Proofs and correction exercises on *The Elements of Statistical Learning*

Esther Boccara

3 août 2017

Chapter 2 : Overview of Supervised Learning

Exercise 1

In our context, each of our output variables $(Y_1, ..., Y_n)$ is categorical and belongs to some set of cardinal K. Let denote them as follows :

For $i = 1, ..., n$, $Y_i \in \{0,1\}^K$ and $\sum_j Y_i(j) = 1$. Given some new input $x \in \mathbb{R}^p$, we want to predict the group it belongs to. Therefore, we estimate the vector of probabilities $\widehat{p} := (\widehat{p}_1, ..., \widehat{p}_K)$, where $\widehat{p}_k$ denotes the (estimated) probability that x belongs to group k, with $k = 1, ..., K$. The predicted class is the integer $\bar{k}$ that maximizes the $\widehat{p}_k$-s. Furthermore, for all integer $k = 1, ..., K$, we denote by t_k a vector of all zeros except a 1 at k-th position.

Show that $\bar{k} = \arg\max_k \widehat{p}_k = \arg\min_k \|t_k - \widehat{p}\|$.

Exercise 1 - Solution

$$\|t_k - \widehat{p}\|^2 = \sum_{j=1}^{K} (t_k(j) - \widehat{p}_j)^2 = (t_k(k) - \widehat{p}_k)^2 = (1 - \widehat{p}_k)^2$$

is a 2-d degree polynomial defined on $[0,1]$ and decreasing on that set. Therefore, $\|t_k - \widehat{p}\|$ is minimal when $\widehat{p}_k$ is maximal.

Exercise 3 - Solution

For all generated points $x_1, ..., x_N$, define $r_i := \|x_i\|$ their distance to the origin. By definition, the median distance m from the closest point to the origin must satisfy $\mathbb{P}(\min_i r_i \geq m) = \frac{1}{2}$. Denote by λ_p the Lebesgue measure in $\mathbb{R}^p$ and $\mathbf{B}(0,a)$ the ball of radius a centered at 0 for the Euclidian distance. By independence of the generation of points,

$$\begin{aligned}\mathbb{P}(\min_i r_i \geq m) = \prod_{i=1}^{N} \mathbb{P}(r_i \geq m) = \prod_{i=1}^{N} (1 - \mathbb{P}(r_i \leq m)) = \prod_{i=1}^{N} (1 - \frac{\lambda_p(\mathbf{B}(0,m))}{\lambda_p(\mathbf{B}(0,1))}) \\ = (1 - \frac{\pi^{p/2} m^p}{\Gamma(p/2+1)} \frac{\Gamma(p/2+1)}{\pi^{p/2} 1^p})^N = (1 - m^p)^N\end{aligned}$$

By solving $(1 - m^p)^N = 1/2$ we find the solution $m = (1 - 2^{-1/N})^{1/p}$.

Exercise 7 - Solution

(a) For linear regression, the estimator can be written as $\widehat{f}(x_0) = x_0^T \widehat{\beta}$. Since $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, we have $\widehat{\beta}_j = \sum_{i=1}^N \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)_{j,i} y_i$. Therefore,

$$\begin{aligned}
\widehat{f}(x_0) &= x_0^T \widehat{\beta} \\
&= \sum_{j=1}^p x_{0,j}\widehat{\beta}_j = \sum_{j=1}^p x_{0,j} \sum_{i=1}^N \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)_{j,i} y_i \\
&= \sum_{i=1}^N \sum_{j=1}^p x_{0,j} \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)_{j,i} y_i \\
&=: \sum_{i=1}^N l_i(x_0, \mathcal{X}) y_i,
\end{aligned}$$

where the weights are defined as $l_i(x_0, \mathcal{X}) = \sum_{j=1}^p x_{0,j} \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)_{j,i}$ for $i = 1, ..., N$.
We use the same technique for k-NN. For a fixed positive integer k, the estimator is :

$$\begin{aligned}
\widehat{f}(x_0) &= \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i \\
&= \sum_{i=1}^N \frac{1}{k} \mathbb{1}_{\{x_i \in N_k(x_0)\}} y_i \\
&=: \sum_{i=1}^N l_i(x_0, \mathcal{X}) y_i
\end{aligned}$$

where the weights are defined as $l_i(x_0, \mathcal{X}) = \frac{1}{k}\mathbb{1}_{\{x_i \in N_k(x_0)\}}$.

