Applied Inductive Learning
Project 2 - bias and variance analysis

November 2016

The goal of this project is to help you better understand the important notions of bias and variance. The first part of the project is purely theoretical, while the second part requires to perform experiments with scikit-learn. Each project should be executed by groups of two students. We expect from each group to do:

- 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 python scripts you 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.run.montefiore.ulg.ac.be`) before *November 22, 23:59 GMT+2*. You must use you sXXXXXX ids as group name.

1 Theoretical questions

1.1 Bayes model and residual error in classification

Let us consider a classification problem where each example is described by two input features x_1 and x_2, and is associated to a class $y \in \{0,1\}$. To draw an example from the distribution $p(x_1, x_2, y)$, we proceed as follows:

- A class y is drawn uniformly at random from $\{0,1\}$.
- x_1 and x_2 are then computed as follows:

$$\begin{aligned} x_1 &= r\cos\alpha \\ x_2 &= r\sin\alpha, \end{aligned}$$

  where $\alpha \sim \mathcal{U}(0, 2\pi)$ and r's distribution depends on y as follows:

  - If $y = 0$, $r \sim Exp(1.5)$
  - If $y = 1$, $r \sim Exp(0.5)$

  where $Exp(\lambda)$ denotes an exponential distribution of parameter λ (probability density function is $\lambda exp(-\lambda r)$).

Figure 1 shows a sample of 200 examples drawn according to this procedure. For this procedure:

(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) Compute the generalization error of the Bayes model, ie. $E_{x_1,x_2,y}\{1(y \neq h_B(x_1, x_2))\}$. Justify your answer.

NB: If the solution to these questions involves the computation of integrals or intersections between functions, you can compute them numerically using any software you want.

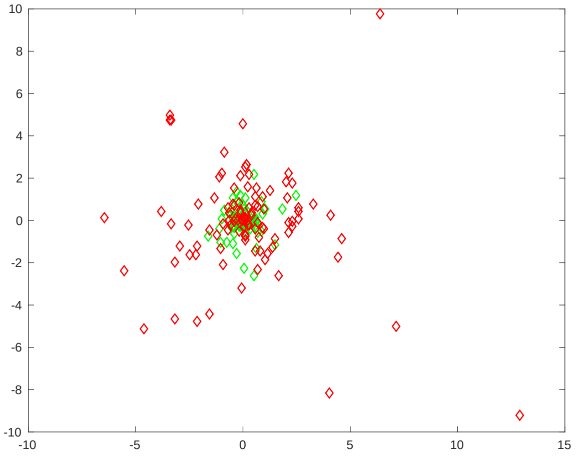

Figure 1: A sample of 200 points drawn from the distribution of Section 1.1. Red points correspond to $y = 1$ and green points to $y = 0$.

1.2 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:

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

where $\hat{y}(\mathbf{x}; LS, k)$ denotes the prediction of the kNN method at point $\mathbf{x}$ for a learning sample LS (of size N), $\mathbf{x}_{(l)}$ denotes the lth 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.

2 Empirical analysis

Let us consider a regression problem (see Figure 2) where each sample (x, y) is generated as follows:

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

$$y = x\left(\sin(x) + \cos(x)\right)^2 + \epsilon,$$

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

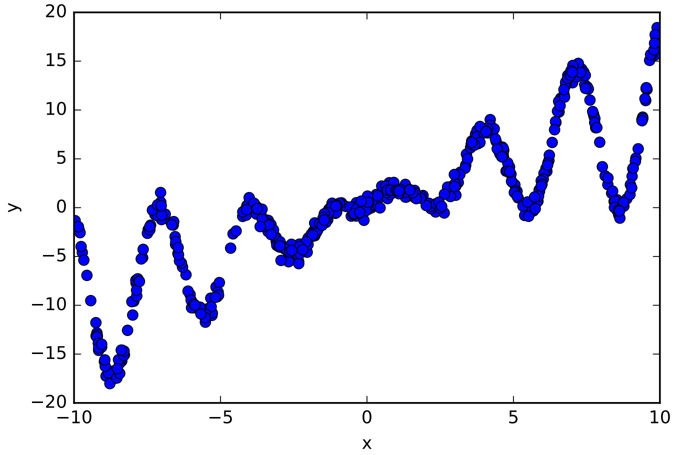

Figure 2: Illustration of the relation between x and y.

(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;
- the standard deviation of the noise ϵ;

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