots.
- The scripts you may have implemented to answer the questions of the second part.

The report and the scripts should be submitted as a tar.gz file on Montefiore's submission plateform (`http://submit.montefiore.ulg.ac.be`) before *November 21, 23:59 GMT+2*. You must concatenate your sXXXXXX ids as group name.

1 Theoretical questions

1.1 Bayes model and residual error in classification

As in the first project, let us consider a binary classification dataset with the output $y \in \{-1, +1\}$ and two real input variables x_0 and x_1. Each example is sampled by first selecting its class at random (with an equal probability for each class) and then sampling its input values from a class-dependent gaussian distribution:

- for the negative examples, a circular gaussian distribution centered at $(x_0, x_1) = (\mu_0^-, \mu_1^-) = (0, 0)$ and with a covariance matrix $\begin{pmatrix} \sigma_n^2 & 0 \\ 0 & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.
- for the positive samples, an ellipsoidal gaussian distribution centered at $(x_0, x_1) = (\mu_0^+, \mu_1^+) = (0, 0)$ and with a covariance matrix $\begin{pmatrix} \sigma_{p,0}^2 & 0 \\ 0 & \sigma_{p,1}^2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$.

Questions:

(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) Estimate the residual error, i.e., the generalization error of the Bayes model:

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

Justify your answer.

Remark: you can compute this error numerically using any software you want but you must explain how exactly the value is computed.

(c) Show and discuss how the Bayes model changes in the two following cases:

i. The gaussian distributions are no longer concentric, i.e., $(\mu_0^-, \mu_1^-) \neq (\mu_0^+, \mu_1^+)$;

ii. The diagonal terms $\sigma_{p,0}^2$ and $\sigma_{p,1}^2$ of the covariance matrix of the positive class become both greater or lower than σ_n^2.

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 $y = ax + \epsilon$, where $a \in \mathbb{R}$ is a constant and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ 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.

We use a very simple learning algorithm that assumes (wrongly) that the output y does not depend on the input x and therefore tries to estimate it as a constant. In other words, this algorithm considers models of the form

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

where μ is some constant to be estimated from LS.

Questions:

(a) Assuming that μ is learned to minimize the training error, compute the Bayes model and its residual error, and then the mean squared bias and variance of this learning algorithm as a function of the learning sample size N.

(b) Explain the impact of a and σ on the different terms.

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 $[-4, 4]$
- The output y is given by

$$y = \sin x + 0.5 \sin 3x + \epsilon,$$

  where $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ (with $\mu = 0$ and $\sigma^2 = 0.01$) is a noise variable.

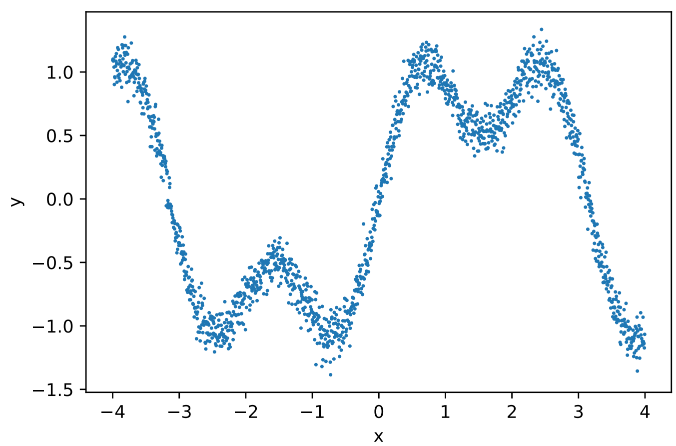

Figure 1: Illustration of the relation between x and y.

(a) Describe an experimental 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.

(b) Using this protocol, estimate and plot the residual error, the squared bias, the variance, and the expected error as a function of x for one *linear* and one *non-linear* regression method of your choice. Comment your results.

(c) Adapt the protocol of question (a) to estimate the mean values of the previous quantities over the input space.

(d) Use this protocol to study the *mean* values of the squared error, the residual error, the squared bias and the variance for the same algorithms as in question (b) as a function of:

- the size of the learning set;
- the model complexity[1];
- the standard deviation of the noise ϵ;

Explain your observations and support your conclusions with the appropriate plots.

[1] Depending on the selected methods, complexity can be measured explicitely by the size of the model or implicitely by some appropriate regularization parameters.