(b) In this case, $\mathcal{Y}$ varies but $\mathcal{X}$, $f(x_0)$, and x_0 are fixed.

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}[(f(x_0) - \widehat{f}(x_0))^2] &= f(x_0)^2 - 2f(x_0)\mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)] + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)^2] \\
&= \left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)]\right)^2 + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)^2] - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)]^2 \\
&= \text{Bias}\left(\widehat{f}(x_0)\right)^2 + \text{Var}_{\mathcal{Y}|\mathcal{X}}\left(\widehat{f}(x_0)\right)
\end{aligned}$$

Let now apply this to our framework. Since $\widehat{f}(x_0)$ depends on x_0 and $\mathcal{X}$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)] &= \sum_{i=1}^N l_i(x_0, \mathcal{X}) \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[y_i] \\
&= \sum_{i=1}^N l_i(x_0, \mathcal{X}) \left(f(x_i) + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\epsilon_i]\right) \\
&= \sum_{i=1}^N l_i(x_0, \mathcal{X}) f(x_i)
\end{aligned}$$

whereas

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)^2] &= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\sum_{i=1}^N l_i^2(x_0,\mathcal{X})y_i^2 + \sum_{i\neq j} l_i(x_0,\mathcal{X})l_j(x_0,\mathcal{X})y_iy_j] \\
&= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\sum_{i=1}^N l_i^2(x_0,\mathcal{X})(f(x_i)+\epsilon_i)^2 + \sum_{i\neq j} l_i(x_0,\mathcal{X})l_j(x_0,\mathcal{X})(f(x_i)+\epsilon_i)(f(x_j)+\epsilon_j)] \\
&= \sum_{i=1}^N l_i^2(x_0,\mathcal{X})(f(x_i)^2+\sigma^2) + \sum_{i\neq j} l_i(x_0,\mathcal{X})l_j(x_0,\mathcal{X})f(x_i)f(x_j) \\
&= \sum_{i=1}^N l_i^2(x_0,\mathcal{X})\sigma^2 + \sum_{i,j} l_i(x_0,\mathcal{X})l_j(x_0,\mathcal{X})f(x_i)f(x_j)
\end{aligned}
$$

Thus,

$$
\text{Bias}\left(\widehat{f}(x_0)\right)^2 = (f(x_0) - \sum_{i=1}^N l_i(x_0,\mathcal{X})f(x_i))^2
$$

and

$$
\text{Var}_{\mathcal{Y}|\mathcal{X}}\left(\widehat{f}(x_0)\right) = \sum_{i=1}^N l_i^2(x_0,\mathcal{X})\sigma^2.
$$

(c) The calculation is the same, except that $\mathcal{Y}$ and $\mathcal{X}$ vary whereas $f(x_0)$ and x_0 are fixed.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Y},\mathcal{X}}[(f(x_0)-\widehat{f}(x_0))^2] &= f(x_0)^2 - 2f(x_0)\mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)] + \mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)^2] \\
&= \left(f(x_0) - \mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)]\right)^2 + \mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)^2] - \mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)]^2 \\
&= \text{Bias}\left(\widehat{f}(x_0)\right)^2 + \text{Var}_{\mathcal{Y},\mathcal{X}}\left(\widehat{f}(x_0)\right)
\end{aligned}
$$

Noticing that $\mathbb{E}_{\mathcal{Y},\mathcal{X}}[V(X)] = \int \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[V(X)]dX$ with X being a given training set and based on the assumption given in the problem,

$$
\mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)] = \int_{\mathbb{R}^N} \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)]h(x_1)...h(x_N)dx_1...dx_N
$$

and

$$
\mathbb{E}_{\mathcal{Y},\mathcal{X}}[\widehat{f}(x_0)^2] = \int_{\mathbb{R}^N} \mathbb{E}_{\mathcal{Y}|\mathcal{X}}[\widehat{f}(x_0)^2]h(x_1)...h(x_N)dx_1...dx_N.
$$

(d) Not clear