EECS 545 MACHINE LEARNING
HOMEWORK 1 WRITE-UP

Yi Yang (davidyy@umich.edu)

Problem 1

Part b) Stochastic Gradient Decent. The trained weight vector $\hat{\mathbf{w}}_{\mathbf{sgd}}$ is:

$$\hat{\mathbf{w}}_{\mathbf{sgd}} = [22.56, -0.89, 1.13, 0.18, 0.64, -2.05, 2.68, 0.29, -3.04, 2.96, -2.4, -1.97, 0.9, -3.95]^T \tag{1}$$

where the first entry **22.56** is the learned bias term. The train error converges to **23.2** and test error converges to **10.68**.The following picture shows the learning process vs. epochs

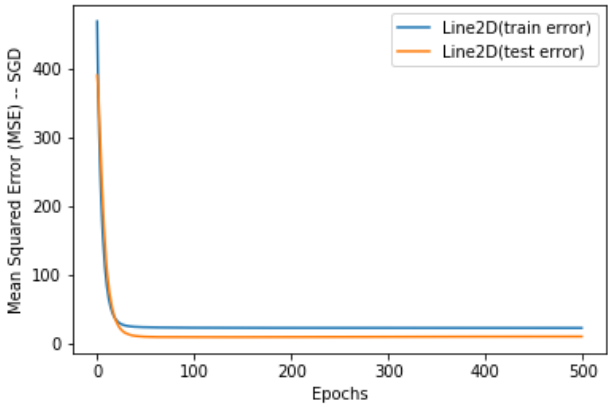

Figure 1. Training and Testing error vs. Epoches

Part c) Batch Gradient Decent. The trained weight vector $\hat{\mathbf{w}}_{\mathbf{bgd}}$ is:

$$\hat{\mathbf{w}}_{\mathbf{bgd}} = [22.56, -0.89, 1.13, 0.18, 0.65, -2.04, 2.69, 0.29, -3.03, 2.96, -2.41, -1.97, 0.91, -3.96]^T \tag{2}$$

where the first entry **22.56** is the learned bias term. The train error converges to **23.2** and test error converges to **10.70**.The following picture shows the learning process vs. epochs

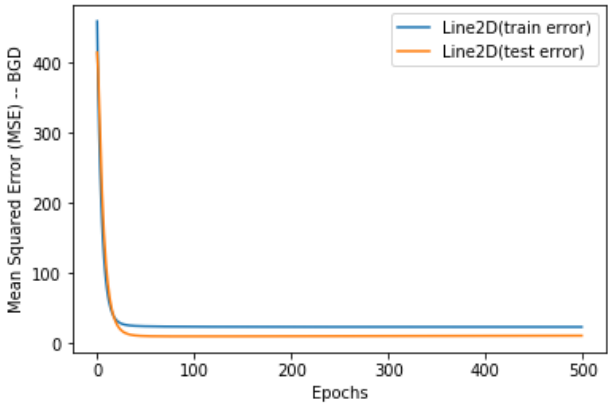

Figure 2. Training and Testing error vs. Epoches

Date: January 29, 2018.

Part d) Closed Form Solution. The trained weight vector $\hat{\mathbf{w}}_{\mathbf{cl}}$ is:

$$\hat{\mathbf{w}}_{\mathbf{cl}} = [22.56, -0.9, 1.14, 0.22, 0.64, -2.04, 2.68, 0.3, -3.02, 3.14, -2.6, -1.97, 0.91, -3.96]^T \tag{3}$$

where the first entry **22.56** is the learned bias term. The train error converges to **23.2** and test error converges to **10.97**. The following picture shows the learning process vs. epochs

Part e) Random Split Consequences. Notice that in previous three questions, we have testing error smaller than training error. It may be due to the bias in the test set. In this case, we created 100 random split data set, and in each of them, we compute the closed form solution, training and test error. The **mean training error** is **12.14**, while the **mean testing error** is **403.6**

