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

November 2019

The goal of this second assignment is to help you to better understand the important notions of bias and variance. 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 of two students and submitted as a tar.gz file on Montefiore's submission plateform (http://submit.montefiore.ulg.ac.be) before November 17, 23:59 GMT+2.

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 $\mathbf{x}^i = (x_0^i, x_1^i)$ is generated by first selecting its class at random (with an equal probability for each class) and then computing its input values as follows:

$$x_0^i = r^i \cos(\alpha^i),\ x_1^i = r^i \sin(\alpha^i)$$

where r^i follows a class-dependent distribution, i.e., $r^i \sim \mathcal{N}(R^-, \sigma^2)$ for the negative class and $r^i \sim \mathcal{N}(R^+, \sigma^2)$ for the positive one, $\alpha^i \sim \mathcal{U}(0, 2\pi)$ for both classes and $\sigma^2 = 0.1$.

(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 formation 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 if $R^+ = 2$ and $R^- = 1$.

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

2 Bias and variance of the kNN algorithm

Let us consider a unidimensional regression problem $y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and let $LS = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\}$ denote the learning sample (of fixed size N). To simplify the analysis, we further assume that the input values $\mathbf{x}^i$ of the N learning sample examples are fixed in advance, i.e., only their outputs y^i are random.

(a) Show that the generalization error of the k-Nearest Neighbours algorithm at some point $\mathbf{x}$ can be decomposed as follows

$$E_{LS}\{E_{y|\mathbf{x}}\{(y - \hat{y}(\mathbf{x}; LS, k))^2\}\} = \sigma^2 + \left[f(\mathbf{x}) - \frac{1}{k}\sum_{l=1}^{k} f(\mathbf{x}_{(l)})\right]^2 + \frac{\sigma^2}{k} \quad (1)$$

where $\hat{y}(\mathbf{x}; LS, k)$ is the prediction of the kNN method for a point $\mathbf{x}$ given a learning sample LS (of size N), $x_{(l)}$ denotes the l^{th} nearest neighbours of $\mathbf{x}$ in LS and k is the number of neighbours.

(b) Using question (a), discuss the effect of the number of neighbours k on each term of the bias-variance decomposition (Equation 1).

3 Bias and variance estimation

(a) Let us consider a regression problem for which you can generate an infinite number of samples, 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.
Important remark: The generator of samples can only be used as a black box and it does 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.

(c) If you only have a finite number of samples (and you can not use the generator), are your protocols still appropriate? Discuss.

Let us consider a function *make_data* that can generate a set of N samples (x_r, y) where

- $x_r \sim \mathcal{U}(-10, 10)$;
- $y = f(x_r) + \frac{1}{10}\epsilon$, where $f(x_r) = \sin(x_r) * e^{-\frac{x_r^2}{16}}$ and $\epsilon \sim \mathcal{N}(0, 1)$.

Note that some irrelevant variables $x_j \sim \mathcal{U}(-10, 10)$ can be added to the problem such that $f(x_r) = f(\mathbf{x})$ where $\mathbf{x} = (x_r, x_{j_1}, x_{j_2}, \ldots, x_{j_q})$ where q is the number of added irrelevant variables.

(d) Use your protocol (question 3(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 (one linear and one non-linear) of your choice. Comment your results.

(e) Use your protocol (question 3(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 number of irrelevant variables added to the problem.

Comment your results.