University of Stavanger

Stavanger, January 2, 2020

Theoretical exercise 6

ELE520 Machine learning

Problem 1

Bias and variance are most easily understood in the context of regression or curve fitting. Suppose there is a true (but unknown) function $F(\boldsymbol{x})$ with continuous valued output with noise, and we seek to estimate it based on N samples in a set $\mathcal{D}$ generated by $F(x)$. The regression function estimated is denoted $g(\boldsymbol{x};\mathcal{D})$ and we are interested in the dependence of this approximation on the training set $\mathcal{D}$. The estimate will vary with $\mathcal{D}$ and we can the effectiveness of the estimator by averaging over all training sets $\mathcal{D}$ according to

$$\mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - F(\boldsymbol{x}))^2]. \tag{1}$$

Show how

$$\underbrace{(\mathbb{E}_{\mathcal{D}}[g(\boldsymbol{x};\mathcal{D}) - F(\boldsymbol{x})])^2}_{bias^2} + \underbrace{\mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[g(\boldsymbol{x};\mathcal{D})])^2]}_{variance} \tag{2}$$

can be derived from (1). Can *bias* be negative? Can *variance* be negative?.

The following relationship might be useful:

$$\mathbb{E}_{\mathcal{D}}[g^2(\boldsymbol{x};\mathcal{D})] = \mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[g(\boldsymbol{x};\mathcal{D})])^2] + (\mathbb{E}_{\mathcal{D}}[g(\boldsymbol{x};\mathcal{D})])^2 \tag{3}$$

Problem 2

Prove that the jackknife estimate of the mean, the mean

$$\mu_{(\cdot)} = \frac{1}{N}\sum_{i=1}^{N}\mu_{(i)} \tag{1}$$

of the leave-one-out means

$$\mu_{(i)} = \frac{1}{N-1}\sum_{j\neq i}^{N} x_j. \tag{2}$$

is equivalent to the traditional estimate of the mean,

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{3}$$

Trygve Eftestøl, Professor
Faculty of Science and Technology
Department of Electrical and Computer Engineering
Kjølv Egelands hus
University of Stavanger
N-4036 Stavanger.
Telephone: +47 51 83 10 00.
Telefax: +47 51 83 17 50.
E-mail: trygve.eftestol@uis.no
www.ux.his.no/~trygve-e

Problem 3

Let $\mathbf{x}_1 = (0\ 0)^T$, $\mathbf{x}_2 = (1\ 1)^T$, $\mathbf{x}_3 = (1\ 0)^T$ and $\mathbf{x}_4 = (2\ 0.5)^T$. Consider the three following clusterings:

1. $\mathcal{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2\}$, $\mathcal{X}_2 = \{\mathbf{x}_3, \mathbf{x}_4\}$
2. $\mathcal{X}_1 = \{\mathbf{x}_1, \mathbf{x}_4\}$, $\mathcal{X}_2 = \{\mathbf{x}_2, \mathbf{x}_3\}$
3. $\mathcal{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\mathcal{X}_2 = \{\mathbf{x}_4\}$

Find the clustering that minimises the criterion based on sum-of-squared-error, J_e,

$$J_e = \sum_{i=1}^{M} \sum_{\mathbf{x} \in \mathcal{X}_i} \|\mathbf{x} - \mathbf{m}_i\|^2, \tag{1}$$

where $\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{x}$ and n_i is the number of samples in $\mathcal{X}_i$.

Problem 4

We have measured 7 instances of feature vectors, hereafter referred to as the training vectors. The result of the measurements were:

```
(1,1) (3,1) (2,3) (1,4.5) (2.5,1.5) (3,3) (3,4)
```

.

a) Use the "basic iterative minimum-squared-error clustering"-algorithm to cluster the data set into $M = 2$ groups.

   Use these initial cluster means

$$\begin{aligned} \mathbf{m}_1 &= (2\ 3)^T \\ \mathbf{m}_2 &= (3\ 1)^T. \end{aligned} \tag{1}$$

   You are supposed to select the samples by random. To be able to campere your results to those of the solution, it is recommended to consider the samples according to the following sequences:

   - *Iteration cycle 1:* $1, 2, 3, 7, 6, 4, 5$
   - *Iteration cycle 2:* $4, 7, 3, 2, 5, 6, 1$.

b) Define the *criterion function* used in the previous subtask and explain how it eveolves towards convergence. Explain briefly why the algorithm is guaranteed to converge.