Applied Inductive Learning
Project 2 - bias and variance analysis

November 2014

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 each group to send us:

- 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. A separate script should be provided for each subquestion.

A zip archive with all files must be sent before November 23, 23:59 to a.joly@ulg.ac.be with subject "[AIL][Project-2]NAME1-NAME2" where NAME1 and NAME2 are the names of the two members of the group.

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 $y \in \{A, B, C\}$. 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 $\{A, B, C\}$.
- Depending on the drawn class, feature values of the example are then drawn as follows\(^1\):
  - $y = A$: $x_1 \sim \mathcal{N}(0.2, 0.05^2)$, $x_2 \sim \mathcal{N}(0.5, 0.2^2)$
  - $y = B$: $x_1 \sim \mathcal{N}(0.5, 0.2^2)$, $x_2 \sim \mathcal{N}(0.5, 0.2^2)$
  - $y = C$: $x_1 \sim \mathcal{N}(0.8, 0.1^2)$, $x_2 \sim \mathcal{N}(0.5, 0.2^2)$

($x_1$ and $x_2$ are thus independent conditionally to $y$.)

(a) For this problem, derive an analytical formulation of the Bayes model $h_B(x_1, x_2)$ corresponding to the zero-one error loss.

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

\(^1\mathcal{N}(\mu, \sigma^2)$ denotes a gaussian distribution of mean $\mu$ and standard deviation $\sigma$.\)
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 \(y = x + \epsilon\), where \(\epsilon \sim \mathcal{N}(0, 1^2)\) 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 try to estimate it as a constant. In other words, this algorithm considers models of the form \(\hat{f}_{LS}(x) = \mu\), where \(\mu\) is some constant to be estimated from \(LS\).

(a) Show that the constant \(\mu\) that minimizes the mean square error on \(LS\) is given by:

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} y_i \]

(b) 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 \(N\).

2 Empirical analysis

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

- The input \(x\) is drawn uniformly in \([0, 8]\)
- The output \(y\) is given by

\[ y = \frac{\sin(x - 2)}{x - 2} x^3 + \epsilon, \]

where \(\epsilon \sim \mathcal{N}(0, 1)\) is a noise variable.

(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 squared error, the mean squared bias and the mean 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 \(\epsilon\).

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