10-701/15-781, Machine Learning: Homework 3

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on February 3, 2010

1 Linear regression, and bias-variance trade-off[20pt, Ni Lao]

1.1 Least square regression [4 pt]

Using SVD we can decompose X as $X = UDV^T$, where D is a $p \times p$ diagonal matrix, V is a $p \times p$ unitary matrix, U is a $n \times p$ matrix, which is the first p columns of a unitary matrix. Here we assume that $n \geq p$.

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y && (1)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) && (2)\\
&= \beta + VD^{-1}U^T\epsilon && (3)
\end{aligned}$$

Therefore, $\hat{\beta} \sim \mathcal{N}(\beta, VD^{-2}V^T\sigma^2) \sim \mathcal{N}(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$.

1.2 Ridge regression [4 pt]

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T y && (4)\\
&= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) && (5)\\
&= \beta + (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(-\lambda\beta + \mathbf{X}^T\epsilon) && (6)\\
&= \beta - \lambda V(D^2 + \lambda I)^{-1}V^T\beta + VD(D^2 + \lambda I)^{-1}U^T\epsilon && (7)
\end{aligned}$$

Therefore, $\hat{\beta} \sim \mathcal{N}(\beta - \lambda V(D^2 + \lambda I)^{-1}V^T\beta, VD^2(D^2 + \lambda I)^{-2}V^T\sigma^2)$.

1.3 The bias-variance trade-off [4 pt]

$$\begin{aligned}
e(\lambda) &= \hat{Y}^* - Y^* && (8)\\
&= \mathbf{X}\hat{\beta} - (\mathbf{X}\beta + \epsilon^*) && (9)\\
&= -\lambda UD(D^2 + \lambda I)^{-1}V^T\beta + UD^2(D^2 + \lambda I)^{-1}U^T\epsilon - \epsilon^* && (10)
\end{aligned}$$

1.4 [4 pt]

Since

$$e(\lambda) \sim \mathcal{N}\left(0, U\left(\frac{\lambda D}{D^2 + \lambda I}\right)^2 U^T\alpha^2 + U\left(\frac{D^2}{D^2 + \lambda I}\right)^2 U^T\sigma^2 + \sigma^2 I\right) \tag{11}$$

we have

$$R(\lambda) = E[e(\lambda)^T e(\lambda)] \tag{12}$$

$$= \sum_{i=1..p} \left[\left(\frac{\lambda d_i}{d_i^2 + \lambda} \right)^2 \alpha^2 + \left(\frac{d_i^2}{d_i^2 + \lambda} \right)^2 \sigma^2 \right] + \sum_{i=1..n} \sigma^2, \tag{13}$$

where we define $d_i = D_{i,i}$.

1.5 [4 pt]

$$\frac{\partial R(\lambda)}{\partial \lambda} = 2 \sum_{i=1..p} \frac{\alpha^2 \lambda d_i^4 - \sigma^2 d_i^4}{(d_i^2 + \lambda)^3} \tag{14}$$

It is zero when $\lambda = \sigma^2/\alpha^2$