6.867: Homework 1

1. Gradient descent

As in the past few years, ICML will rely exclusively on electronic formats for submission and review.

1.1. Templates for Papers

2. Linear basis function regression

We consider linear basis function regression as a method to benchmark the robustness of the gradient descent solution presented above. By using the closed-form maximum likelihood equation, we can calculate the maximum likelihood weight vector for our list of basis functions to approximate the data in the form of our basis. In this scenario, we are using data generated by $y(x) = \cos(\pi x) + 1.5\cos(2\pi x) + \epsilon(x)$, where $\epsilon(x)$ is some added noise to the dataset. Running linear regression on a simple polynomial basis of order M, where $\phi_0(x) = x^0$, $\phi_1(x) = x^1$, $\phi_2(x) = x^2$, ..., $\phi_M(x) = x^M$, we calculate the maximum likelihood weight vector by the following:

$$w_{ML} = (\Phi^T\Phi)^{-1}\Phi^T y$$

where w_{ML} is the maximum likelihood weight vector and Φ is given by:

$$\Phi = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \phi_2(x_0) & \dots & \phi_M(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_M(x_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \phi_2(x_n) & \dots & \phi_M(x_n) \end{bmatrix}$$

Our choice of M, the degree of our polynomial basis, largely determines the fit of the regression to the data (Figure 2). More specifically, as small values of M, the polynomial basis cannot adequately capture all the data points. At higher values of M however, overfitting occurs in which the weight vector performs well on the training data, but is not well generalized to new data. The value of M therefore must be carefully considered in order to prevent too high variability in our generated regression polynomial.

We can also instead choose our set of basis functions to be the set of cosine functions, where $\phi_1(x) = \cos(\pi x)$, $\phi_2(x) = \cos(2\pi x)$, ..., $\phi_M(x) = \cos(M\pi x)$. This is again calculated for multiple values of M and compared in Figure 3. Interestingly, even when we use the same family of basis functions as used to generate the initial data ($M = 2$), due to the noise $\epsilon(x)$ added to the initial dataset, the maximum likelihood weight vector does not identically match the actual function used:

$$w = \begin{bmatrix} 1 \\ 1.5 \end{bmatrix}$$

$$w_{MLE} = \begin{bmatrix} 0.779 \\ 1.174 \end{bmatrix}$$

where w is the actual weight vector and w_{MLE} is the maximum likelihood estimated weight vector.

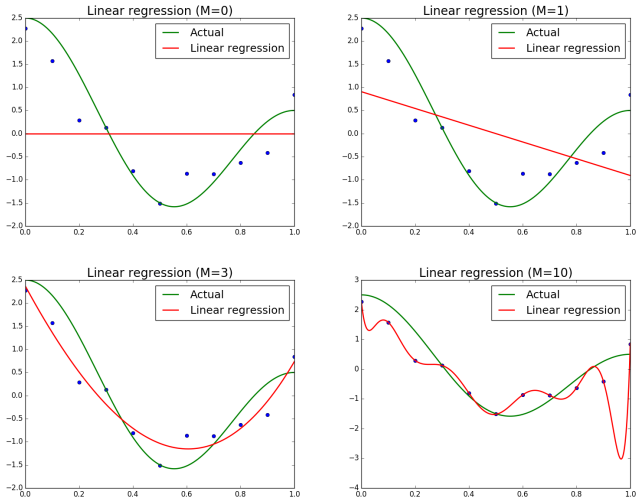

Figure 1. Linear regression for varying values of M.

3. Ridge regression

As illustrated in the previous section, some values of M can result in overfitting to the training set. Thus, in an attempt to reduce overfitting, we implement ridge regression, which adds a regularization parameter λ to the regression function. In this case, we have chosen a regularizer to serve as a weight decay, in order to encourage weight vector values to tend towards zero unless otherwise supported by the data.

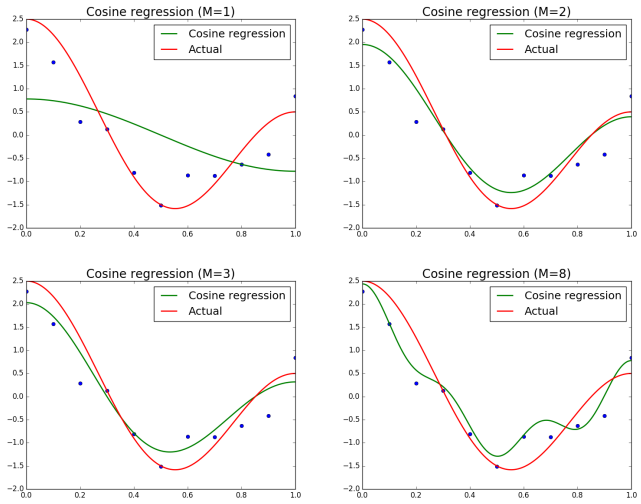

Figure 2. Linear regression for varying values of M.

