Machine Learning	**Homework 2**
CSE 546 Fall 2017	**October 21, 2017**
Swati Padmanabhan	

NOTE: Collaborated with Robbie Weber.

1. Least Squares. Dimensions: $x \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ and $\varepsilon \in \mathbb{R}^n$. Therefore, $XX^T \in \mathbb{R}^{n \times n}$, $X^TX \in \mathbb{R}^{d \times d}$.

   a. The given expression is $\hat{w} = (X^TX)^{-1}X^Ty$. Substitute the given expression for y,

   $$\hat{w} = (X^TX)^{-1}X^T(Xw^* + \varepsilon) = w^* + (X^TX)^{-1}X^T\varepsilon. \tag{1.1}$$

   b. Define

   $$U = (X^TX)^{-1}X^T. \tag{1.2}$$

   Let U_i represent the ith row of U. Then from Equation (1.1):

   $$\begin{aligned} \mathop{\mathbf{E}}_{\varepsilon} \hat{w} &= w^* + [\mathop{\mathbf{E}}_{\varepsilon} U_i^T\varepsilon] \\ &= w^*, \end{aligned} \tag{1.3}$$

   where the last equality is because $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for all $i = 1, \ldots, n$, so $\mathbf{E}_\varepsilon U_i^T\varepsilon = \sum_{j=1}^n U_{ij} \mathbf{E}_{\varepsilon_j} \varepsilon_j = 0$. Again from Equation (1.1)

   $$\begin{aligned} \mathop{\mathbf{E}}_{\varepsilon} \|\hat{w} - w^*\|_2^2 &= \mathop{\mathbf{E}}_{\varepsilon} \|U\varepsilon\|_2^2 \\ &= \mathbf{E}\, \varepsilon^T U^T U \varepsilon \\ &= \mathbf{E}\, \operatorname{Tr} \varepsilon^T U^T U \varepsilon \\ &= \mathbf{E}\, \operatorname{Tr} \varepsilon\varepsilon^T U^T U \\ &= \operatorname{Tr} \mathbf{E}\, \varepsilon\varepsilon^T U^T U \\ &= \operatorname{Tr} \sigma^2 U^T U \\ &= \sigma^2 \operatorname{Tr} U^T U \\ &= \sigma^2 \operatorname{Tr}\big(X^TX\big)^{-1}. \end{aligned} \tag{1.4}$$

   c. For the first expectation:

   $$\begin{aligned} \mathop{\mathbf{E}}_{\tilde{\varepsilon}}(\tilde{y}) &= \mathop{\mathbf{E}}_{\tilde{\varepsilon}}(Xw^* + \tilde{\varepsilon}) \\ &= Xw^* + \mathop{\mathbf{E}}_{\tilde{\varepsilon}} \tilde{\varepsilon} = Xw^*. \end{aligned} \tag{1.5}$$

   where the last expression follows from substituting $\hat{w}$ from Equation (1.1) and Defini-

tion 1.2. Next, to make calculations simpler, we observe that:

$$\begin{aligned}\tilde{y}_i - x_i^T\hat{w} &= (x_i^T w^* + \tilde{\varepsilon}_i) - x_i^T\hat{w} \\ &= x_i^T(w^* - \hat{w}) + \tilde{\varepsilon}_i \\ &= -x_i^T U\varepsilon + \tilde{\varepsilon}_i. \\ \therefore \sum_{i=1}^n (\tilde{y}_i - x_i^T\hat{w})^2 &= \sum_{i=1}^n (-x_i^T U\varepsilon + \tilde{\varepsilon}_i)^2 \\ &= \|XU\varepsilon - \tilde{\varepsilon}\|_2^2.\end{aligned}$$

We first compute the inner expectation. Let $M = XU$.

$$\begin{aligned}\mathop{\mathbf{E}}_{\varepsilon} \sum_{i=1}^n (\tilde{y}_i - x_i^T\hat{w})^2 &= \mathop{\mathbf{E}}_{\varepsilon} \|XU\varepsilon - \tilde{\varepsilon}\|_2^2 \\ &= \mathop{\mathbf{E}}_{\varepsilon} \|M\varepsilon\|_2^2 + \mathop{\mathbf{E}}_{\varepsilon} \|\tilde{\varepsilon}\|_2^2 \\ &= \mathop{\mathbf{E}}_{\varepsilon} \varepsilon^T M^T M\varepsilon + \|\tilde{\varepsilon}\|_2^2 \\ &= \mathop{\mathbf{E}}_{\varepsilon} \operatorname{Tr} \varepsilon^T M^T M\varepsilon + \|\tilde{\varepsilon}\|_2^2 \\ &= \operatorname{Tr} \mathop{\mathbf{E}}_{\varepsilon} M^T M\varepsilon\varepsilon^T + \|\tilde{\varepsilon}\|_2^2 \\ &= \operatorname{Tr} M^T M\sigma^2 I + \|\tilde{\varepsilon}\|_2^2 \\ &= \operatorname{Tr}\big((X^TX)^{-1}X^TX(X^TX)^{-1}X^TX\big)\sigma^2 + \|\tilde{\varepsilon}\|_2^2 \\ &= n\sigma^2 + \|\tilde{\varepsilon}\|_2^2.\end{aligned}$$

