Principles of Statistical Machine Learning
Introduction to Nearest Neighbors Learning

Fokoué Ernest, PhD
Professor of Statistics

@ErnestFokoue

To understand God's thoughts, one must study statistics, the measure of His purpose
Florence Nightingale

Exercise: 1NN Classification I

1. Consider the kNearest Neighbors learning machine with $k = 1$, also known as 1NN or NN learning machine. Given a sample of size n, namely $\mathscr{D}_n = \left\{(\mathbf{x}_i, \mathrm{y}_i) \overset{iid}{\sim} p_{\mathrm{xy}}(\mathbf{x}, \mathrm{y}),\ \mathbf{x}_i \in \mathscr{X}, \mathrm{y}_i \in \{1, 2, \cdots, G\}\right\}$,
   - Find the in sample prediction $\widehat{f}_{\mathtt{NN}}(\mathbf{x}_{i_m})$ for $\mathbf{x}_{i_m}$, where $i_m = \lceil (n+1)/2 \rceil$. If $\mathbf{x}_i \in \mathscr{D}_n$, then its nearest neighbor is itself $\mathbf{x}_i$, so that $\widehat{f}_{\mathtt{NN}}(\mathbf{x}_i) = \mathrm{y}_i$, because

$$\min_{\mathbf{x}_j \in \mathscr{D}_n} d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_i) = 0$$

Therefore, the nearest neighbor to $\mathbf{x}_i$ in $\mathscr{D}_n$ is $\mathbf{x}_i$

$$\mathbf{x}_i = \underset{\mathbf{x}_j \in \mathscr{D}_n}{\mathtt{argmin}}\, d(\mathbf{x}_i, \mathbf{x}_j).$$

Since $\mathbf{x}_{i_m} \in \mathscr{D}_n$, it follows that

$$\widehat{f}_{\mathtt{NN}}(\mathbf{x}_{i_m}) = \mathrm{y}_{i_m}.$$

Exercise: 1NN Classification II

- *Find the in-sample error rate under the zero-one loss.*
  Since $\widehat{f}_{\texttt{NN}}(\mathbf{x}_i) = \mathrm{y}_i$, it holds true that $\mathbb{1}(\mathrm{y}_i \neq \widehat{f}_{\texttt{NN}}(\mathbf{x}_i)) = 0$, for all $\mathbf{x}_i \in \mathscr{D}_n$. Therefore, it immediately follows that

$$\widehat{R}_n(\widehat{f}_{\texttt{NN}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\mathrm{y}_i \neq \widehat{f}_{\texttt{NN}}(\mathbf{x}_i)) = 0, \quad \forall \mathscr{D}_n \subset \mathscr{X} \times \mathscr{Y}.$$

- *Argue on a reasonable VC Dimension for*

$$\mathscr{H} = \{\mathbf{x}_i \mapsto f_{\texttt{1NN}}(\mathbf{x}_i) = \mathrm{y}_i, \quad \forall (\mathbf{x}_i, \mathrm{y}_i) \in \mathscr{D}_n\}$$

Since $\widehat{R}_n(\widehat{f}_{\texttt{NN}}) = 0, \forall n \in \mathbb{N}$, it follows naturally that

$$\texttt{VCdim}(\mathscr{H}) = +\infty.$$

This result has a lot of consequences on the generalizability of the nearest neighbors learning machine when $k = 1$.

Exercise: 1NN Classification III

2. *Consider the kNearest Neighbors learning machine with $k = 1$, also known as 1NN or NN learning machine. Given a sample of size n, namely $\mathscr{D}_n = \{(\mathbf{x}_i, \mathrm{y}_i) \overset{iid}{\sim} p_{\mathrm{xy}}(\mathbf{x}, \mathrm{y}),\ \mathbf{x}_i \in \mathscr{X}, \mathrm{y}_i \in \{0, 1\}\}$, with $\pi = \Pr[Y_i = 1] = \mathbb{E}[Y_i]$ for all $i = 1, \cdots, n$.*
   - *What is the bias of this learning machine?*
     Now, $\forall \mathbf{x}_i \in \mathscr{D}_n$, the point-wise bias $\widehat{f}_{\mathtt{NN}}(\mathbf{x}_i)$ of $\widehat{f}_{\mathtt{NN}}$ is

$$\mathtt{Bias}_n(\widehat{f}_{\mathtt{NN}}(\mathbf{x}_i)) = \mathbb{E}[\widehat{f}_{\mathtt{NN}}(\mathbf{x}_i)] - f^\star(\mathbf{x}_i)$$

where $f^\star(\mathbf{x}_i)$ is the theoretical (true) underlying response yield by the generator of the data, corresponding to the Bayes learning machine prediction.

1. *Consider the kNearest Neighbors learning machine with $k = n$, also known as nNN learning machine. Given a sample of size n, namely* $\mathscr{D}_n = \{(\mathbf{x}_i, \mathrm{y}_i) \overset{iid}{\sim} p_{\mathrm{xy}}(\mathbf{x}, \mathrm{y}),\ \mathbf{x}_i \in \mathscr{X}, \mathrm{y}_i \in \mathbb{R},\ i = 1, \cdots, n\}$,
   - *Find the in sample prediction $\widehat{f}_{\mathtt{nNN}}(\mathbf{x}_{i_m})$ for $\mathbf{x}_{i_m}$, where $i_m = \lceil (n+1)/2 \rceil$.* $\forall \mathbf{x}_i \in \mathscr{D}_n$, *it is true that* $\widehat{f}_{\mathtt{nNN}}(\mathbf{x}_i) = \mathtt{constant} = \bar{\mathrm{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{y}_i$, *therefore it follows that*
   $$\widehat{f}_{\mathtt{nNN}}(\mathbf{x}_{i_m}) = \bar{\mathrm{y}}, \quad \textit{since } \mathbf{x}_{i_m} \in \mathscr{D}_n.$$
   - *Find the in-sample error rate under the squared error loss* $\mathcal{L}(\mathrm{y}, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$.
   $$\widehat{R}_n(\widehat{f}_{\mathtt{nNN}}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\mathrm{y}_i, \widehat{f}_{\mathtt{nNN}}(\mathbf{x}_i)) = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{y}_i - \bar{\mathrm{y}})^2 = \left(\frac{n-1}{n}\right) S_{\mathrm{y}}^2$$
   - *What is the variance of this learning machine? Since* $\widehat{f}_{\mathtt{nNN}}(\mathbf{x}_i) = \mathtt{constant} = \bar{\mathrm{y}}$ *for all* $\mathbf{x}_i \in \mathscr{D}_n$, *it follows that*
   $$\mathtt{Variance}(\widehat{f}_{\mathtt{nNN}}) = 0 \quad \textit{smallest possible value.}$$