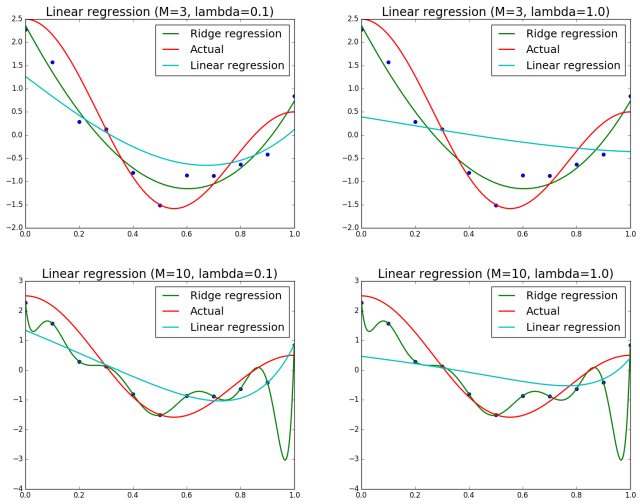

Figure 3. Ridge regression for values of M and λ.

For any given value of M therefore, with greater weight decay regularization coefficients, the weight vector decays more strongly towards zero. In other words, for higher values of λ, the data must more strongly support greater weight vector values in order to acehive the same magnitude of coefficients in the weight vector.

This is illustrated in Figure 4, where the first row illustrates the fit of ridge regression using a polynomial basis of degree 3 with $\lambda = 0.1$ and 1 and the second row shows the fit of the ridge regression using a polynomial basis of degree 10. As illustrated, the weight vector coefficients in the regression are dampened towards zero with greater values of the regularization coefficient. This is compared with the regression computed with a $\lambda = 0$ to illustrate the decay effect of the regulaization parameter.

Using additional data sets for validation and testing, we can select for the best M and λ parameters in our regression model using a polynomial basis. More specifically, for many combinations of values of M and λ, the maximum likelihood weight vector w_{MLE} was calculated based on the training data set. Through model selection on the validation data set, the best model and optimal values for M and λ were chosen based on the evaluation metric. This optimal regression model was then run on the testing data set to evaluate the performance of the model on the new data. The maximum likelihood weight vectors were generated for all $M \in [0, 10]$ and $\lambda \in \{0, 0.1, 0.2, ..., 1.4, 1.5\}$, and this was performed multiple times with data set A as the training data and data set B as the testing data, as well as vice versa. The performance of the regression model was evaluated using the sum of squared errors (SSE) and the mean squared errors (MSE); qualitatively, both evaluation metrics result in the same conclusions, although quantitatively, some additional information can be gleaned from each evaluation metric.

When data set A was used as the training data and data set B was used as the testing data, the validation step indicated that $M = 2$ and $\lambda = 0.0$ yielded the optimal regression model with a minimum SSE of 2.35 and a minimum MSE of 0.11. Running this model on the testing data demonstrated that the regression fit the testing data remarkably well, excepting a single outlier in the testing data. Thus, despite a relatively low MSE of 2.58 on the testing data, the SSE was 25.75 due to the significant error in the outlier data point. The training, validation, and testing steps are illustrated in Figure 6, in which the best regression line chosen in model selection is plotted against each data set used for training, validation, and testing, respectively.

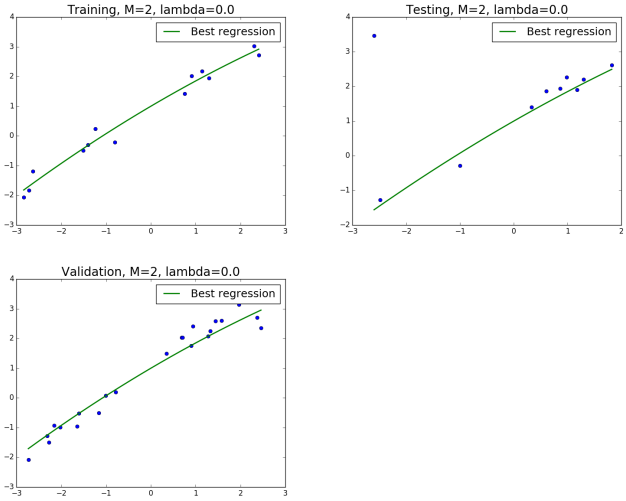

Figure 4. Ridge regression for values of M and λ.