Problem 2

Part a) Linear Regression with Different Feature Orders. In this problem, we vary the order of features from $\{0, 1, 2, 3, 4\}$. We used closed form solution to compute the weight vector for order list in the set above:

$$\hat{\mathbf{w}}_{\mathbf{cl0}} = 22.94$$

$$\hat{\mathbf{w}}_{\mathbf{cl1}} = [22.56, -0.9, 1.14, 0.22, 0.64, -2.04, 2.68, 0.3, -3.02, 3.14, -2.6, -1.97, 0.91, -3.96]^T$$

$$\hat{\mathbf{w}}_{\mathbf{cl2}} = [22.64, 6.36, -2.58, 12.06, 23.74 - 10, 2.76, 1.5, 15.16, 1.96, 0.33, -0.57, 0.92, 1.54, -3.17, -1.06, 0.61, 0.33, -5.13, -12.84, -0.75, -5.1, 6.58, -19.11, -13.75, 3.06, -10.72]^T$$

$$\hat{\mathbf{w}}_{\mathbf{cl3}} = [22.26, -4.8, 14.15, -5.34, 5.64, -2.15, 16.88, -81.31, 125.16, 19.5, -14.59, -8.63, 19.91, 6.18, -8.84, -5.95, 26.22, 69.49, -139.24, 0.19, 0.19, -6.92, 9.36, 4.26, -5.41, -4.16, 8.34, -6.84, 1.47, -2.54, 0.19, 67.1, -17.85, 3.09, -14.38, 5.09, -52.09, -16.21, 0.14, -13.67]^T$$

$$\hat{\mathbf{w}}_{\mathbf{cl4}} = [2.65e01, 1.48e01 - 3.73e01, 3.67e01, 1.67, -8.74, 7.78, -5.96e02, 1.71e03, -1.63e03, 1.64e03, -3.01e03, 1.91e03, -7.54e02, 9.75e02, -3.75e02, 1.16e01, -4.05e01, 5.21e01, -1.26e01, 3.26e01, -2.65e01, -4.48e02, 1.23e03, -1.11e03, -4.91e01, 1.87e02, -2.32e02, 1.33e-01, 1.33e-01, 1.33e-01, 9.73, -2.97e01, 2.59e01, 1.91, -9.39e-01, -1.74, -2.33e-01, -4.55, 9.60, -7.87, 9.19e-01, -6.22, 1.33e-01, 9.17e01, 3.32e02, 6.62, -2.67e01, 4.82e01, -4.31e02, 5.14e02, -4.95e-01, -1.86e01]^T$$

The training error and testing error are shown in the following plot.

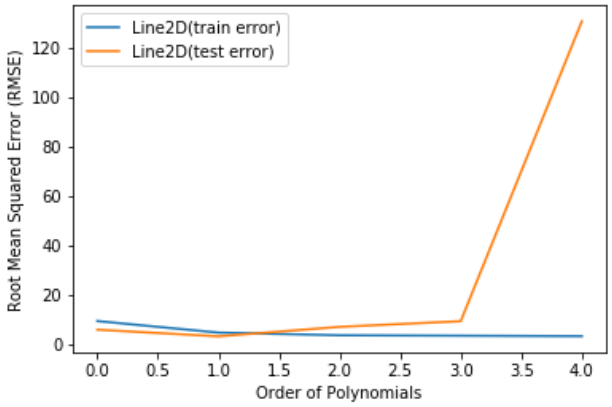

Figure 3. Training and Testing error vs. Polynomial Orders

Part b) Linear Regression with Different Split Ratio. In this problem, we vary the ratio of training set and testing set. The trained parameter vector $\hat{\mathbf{w}}_{\mathbf{cl}}$ is:

$$\hat{\mathbf{w}}_{\mathbf{cl,p=0.2}} = [8.71, -30.46, 0.14, -1.19, -2.37, 2.61, 6.11, -2.27, -0.04, -1.12, -1.18, 0.4, 088, 0.28]^T \tag{4}$$

$$\hat{\mathbf{w}}_{\mathbf{cl,p=0.4}} = [27.7, 10.6, 0.64, 0.0257, -0.007, -0.837, 6.43, -1.15, -2.47, 4.15, -3.05, -1.47, 1.70, -1.26]^T \tag{5}$$

$$\hat{\mathbf{w}}_{\mathbf{cl,p=0.6}} = [26.01, 10.55, 0.3, 0.13, 0.17, -0.94, 6.41, -1.32, -2.12, 1.08, -2.51, -1.33, 1.53, -0.83]^T \tag{6}$$

$$\hat{\mathbf{w}}_{\mathbf{cl,p=0.8}} = [23.13, -1.64, 1.03, 0.36, 0.48, -1.73, 3.34, 0.08, -2.74, 3.99, -2.62, -1.75, -0.2, -3.8]^T \tag{7}$$

$$\hat{\mathbf{w}}_{\mathbf{cl,p=1}} = [2.96, -1.24, 0.84, -3.81, -0.81, -0.43, 1.22, -0.36, 0.42, -2.91, -1.97, -4.32, 1.31, -3.18]^T \tag{8}$$

