Jiaqi Li Net ID: jl9555 University ID: N14088502
Xiahao Zhang NetID:xz2456 University ID:N14493597

CS6923 Homework Assignment 4*

1、

(a): Accorting to the definition: $E_{ridge} = \sum_{i=1}^{N}(\mathrm{y}^{(\mathrm{i})} - \mathbf{w}^T\mathbf{x}^{(\mathrm{i})})^2 + \lambda\sum_{j=1}^{d} w_j^2$

$$= \|\mathbf{X}\mathbf{w} - y\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

$$= (\mathbf{X}\mathbf{w} - \mathrm{y})^T(\mathbf{X}\mathbf{w} - \mathrm{y}) + \lambda\mathbf{w}^T\mathbf{w}$$

So the gradient is $\nabla E_{\mathrm{ridge}} = 2(\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}) + 2\lambda\mathbf{I}\mathbf{w}$

(b): If we should set the gradient to 0: $\nabla E_{ridge} = 0$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{I}^{(d+1)\times(d+1)} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \ldots & 1 \end{bmatrix}$

2、

(a): Let's make clear the definitions first. The size of D_{val} is K. Let μ be the true error of g_h, v is the average error on D_{val}. For every single input $x^{(i)}$ in validation set, we have an expected value $y^{(i)}$. Since we are doing a binary classification, [a, b] is equivalent to [0, 1]. At last, according to the Hoeffding inequality, we have:

$$P[|v - \mu| > \varepsilon] \le 2e^{-\frac{2\varepsilon^2 K}{(\mathrm{b}-a)^2}}$$

When ε=0.1, K=100, b=1, a=0 we have:

$$P[|v - \mu| > 0.1] \le 2e^{-\frac{2\times(0.1)^2\times 100}{(1)^2}} \approx 0.27$$

So we have a probability of 1 - 0.27 = 0.73 that the true error of h is within 0.1 of its average error on D_{val}.

(b): Like what we did in part (a), we have:

$$P[|v - \mu| > 0.1] \le 2e^{-\frac{2\times(0.1)^2\times 200}{(1)^2}} \approx 0.037$$

So we have a probability of 1 - 0.037 = 0.963 that the true error of h is within 0.1 of its average error on D_{val}.

(c): We want to calculate the probability that the differences between the average error and the true error for both models are less than 0.1.

$$P[|v-\mu| \text{ of any one of the models we choose is greater than } 0.1]$$

$$\leq P[|v_1-\mu_1| \text{ of model}_1 \text{ is greater than } 0.1] + P[|v_2-\mu_2| \text{ of model}_2 \text{ is greater than } 0.1]$$

$$\leq 2e^{-\frac{2\times(0.1)^2\times 100}{1^2}} + 2e^{-\frac{2\times(0.1)^2\times 100}{1^2}} \approx 0.54$$

Thus we have a probability of 1 - 0.54 = 0.46 that the model selected is within 0.1 of its average error on D_{val}.

3、

In this question, since we assume that w_i for i = 1…d are independent, identically and distributed according to a gaussian distribution, $wi \sim N(0,\rho^2)$ for i = 1...d. Then we have:

$$P(\mathbf{w}) = (\frac{1}{\sqrt{2\pi\rho^2}})^d \exp(\sum_{j=1}^{d} \frac{-(w^{(i)})^2}{2\rho^2})$$

Thus, we will have a new $\mathbf{w}_{MAP}$:

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(\mathrm{y}^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) p(\mathbf{w})$$

$$= \prod_{\mathrm{i}=1}^{N} [\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(\mathrm{y}^{(i)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)})^2}{2\sigma^2}] \times (\frac{1}{\sqrt{2\pi\rho^2}})^d \exp(\sum_{j=1}^{d} \frac{-(w^{(i)})^2}{2\rho^2})$$

Note that maximizing this value is the same as maximizing $\ell(\mathbf{w}) = \log \mathrm{L}(\mathbf{w})$:

$$\ell(\mathrm{w}) = \log(\prod_{\mathrm{i}=1}^{N} [\frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(\mathrm{y}^{(i)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)})^2}{2\sigma^2}] \times (\frac{1}{\sqrt{2\pi\rho^2}})^d \exp(\sum_{j=1}^{d} \frac{-(w^{(i)})^2}{2\rho^2}))$$

$$= \mathrm{N} \times \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2} \sum_{i=1}^{N} -(\mathrm{y}^{(i)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)})^2 + d \times \log \frac{1}{\sqrt{2\pi\rho^2}} + \frac{1}{2\rho^2} \sum_{j=1}^{d} -(w^{(i)})^2$$

This is the same as minimizing the following equation:

$$\frac{1}{2\sigma^2} [\sum_{i=1}^{N} (\mathrm{y}^{(i)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)})^2 + \frac{\sigma^2}{\rho^2} \sum_{j=1}^{d} (w^{(i)})^2]$$

Compared to the Ridge Regression L_2 regularization, the ridge regression estimate is the MAP estimate, where **w** has the prior distribution as described in the question and

$$\lambda = \frac{\sigma^2}{\rho^2} \quad .$$