EECS 545 Homework 1 Solution

Problem 1

Standardization

We need to need to state the difference between standardization and adding ones. Standardization is widely used as a feature-scaling method. It does following transformation,

$$x' = \frac{x - \bar{x}}{\sigma}$$

where $\bar{x}$ stands for mean value and σ for standard deviation. It will force values of one feature have zero-mean and unit-variance. Specially, for linear models, if we do standardization on features, the bias term must be the mean value of continuous labels.

$$\begin{aligned}
\sum_{i=1}^{N}(w^T(x_n - \bar{x}) + b - t_n)^2 &= \sum_{i=1}^{N}\left[(w^T x_n - w^T\bar{x})^2 + 2(b - t_n)(w^T x_n - w^T\bar{x}) + (b - t_n)^2\right] \\
&= \sum_{i=1}^{N}(w^T x_n - w^T\bar{x})^2 + \sum_{i=1}^{N} 2(b - t_n)(w^T x_n - w^T\bar{x}) + \sum_{i=1}^{N}(b - t_n)^2 \\
&= \sum_{i=1}^{N}(w^T x_n - w^T\bar{x})^2 - \sum_{i=1}^{N} t_n(w^T x_n - w^T\bar{x}) + \sum_{i=1}^{N}(b - t_n)^2 \\
b_{min} = arg\min_b \sum_{i=1}^{N}(b - t_n)^2 = \bar{t}
\end{aligned}$$

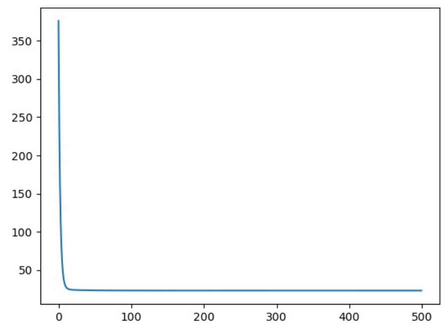
An alternative method is to add extra additional ones in features, like $x' = [x, 1]$ to include bias term in weight vector. However, if the objective function contains a regularization term, we need to remove this bias-like weight factor.

Stochastic Gradient Descent

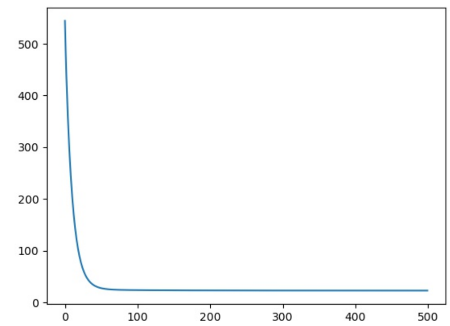
Learning Rate: 5e-4
Weight Vector:

$$w = [-0.803, 0.975, -0.169, 0.737, -1.721, 2.953, 0.152, -2.925, 1.584, -1.13, -1.877, 0.92, -3.93]^T$$

Bias: 22.94087791
Train error: 23.463031612134426
Test error: 9.674516161718294



(a) Stochastic Gradient Descent Training MSE



(b) Batch Gradient Descent Training MSE

Batch Gradient Descent

Learning Rate: 5e-2
Weight Vector:

$$w = [-0.802, 0.981, -0.164, 0.735, -1.714, 2.948, 0.154, -2.929, 1.608, -1.16, -1.874, 0.918, -3.937]^T$$

Bias: 22.94078326
Train error: 23.45431171000535
Test error: 9.693555099802818

Closed Form Solution

Weight vector:

$$w = [-0.937, 1.19, 0.218, 0.67, -2.105, 2.751, 0.308, -3.124, 2.961, -2.455, -2.007, 0.906, -4.057]^T$$

Bias: 22.94100877
Train error: 23.1915564692
Test error: 10.9665431668

Random Split Dataset

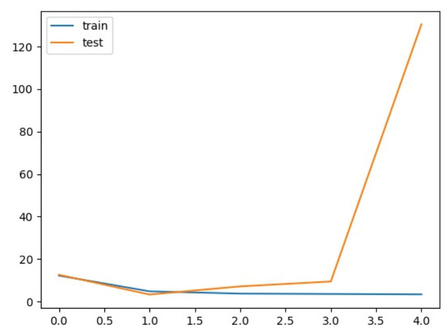
Mean train error: 21.508138
Mean test error: 26.661521
Obviously data is given in order. Test error usually larger than train error.

Problem 2

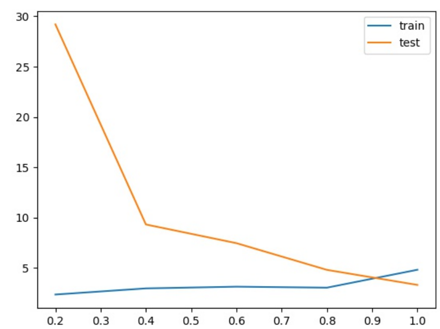
Polynomial Features & Partial Training Data

Some things need to be noticed:

- We need to plot RMSE instead of MSE in this problem.
- Feature-scaling(normalization, standardization) is still used here.
- Your plot should combine curves for test errors and train errors in the same figure. And a legend is required.