Taking outer expectation:

$$\begin{aligned}\mathop{\mathbf{E}}_{\tilde{\varepsilon}} \mathop{\mathbf{E}}_{\varepsilon} \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - x_i^T\hat{w})^2 &= \mathop{\mathbf{E}}_{\tilde{\varepsilon}} \frac{1}{n}(n\sigma^2 + \|\tilde{\varepsilon}\|_2^2) \\ &= \sigma^2 + \frac{1}{n} \mathop{\mathbf{E}}_{\tilde{\varepsilon}} \sum_{i=1}^n \tilde{\varepsilon}_i^2 \\ &= 2\sigma^2. \qquad (1.6)\end{aligned}$$

2. a. Set $z_i = \mathbf{1}_{(f(x_i)\neq y_i)}$. Then we have by Hoeffding's inequality,

$$\mathbf{Prob}\left(|\hat{L}(\tilde{f}) - L(\tilde{f})| \leq A\right) \geq 1 - 2\exp\big(-2NA^2\big).$$

Set the RHS to $1 - \delta$. This gives

$$A = \sqrt{\frac{1}{2N}\log\left(\frac{2}{\delta}\right)}.$$

b. Unfortunately yes. The same confidence interval holds, since there was nothing inherently about f or $\tilde{f}$ that was used to compute the conjfidence interval. Just the fact that the empirical loss is computed by averaging over N samples.

c. Ah, okay, this one is better. The proposed function $\hat{f}$ is, by definition, the one that brings the empirical loss closer to the true loss. So yes, the confidence interval *is* changed!

d.

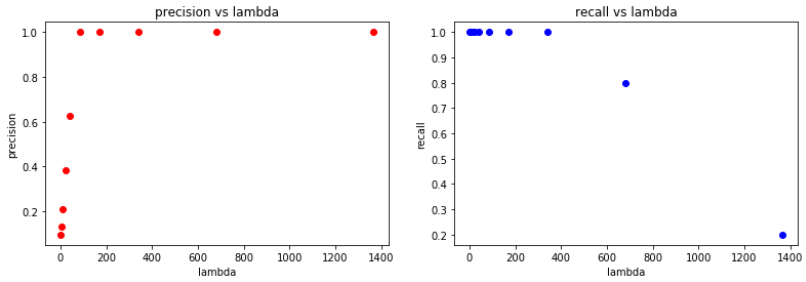

Figure 1: Precision and Recall Curves.

3. The optimal λ for perfect precision and recall is 400. As λ increases, we are forcing far too many elements to become zero (more than are actually zero). That's why the number of *correct* non-zeros decreases (which is exactly recall). The precision, on the other hand, increases because the elements being set to zero are actually the ones that are zero. On synthetic data, this algorithm runs fast and is able to recover the zero patterns perfectly and get very good results on the non-zeros (signs are correct, values are very close to true ones).

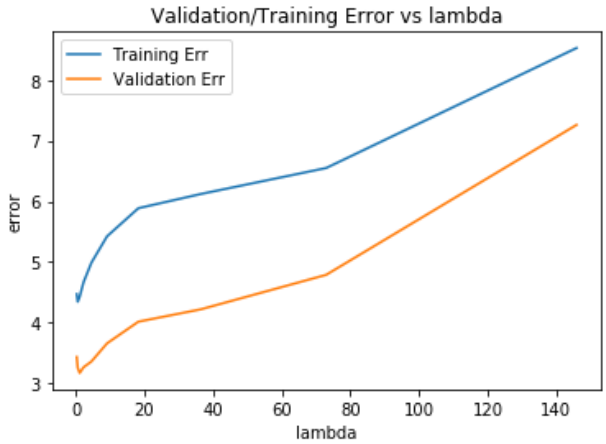

Figure 2: Training and Validation Errors.

4. My best λ is 4.5 and the mean squared test error is 6.4.

5. Using chain rule and some basic algebra:

$$\nabla_w J(w,b) = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\big(-y_i(b+x_i^T w)\big)}{1+\exp\big(-y_i(b+x_i^T w)\big)}(-y_i x_i) + 2\lambda w.$$

Simplifying in terms of μ, this equals

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_i(w,b)-1)y_i x_i + 2\lambda w.$$