EE511 Machine Learning, Winter 2018: Homework 1 Solutions

Due: Wednesday, Janunary 17^{th}, beginning of class

1 MLE [25 points]

1. *(8 points)* the joint log-likelihood function of E and H given λ and p.

$$Pr\{E=(E_1,...,E_n),H=(H_1,...,H_n)|\lambda,p\}=\prod_{i=1}^{n}\frac{\lambda^{E_i}}{E_i!}e^{-\lambda}\binom{E_i}{H_i}p^{H_i}(1-p)^{(E_i-H_i)}$$

$$\ln Pr\{E=(E_1,...,E_n),H=(H_1,...,H_n)|\lambda,p\}=\ln\prod_{i=1}^{n}\frac{\lambda^{E_i}}{E_i!}e^{-\lambda}\binom{E_i}{H_i}p^{H_i}(1-p)^{(E_i-H_i)}=$$

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

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

2. *(12 points)* the MLE for λ and p in the general case.

$$\frac{\partial}{\partial\lambda}\sum_{i=1}^{n}\ln[\frac{\lambda^{E_i}}{E_i!}]+\ln[e^{-\lambda}]+\cancelto{0}{\ln[\binom{E_i}{H_i}]}+\cancelto{0}{\ln[p^{H_i}]}+\cancelto{0}{\ln[(1-p)^{(E_i-H_i)}]}=$$

$$\sum_{i=1}^{n}\frac{E_i}{\lambda}-1=0\implies\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}E_i$$

$$\frac{\partial}{\partial p}\sum_{i=1}^{n}\cancelto{0}{\ln[\frac{\lambda^{E_i}}{E_i!}]}+\cancelto{0}{\ln[e^{-\lambda}]}+\cancelto{0}{\ln[\binom{E_i}{H_i}]}+\ln[p^{H_i}]+\ln[(1-p)^{(E_i-H_i)}]=$$

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

$$\sum_{i=1}^{n}\frac{H_i}{p}-\frac{(E_i-H_i)}{(1-p)}=0\implies\hat{p}=\frac{\sum_{i=1}^{n}H_i}{\sum_{i=1}^{n}E_i}$$

3. *(5 points)* the MLE for λ and p using the observed values of E and H.
With $E=(8,9,6,4,1,5,2,12,9,7)$ and $H=(5,6,4,3,0,5,2,9,8,6)$,

$$\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}E_i=6.3$$

$$\hat{p}=\frac{\sum_{i=1}^{n}H_i}{\sum_{i=1}^{n}E_i}=0.7619$$

2 Regularization Constants [15 points]

2.1 LASSO Regression

1. Suppose our λ is much too small; that is,

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

   How will this affect the magnitude of:

   (a) *(1 point)* The error on the training set? **decreases**

   (b) *(1 point)* The error on the testing set? **increases**

   (c) *(1 point)* $\hat{w}$? **increases**

   (d) *(1 point)* The number of nonzero elements of $\hat{w}$? **increases**

2. Suppose instead that we overestimated on our selection of λ. What do we expect to be the magnitude of:

   (a) *(1 point)* The error on the training set? **increases**

   (b) *(1 point)* The error on the testing set? **increases**

   (c) *(1 point)* $\hat{w}$? **decreases**

   (d) *(1 point)* The number of nonzero elements of $\hat{w}$? **decreases**

2.2 Ridge Regression

1. *(2 points)*

$$\frac{\partial}{\partial \hat{w}_i}\lambda \left\|\hat{w}\right\|_1 = \begin{cases} -\lambda \text{ if } \hat{w}_i < 0 \\ \lambda \text{ if } \hat{w}_i > 0 \end{cases}$$

2. *(2 point)*

$$\frac{\partial}{\partial \hat{w}_i}\lambda \left\|\hat{w}\right\|_2^2 = 2\lambda w_i$$

3. *(3 points)*
   For large λ, L_1 regulariaztion leads to sparse w_i (many zero weights) whereas L_2 regularization leads to small w_i