EE 511 HW #1

P1. MLE

① $P(\underline{E}, \underline{H}; \lambda, p) = \prod_{i=1}^{n} \left(\frac{\lambda^{E_i}}{E_i!} e^{-\lambda} \cdot p^{H_i} (1-p)^{E_i - H_i} \right)$

$\ln P(\underline{E}, \underline{H}; \lambda, p) = \sum_{i=1}^{n} \ln \left[\frac{\lambda^{E_i}}{(E_i)!} e^{-\lambda} \cdot p^{H_i} (1-p)^{E_i - H_i} \right]$

$= \sum_{i=1}^{n} \left[E_i \ln\lambda + (-\lambda) - \ln\frac{1}{(E_i)!} + H_i \ln p + (E_i - H_i) \ln(1-p) \right]$

② $\frac{\partial \ln P(E, H; \lambda, p)}{\partial \lambda} = \sum_{i=1}^{n} \left[\frac{E_i}{\lambda} - 1 \right] = 0 \Rightarrow \frac{\Sigma E_i}{\lambda} = n \Rightarrow \hat{\lambda} = \frac{\Sigma E_i}{n}$

$\frac{\partial \ln P(E, H; \lambda, p)}{\partial p} = \sum_{i=1}^{n} \left[H_i \cdot \frac{1}{p} - (E_i - H_i) \frac{1}{1-p} \right] = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} H_i}{\sum_{i=1}^{n} E_i}$

③ $\Sigma E_i = 8 + 9 + 6 + 4 + 1 + 5 + 2 + 12 + 9 + 7 = 63$

$\Sigma H_i = 5 + 6 + 4 + 3 + 0 + 5 + 2 + 9 + 8 + 6 = 48$

$\therefore \hat{\lambda}_{MLE} = \frac{63}{10} = 6.3 \qquad \hat{p}_{MLE} = \frac{48}{63} = \frac{16}{21}$

P2: $\sum_{i=1}^{n}(y_i - X\hat{\omega})^2 + \lambda\|w\|_1 \approx \sum_{i=1}^{n}(y_i - X\hat{\omega})^2$

2.1 (a) Small.

With λ approaching to 0, we ignore the regularization part.
The error on the training set should be small because of optimizing MLE

(b) Large.

If optimizing training set only by ordinary least square method, the error on the testing set should be large

(c) Large.

With a small λ, it's nonregularized.

(d) Large.

λ is too small to force relatively small weights to real zero.

Overestimate of λ

(a) Large

With such a large λ, too many weights become zero. The error on the training set became large.

(b) Large

Overfitting means that the model can't generalize to test set.

(c) Small.

λ is so large that most weights are forced to be zero.

(d) Small

Too many weights are forced to be real zero for overestimated λ.

2.2

1.

$$\lambda \|w_i\| = \lambda \sum_{i=1}^{d} |w_i|$$

$$\frac{\partial E_R}{\partial w_i} = \begin{cases} \lambda & w_i > 0 \\ -\lambda & w_i < 0 \end{cases}$$

2.

$$\lambda \|w\|_2^2 = \lambda \sum_{i=1}^{d} (w_i)^2$$

$$\frac{\partial E_R}{\partial w_i} = 2\lambda w_i$$

3. If $w_i \neq \pm\frac{1}{2}$, their behaviors differ.

When λ is very large:

For Lasso, such a large λ and linear penalty pushes more weights to zero. It allows for a type of feature selection.

For Ridge, relatively small weights may trade off such a large λ.