Homework 4: Solutions
600.482/682 Deep Learning
Fall 2018

Kaushik Srinivasan

November 9, 2018

With collaboration with Nathan Vallapureddy

1. For the first approach which is thhe majority vote approach, we count the number of values of $p < 0.5$ - in this case we have 6 p estimates. That means that there is a higher probability of **green** than red, as the probability the class is red is low.

   For the second approach, we take the mean of the samples - in this case the mean is 0.445, which is less than 0.5. Hence p(class is red) $<$ 0.5 – hence **green** is the predominant class.

2. (a)

$$\begin{aligned}
Cov(\hat{w}^{l2}) =& Cov\Big((X^TX + kI)^{-1}\big(X^Ty^{(\text{train})}\big)\Big) \\
=& (X^TX + kI)^{-1}Cov(X^Ty^{(\text{train})})((X^TX + kI)^{-1})^T \\
=& (X^TX + kI)^{-1}X^TCov(y^{(\text{train})})X((X^TX + kI)^{-1})^T \\
Cov(y^{(\text{train})}) =& \sigma^2 I \quad \text{as values are independent} \\
Cov(\hat{w}^{l2}) =& \sigma^2(X^TX + kI)^{-1}X^TX((X^TX + kI)^{-1})^T
\end{aligned}$$

   (b)

$$X^TX = UDU^T$$

$$\begin{gathered}
Cov(\hat{w}^{l2}) = \sigma^2(UDU^T + kI)^{-1}UDU^T((UDU^T + kI)^{-1})^T \\
\text{Let's concentrate on } (UDU^T + kI)^{-1} \qquad U^TU = I \\
= (UDU^T + kI)^{-1} \\
= [U(DU^T + kU^{-1}I)]^{-1} \\
= [U(D + kU^{-1}I(U^T)^{-1})U^T]^{-1} \\
= U[D + kI]^{-1}U^T \\
\textbf{Now put together} \\
\sigma^2(U[D + kI]^{-1}U^T)UDU^T(U[D + kI]^{-1}U^T)^T \\
= \sigma^2(U[D + kI]^{-1}U^T)UDU^TU[[D + kI]^{-1}]^TU^T \\
= \sigma^2(U[D + kI]^{-1})D([[D + kI]^{-1}]^TU^T) \\
= \sigma^2U[D + kI]^{-1}D[D + kI]^{-1}U^T
\end{gathered}$$

(c) Now let us simplify element wise

$$\begin{aligned}
P =& U[D+kI]^{-1}D[D+kI]^{-1} \\
P_{ij} =& \sum_{l=1}^{n} U_{il}([D+kI]^{-1}D[D+kI]^{-1})_{lj} \\
=& U_{ij}([D+kI]^{-1}D[D+kI]^{-1})_{jj} \\
=& \frac{U_{ij}D_{jj}}{(D_{jj}+k)^2}
\end{aligned}$$

Since P will be a diagonal, we can let $R = \sigma^2 PU^T$. We will need to ensure $k = \{x : x \neq D_{jj}, \forall j\}$. Hence we will get:

$$\begin{aligned}
R_{ii} &= \sigma^2 \sum_{l=1}^{n} R_{il}U_{li}^T \\
&= \sigma^2 \sum_{l=1}^{n} R_{il}U_{il} \\
&= \sigma^2 \sum_{l=1}^{n} \frac{U_{il}D_{ll}}{(D_{ll}+k)^2}U_{il} \\
&= \sigma^2 \sum_{l=1}^{n} \frac{U_{il}^2 D_{ll}}{(D_{ll}+k)^2}
\end{aligned}$$

(d) Non regularized version

$$\begin{gathered}
Cov(\hat{w}) = Cov((X^TX)^{-1}X^Ty^{(\text{train})}) \\
Cov(\hat{w}) = (X^TX)^{-1}X^TCov(y^{(\text{train})})X((X^TX)^{-1})^T \\
Cov(y^{(\text{train})}) = \sigma^2 I \quad \text{as values are independent} \\
Cov(\hat{w}) = \sigma^2(X^TX)^{-1}X^TX((X^TX)^{-1})^T \\
\text{Let} \quad X^TX = UDU^T \\
\text{Hence} \qquad Cov(\hat{w}) = \sigma^2(UDU^T)^{-1}UDU^T((UDU^T)^{-1})^T \\
= \sigma^2[(UDU^T)^{-1}U]D[((UDU^T)^{-1}U)^T] \\
\text{expand the inverses} \\
= \sigma^2[((U^T)^{-1}D^{-1}U^{-1})U]D[(((U^T)^{-1}D^{-1}U^{-1})U)^T] \\
= \sigma^2(U^T)^{-1}((U^T)^{-1}D^{-1})^T \\
= \sigma^2(U^T)^{-1}(D^{-1})^TU^{-1} \\
= \sigma^2(UDU^T)^{-1} \\
= \sigma^2(X^TX)^{-1} \\
\hat{R}_{ii} = \sigma^2 \sum_{l=1}^{n} \frac{U_{il}^2}{D_{ll}}
\end{gathered}$$

We know that D is a diagonal matrix that is non-negative (we force our SVD in that manner). This regularized R is strictly less than unregularized $\hat{R}$ as we note ($\frac{D_{ll}}{(D_{ll}+k)^2} < \frac{1}{D_{ll}}$). Hence variance of regularized weight vector is less than variance of unregularized weight vector.