Machine Learning 1 - Exercise 7

Fränz Beckius (374057)
Ivan David Aranzales Acero (399364)
Janek Tichy (584200)
Jeremias Eichelbaum (358685)
Alessandro Schneider (364988)

December 2, 2018

1 Bias and Variance of Mean Estimators

1.a

$$\begin{aligned}
Bias(\hat{\mu}) &= E[(\frac{1}{N}\sum_i^N x_i) - \mu] && (1)\\
&= \frac{1}{N}(\sum_i^N E[x_i]) - \mu && (2)\\
&= \frac{N\mu}{N} - \mu && (3)\\
&= 0 && (4)
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\mu}) &= E[((\frac{1}{N}\sum_i^N x_i) - E[\hat{\mu}])^2] && (5)\\
&= E[((\frac{1}{N}\sum_i^N x_i) - \mu)^2] && (6)\\
&= E[((\frac{1}{N}\sum_i^N x_i)^2 - 2\mu(\frac{1}{N}\sum_i^N x_i) + \mu^2] && (7)\\
&= \mu^2 - 2\mu^2 + \mu^2 && (8)\\
&= 0 && (9)
\end{aligned}$$

$$\begin{aligned}
MSE(\hat{\mu}) &= Bias(\hat{\mu})^2 + Var(\hat{\mu}) && (10)\\
&= 0^2 + 0 = 0 && (11)
\end{aligned}$$

1.b

$$
\begin{aligned}
Bias(\hat{\mu}) &= E[0-\mu] && (12)\\
&= -\mu && (13)\\
& && (14)
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\mu}) &= E[(0-0)^2] && (15)\\
&= 0 && (16)\\
& && (17)
\end{aligned}
$$

$$
\begin{aligned}
MSE(\hat{\mu}) &= Bias(\hat{\mu})^2 + Var(\hat{\mu}) && (18)\\
&= (-\mu)^2 + 0 = \mu^2 && (19)
\end{aligned}
$$

2 Bias-Variance Decomposition for Regression

2.a

We prove that $Error(\hat{f}(x)) = Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) = E[(\hat{f}(x) - f(x))^2]$

$$
\begin{aligned}
Error(\hat{f}(x)) &= Bias(\hat{f}(x))^2 + Var(\hat{f}(x)) && (20)\\
&= (E[\hat{f}(x) - f(x)])^2 + E[(\hat{f}(x) - E[\hat{f}(x)])^2] && (21)\\
&= (E[\hat{f}(x)])^2 + f(x)^2 - 2E[\hat{f}(x)]f(x) + E[\hat{f}(x)^2] - (E[\hat{f}(x)])^2 && (22)\\
&= E[\hat{f}(x)^2] - 2E[\hat{f}(x)]f(x) + f(x)^2 && (23)\\
&= E[(\hat{f}(x) - f(x))^2] && (24)\\
&= Error(\hat{f}(x)) && (25)
\end{aligned}
$$

3 Bias-Variance Decomposition for Classification

3.a

We rewrite the optimization problem as follows

$$
\begin{aligned}
min_R \quad E[D_{KL}(R||\hat{P}_i)] &= min_R \quad E[\sum_{1\leq i\leq c} R_i log(\frac{R_i}{\hat{P}_i})] && (26)\\
&= min_R \quad \sum_{1\leq i\leq c} R_i E[log(\frac{R_i}{\hat{P}_i}))] && (27)\\
&= min_R \quad \sum_{1\leq i\leq c} R_i (log(R_i) - E[log(\hat{P}_i)])] && (28)
\end{aligned}
$$

and then derive the function by R which is the sum of the partial derivatives R_i.

$$\frac{d}{dR_i} = log(R_i) + 1 - E[log(\hat{P}_i)] = 0 \tag{29}$$

$$\implies R_i = exp(E[log(\hat{P}_i)] - 1) \tag{30}$$

$$\implies R_i = \frac{exp(E[log(\hat{P}_i)])}{exp(1)} \tag{31}$$

R is defined as a probability distribution, because of that $\sum_i R_i = 1$ must be true. To assure this property we multiply each R_i with a scaling factor F, which we define as:

$$\sum_i FR_i = 1 \tag{32}$$

$$\implies \sum_i F\frac{exp(E[log(\hat{P}_i)])}{exp(1)} = 1 \tag{33}$$

$$F = \frac{1}{\frac{\sum_i exp(E[log(\hat{P}_i)])}{exp(1)}} \tag{34}$$

Which leads to following definition of R_i

$$R_i = \frac{1}{\frac{\sum_j exp(E[log(\hat{P}_j)])}{exp(1)}} \frac{exp(E[log(\hat{P}_i)])}{exp(1)} \tag{35}$$

$$= \frac{exp(E[log(\hat{P}_i)])}{\sum_j exp(E[log(\hat{P}_j)])} \tag{36}$$

3.b