Machine Learning Foundation HW4

B04902004 王佑安

1.

QUIZ

作業四

20 questions

Your Score 200/200 points (100%)

We keep your highest score.

View Latest Submission

Take it again

2. $\nabla E_{aug}(\mathbf{w}) = \nabla E_{in}(\mathbf{w}) + \nabla(\frac{\lambda}{N}\mathbf{w}^T\mathbf{w})$
   $= \nabla E_{in}(\mathbf{w}) - \frac{2\lambda}{N}\mathbf{w}$
   $\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta\nabla E_{aug}(\mathbf{w}(t))$
   $\Rightarrow \mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta(\nabla E_{in}(\mathbf{w}(t)) + \frac{2\lambda}{N}\mathbf{w}(t))$
   $\Rightarrow \mathbf{w}(t+1) \leftarrow (1 - \frac{2\eta\lambda}{N})\mathbf{w}(t) + \eta\nabla E_{in}(\mathbf{w}(t))$

3. If $\mathbf{w}^T\mathbf{w} = C \geq ||\mathbf{w}_{lin}|| \Rightarrow$ optimal $\mathbf{w}_{REG} = \mathbf{w}_{lin}$
   If $\mathbf{w}^T\mathbf{w} = C < ||\mathbf{w}_{lin}|| \Rightarrow ||\mathbf{w}_{REG}|| \leq C < ||\mathbf{w}_{lin}||$
   $\Rightarrow ||\mathbf{w}_{REG}|| \leq ||\mathbf{w}_{lin}||$

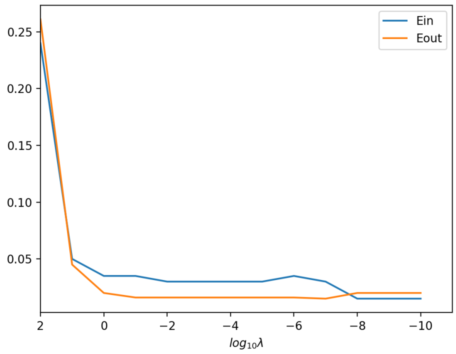
4. constant:
   Leave $(1,0) \Rightarrow h_0(x) = \frac{1}{2}, e_1 = \frac{1}{2}$
   Leave $(-1,0) \Rightarrow h_0(x) = \frac{1}{2}, e_2 = \frac{1}{2}$
   Leave $(\rho,1) \Rightarrow h_0(x) = 0, e_3 = 1$
   $E_{loocv}(constant) = \frac{1}{3}(\frac{1}{2}^2 + \frac{1}{2}^2 + 1^2)$

   linear:
   Leave $(1,0) \Rightarrow h_1(x) = \frac{x+1}{\rho+1}, e_1 = \frac{2}{\rho+1}$
   Leave $(-1,0) \Rightarrow h_1(x) = \frac{x-1}{\rho-1}, e_2 = \frac{2}{\rho-1}$
   Leave $(\rho,1) \Rightarrow h_1(x) = 0, e_3 = 1$
   $E_{loocv}(linear) = \frac{1}{3}((\frac{2}{\rho+1})^2 + (\frac{2}{\rho-1})^2 + 1^2)$

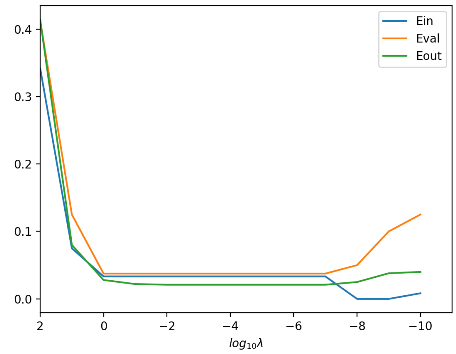
   $\frac{1}{3}(\frac{1}{2}^2 + \frac{1}{2}^2 + 1^2) = \frac{1}{3}((\frac{2}{\rho+1})^2 + (\frac{2}{\rho-1})^2 + 1^2)$
   $\Rightarrow \rho = \sqrt{9 + 4\sqrt{6}}$

5. Let $\hat{\mathbf{X}} = [x_1 ... x_N, \tilde{x}_1 ... \tilde{x}_K]^T, \hat{\mathbf{y}} = [y_1 ... y_N, \tilde{y}_1 ... \tilde{y}_K]^T$
   $\mathbf{w} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}^T\hat{\mathbf{y}})$
   $= (\sum_{n=1}^{N} x_n^T x_n + \sum_{k=1}^{K} x_k^T x_k)^{-1}(\sum_{n=1}^{N} x_n y_n + \sum_{k=1}^{K} x_k y_k)$
   $= (\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$

6. Optimal solution:$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
   $\Rightarrow \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \lambda\mathbf{I}$, $\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} = 0$
   $\Rightarrow \tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}$, $\tilde{\mathbf{y}} = 0$

7. 

   λ 很大時受到的限制太大，所以連 training data 都沒辦法 fit，而 λ 很小時發現 Eout 比 Ein 高，可以推論是因為 Regularization 不夠導致 overfitting。

8. 

   Validation 跟 Test data 都是 train data 裡觀察不到的，因此 λ 太小都會因為 overfitting 導致 error 上升，驗證了 validation 可以判斷是否 overfitting 的性質。

9. (a) Leave positive instance:
   $\mathcal{A}_{major}$ always predict negative, $E_{loocv} = 1$
   $\mathcal{A}_{minor}$ always predict positive, $E_{loocv} = 0$
   Leave negative instance:
   $\mathcal{A}_{major}$ always predict positive, $E_{loocv} = 1$
   $\mathcal{A}_{minor}$ always predict negative, $E_{loocv} = 0$
   $\Rightarrow \mathcal{A}_{minor}$ is better.

   (b) Suppose $\{y_n\}_{n=1}^N$ has mean μ.
   $E_{loocv}(\mathcal{A}_{avg}) = \frac{1}{N}\sum_{n=1}^{N}(\frac{N\mu - y_n}{N-1} - y_n)^2 = \frac{1}{N}\sum_{n=1}^{N}(\frac{N\mu - y_n - Ny_n + y_n}{N-1})^2 = \frac{N}{(N-1)^2}\sum_{n=1}^{N}(\mu - y_n)^2$
   $= \frac{N^2}{(N-1)^2}Var(\{y_n\}_{n=1}^N)$