hw3

Wenhua Bao 2512664
Yue Liu 2803140

July 9, 2020

1 Task1

1.0.1 a

Ridge coefficient is a regulizer. To regularize the psesudo-inverse, we use it. Otherwise, we can't get a weight cofficient, when the matrix $\phi\phi^T$ is inversiable.

1.0.2 a

squared error loss function

$$J(\theta)=\sum_{i=1}^{m}\left(f_{\theta}\left(x^{(i)}\right)-y^{(i)}\right)^{2}$$

add ridge coefficientλ

$$J(\theta)=\frac{1}{2}\left[\sum_{i=1}^{n}\left(f_{\theta}(x)^{(i)}-y^{(i)}\right)^{2}+\sum_{j=1}^{n}\lambda\theta_{j}^{2}\right]$$

$$\begin{aligned}
J(\theta) &= \frac{1}{2}(X\theta-Y)^{\top}(X\theta-Y)+\lambda\theta^{\top}\theta \\
&= \frac{1}{2}(X\theta-Y)^{\top}(X\theta-Y)+\lambda\theta^{\top}\theta \\
&= \frac{1}{2}\left(\theta^{\top}X^{\top}X\theta-\theta^{\top}X^{\top}Y-Y^{\top}X\theta+Y^{\top}Y+\lambda\theta^{\top}\theta\right.
\end{aligned}$$

$$\begin{aligned}
&\tfrac{\partial J(\theta)}{\partial\theta}=X^{\top}X\theta-X^{\top}Y+\lambda\theta=0 \\
&\theta=\left(X^{\top}X+\lambda I\right)^{-1}X^{\top}Y
\end{aligned}$$

1.0.3 3

degree 1:0.3843532873748282

2 Task2

2.1 a

Generative models learn prior distribution to derive posterior distribution then get classification, but discriminative models learn posterior distribution to get classification. discriminative models: Logistical Regression generative models: Bayesian Analysis Discriminative models is easier to learn, because it doesn't need to learn conditional probability and directly to learn posterior.

2.2 2b

there's 19 samples being misclassified.