(a) RMSE of Polynomial Features



(b) RMSE of Partial Training Data

Problem 3

The original objective function is:

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

Closed Form Solution

First, using matrix Φ and vector t to

$$\begin{aligned} E(w) &= \frac{1}{2N} \sum_{i=1}^{N} (w^T \phi(x_n) - t_n)^2 + \frac{\lambda}{2} \|w\| \\ &= \frac{1}{2N} (\Phi w - t)^T (\Phi w - t) + \frac{\lambda}{2} w^T w \\ &= \frac{1}{2N} w^T \Phi^T \Phi w - \frac{1}{N} t^T \Phi w + \frac{1}{2N} t^T t + \frac{\lambda}{2} w^T w \end{aligned}$$

By setting the gradient $E'(w)$ to zero and solving the equation, the closed form solution for w is

$$\begin{aligned} E'(w) = \frac{1}{N} \Phi^T \Phi w - \frac{1}{N} \Phi^T t + \lambda w &= 0 \\ \left(\Phi^T \Phi + \lambda N I\right) w &= \Phi^T t \\ w &= \left(\Phi^T \Phi + \lambda N I\right)^{-1} \Phi^T t \end{aligned}$$

Regularization Result

Select λ as 0.30 and the RMSE on test set is 5.116131249426067.

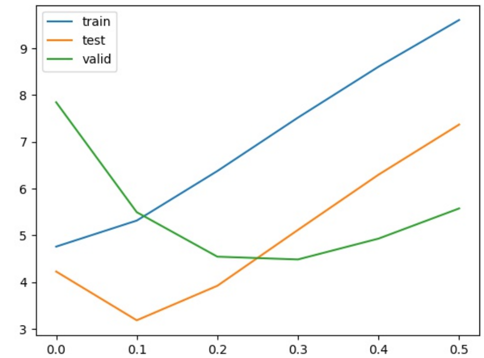


Figure 3: RMSE of Regularization

Unfortunately, if you get a figure like that:

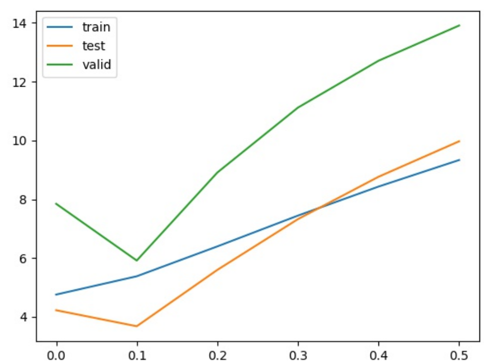


Figure 4: Wrong RMSE of Regularization

Please be aware that the mean and standard deviation in normalization **only** come form **training** set. The training set is the first 90 percent of the old training set used in problem 1.

Problem 4 Weighted Linear Regression

Proper Form

The matrix X is the same definition as usual,

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,D-1} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_N \end{bmatrix}$$

and target values vector t is defined as the same too,

$$t^T = [t_1, t_2 \ldots, t_N]$$

but the matrix R is defined as a diagonal matrix for weights r_i

$$R = \frac{1}{2}\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & r_N \end{bmatrix}$$

Following is the proof of the equality between these two forms.

Proof Obviously,

$$d^T = (Xw - t)^T = [(w^T x_1 - t_1), (w^T x_2 - t_2), \cdots, (w^T x_N - t_N)]$$

Therefore,

$$\begin{aligned} d^T R d &= \frac{1}{2}[r_1(w^T x_1 - t_1), r_2(w^T x_2 - t_2), \cdots, r_N(w^T x_N - t_N)] \times \begin{bmatrix} (w^T x_1 - t_1) \\ (w^T x_2 - t_2) \\ \vdots \\ (w^T x_N - t_N) \end{bmatrix} \\ &= \frac{1}{2} r_1 (w^T x_1 - t_1)^2 + \frac{1}{2} r_2 (w^T x_2 - t_2)^2 + \cdots + \frac{1}{2} r_N (w^T x_N - t_N)^2 \\ &= \frac{1}{2}\sum_{i=1}^{N} r_i (w^T x_i - t_i)^2 \end{aligned}$$

Close Form Solution

$$\begin{aligned} E(w) &= (Xw - t)^T R (Xw - t) \\ &= w^T X^T R X w - 2 w^T X^T R t + t^T R t \\ E'(w) &= 2X^T R X w - 2X^T R t \end{aligned}$$

Let the gradient $E'(w)$ to zero, we get

$$\begin{aligned} X^T R X w &= X^T R t \\ (\sqrt{R}X)^T(\sqrt{R}X) w &= (\sqrt{R}X)^T \sqrt{R} t \\ w &= (\sqrt{R}X)^\dagger \sqrt{R} t \end{aligned}$$

Likelihood

The log-likelihood is

$$\log \prod_{i=1}^{N} p(t_i|x_i; w) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma_i^2}} - \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(t_i - w^T x_i)^2$$

Because we already know the values of σ_i, the first term $\sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma_i^2}}$ is a constant. Therefore, the log-likelihood maximization problem is equal to the minimization problem of the second term, $\sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(t_i - w^T x_i)^2$. By assigning $r_i = \frac{1}{\sigma_i^2}$, we transform this likelihood maximizing problem in to the minimizing problem stated in the beginning.

All $r_i \propto \frac{1}{\sigma_i^2}$ would be regarded as a correct answer.