Introduction to Machine Learning
Project 2 - bias and variance analysis

November 2017

The goal 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 requires to perform experiments with scikit-learn. 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 22, 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 two real input variables where the examples are sampled from two circular gaussian distributions with the same covariance matrices $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, centered at $(+1.5, +1.5)$ for the negative class and $(-1.5, -1.5)$ for the positive class. As in the first project, 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) Estimate the generalization error of the Bayes model, i.e. $E_{x_1,x_2,y}\{\mathbb{1}(y \neq h_b(x_1, x_2))\}$ with $\sigma = 1.6$. Justify your answer.

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

1.2 Bias and variance of ridge regression

Let us consider a regression problem $y = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and let $LS = \{(\mathbf{x}_i, y_i) | i = 1, \dots, N\}$ denote the learning sample (of fixed size N), with $y_i \in I\!R$ and $\mathbf{x}_i \in I\!R^p$. Assuming for simplicity that we know that $f(0) = 0$ (so that no intercept is needed), we want to approximate this function with a linear model defined as $\hat{y}(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$, with $\mathbf{w} \in I\!R^p$. Let us consider the following two ways to train the vector $\mathbf{w}$:

- Ordinary least-square: $\mathbf{w}_{OLS} = \underset{\mathbf{w} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{N} (y_i - \mathbf{x_i^T} \mathbf{w})^2$
- Ridge regression: $\mathbf{w}_R = \underset{\mathbf{w} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{N} (y_i - \mathbf{x_i^T} \mathbf{w})^2 + \lambda \mathbf{w}^T \mathbf{w}$

Let us denote by $\mathbf{X}$ the $n \times p$ data matrix $(\mathbf{x}_1, \dots, \mathbf{x}_N)^T$.

(a) Assuming that $\mathbf{X}$ is orthogonal (i.e., such that $\mathbf{X}^T\mathbf{X} = I$), show that

$$\mathbf{w_R} = \frac{\mathbf{w}_{OLS}}{1 + \lambda}.$$

(b) Even if the data matrix is not orthogonal, let us assume that we use $\hat{y}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{w}_{OLS}}{1+\lambda}$ as our model, with $\mathbf{w}_{OLS}$ defined as above. At a point $\mathbf{x}$ where the model $\mathbf{x}^T\mathbf{w}_{OLS}$ has zero bias:

  i Show how bias and variance are affected by λ

  ii Determine an analytical formulation of the value of λ that minimizes the expected generalization (square) error at x. Explain intuitively the resulting formula.

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
$$y = \frac{\sin x}{e^{-x}} + \epsilon,$$
where $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$ (with $\mu = 0$ and $\sigma^2 = 1$) 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.

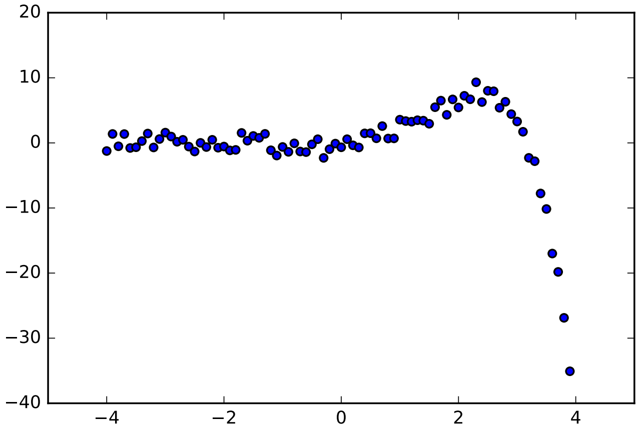

Figure 1: Illustration of the relation between x and y.

(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.