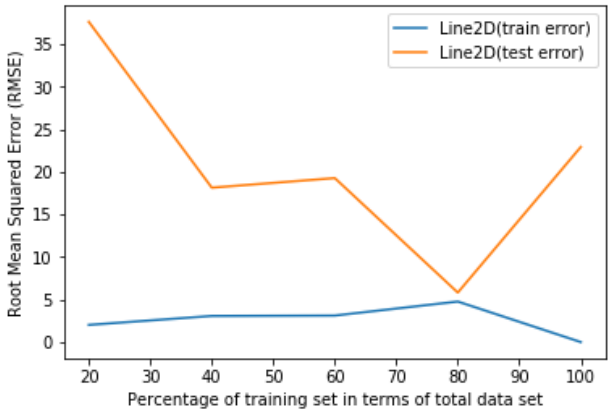

FIGURE 4. Training and Testing error vs. Data Split Ratio

As you can tell from the figure above, the train error converges from the bottom while the test error converges from above. The last data point is skewed because only 1 data point was assigned to the test set.

Problem 3 Regularized Linear Regression

Consider the following regularized linear regression least squares:

$$E(w) = \frac{1}{2N}\sum_{i=1}^{N}(w^T\phi(x_n) - t_n)^2 + \frac{\lambda}{2}||w||^2 \tag{9}$$

Part a) Derive a Closed Form Solution. Here we derive a closed form solution:

$$E(\mathbf{w}) = \frac{1}{2N}\mathbf{w}^T\Phi^T\Phi\mathbf{w} - \frac{1}{N}\mathbf{t}^T\Phi\mathbf{w} + \frac{1}{2N}\mathbf{t}^T\mathbf{t} + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{10}$$

The gradient is:

$$\nabla_{\mathbf{w}}E(\mathbf{w}) = \frac{1}{N}\Phi^T\Phi\mathbf{w} + \lambda\mathbf{w} - \frac{1}{N}\Phi^T\mathbf{t} \tag{11}$$

Setting the gradient to zero, we will get the solution:

$$\hat{\mathbf{w}} = (\Phi^T\Phi + N\lambda I)^{-1}\Phi^T\mathbf{t} \tag{12}$$

Part b) Regularized Linear Regression with Different Hyper-parameter λ. In this section, we split the data set into three different parts: Training Set, Validation Set, and Testing Set. We first use training set and validation set to find best hyper-parameter λ and then test the selected model using the testing set. For the model selection process, we select λ from the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ The following figure shows the train and test error vs. λ

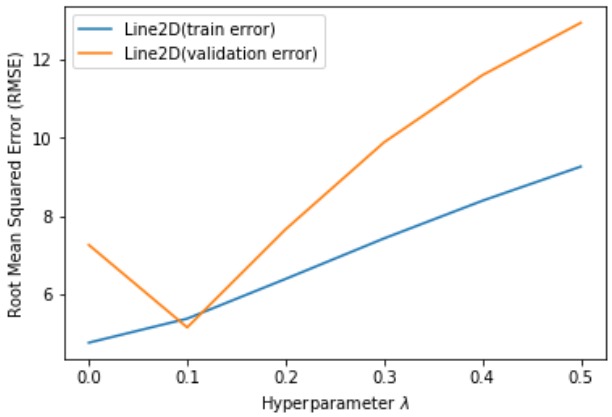

Figure 5. Training and Testing error vs. Hyperparameter λ

As you can see from the figure above, $\lambda = 0.1$ corresponds to the best option. Using $\lambda = 0.1$, we obtain a **test error** of **3.943**

Problem 4 Weighted Linear Regression

Consider a linear regression problem in which we want to weigh different training examples differently. Specifically, suppose we want to minimize

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n(\mathbf{w}^T\mathbf{x_n} - \mathbf{t_n})^2 \tag{13}$$

Part a). Show that E(w) can be also expressed as:

$$E(\mathbf{w}) = (\mathbf{Xw} - \mathbf{t})^T\mathbf{R}(\mathbf{Xw} - \mathbf{t}) \tag{14}$$

Proof: We can omit the coefficient $\frac{1}{2}$ without affecting the optimization result

$$\begin{aligned} E(\mathbf{w}) &= \sum_{n=1}^{N}(\sqrt{r_n}\mathbf{w}^T\mathbf{x_n} - \sqrt{r_n}\mathbf{t_n})^2 \\ &= ||\sqrt{\mathbf{R}}\mathbf{Xw} - \sqrt{\mathbf{R}}\mathbf{t}||^2 \\ &= (\mathbf{Xw} - \mathbf{t})^T\mathbf{R}(\mathbf{Xw} - \mathbf{t}) \end{aligned}$$

where:

$$\mathbf{R} = \begin{bmatrix} r_1 & & \\ & \ddots & \\ & & r_N \end{bmatrix}$$

$$\mathbf{t} = \begin{bmatrix} t_1, \ldots, t_N \end{bmatrix}^T$$

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1M} \\ x_{20} & x_{21} & \ldots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N0} & x_{N1} & \ldots & x_{NM} \end{bmatrix}$$

Part b) Find the optimal w. Find the value of w that minimizes the above loss function by computing the derivative and setting it to zero. Express this optimal $w = w^*$ in terms of X,R and t.

The gradient is:

$$\nabla_w E(w) = 2\mathbf{X^T}\mathbf{R}(\mathbf{Xw} - \mathbf{t}) \tag{15}$$

$$= 2\mathbf{X^T}\mathbf{RXw} - 2\mathbf{X^T}\mathbf{Rt} \tag{16}$$

Setting gradient to zero, we have the following:

$$2\mathbf{X^T}\mathbf{RXw} - 2\mathbf{X^T}\mathbf{Rt} = 0 \tag{17}$$

$$\mathbf{X^T}\mathbf{RXw} = \mathbf{X^T}\mathbf{Rt} \tag{18}$$

$$\hat{\mathbf{w}} = (\mathbf{X^T}\mathbf{RXw})^{-1}\mathbf{X^T}\mathbf{Rt} \tag{19}$$

Part c) Maximum Likelihood Estimator. Suppose we have a training data set $\{(x_i, t_i); i = 1, .., N\}$ of N independent samples. t_i is observed with different variances. Specifically, we suppose that:

$$p(t_i|\mathbf{x_i}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp -\frac{(t_i - \mathbf{w}^T\mathbf{x_i})^2}{2\sigma_i^2} \tag{20}$$

Therefore, since all observations are independent. We have the following rule:

$$p(\mathbf{t}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^{N} p(t_i|\mathbf{x_i}; \mathbf{w}) \tag{21}$$

We take the natural log on both sides of the equation:

$$\ln(p(\mathbf{t}|\mathbf{X};\mathbf{w})) = \ln(\prod_{i=1}^{N} p(t_i|\mathbf{x_i};\mathbf{w})) \tag{22}$$

$$= \sum_{i=1}^{N} \ln(p(t_i|\mathbf{x_i};\mathbf{w})) \tag{23}$$

$$= \sum_{i=1}^{N} \left[\ln\big(\frac{1}{\sqrt{2\pi}\sigma_i}\big) - \frac{(t_i - \mathbf{w}^T\mathbf{x_i})^2}{2\sigma_i^2} \right] \tag{24}$$

Now, we find the maximum likelihood with respect to $\mathbf{w}$.

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left[\ln(p(\mathbf{t}|\mathbf{X};\mathbf{w})) \right] \tag{25}$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left[\ln\big(\frac{1}{\sqrt{2\pi}\sigma_i}\big) - \frac{(t_i - \mathbf{w}^T\mathbf{x_i})^2}{2\sigma_i^2} \right] \tag{26}$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \left[- \frac{(t_i - \mathbf{w}^T\mathbf{x_i})^2}{2\sigma_i^2} \right] \tag{27}$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2} \sum_{i=1}^{N} \left[\frac{1}{\sigma_i^2}(t_i - \mathbf{w}^T\mathbf{x_i})^2 \right] \tag{28}$$

Here we proved that the maximum likelihood method leads to the same results as weighted linear regression method. Specifically, we have a relationship between r_i and σ_i as:

$$r_i = \frac{1}{\sigma_i^2} \tag{29}$$