Department of Electrical Engineering

EE5610/AI5000 - PRML

Linear Regression - 1

Understanding Basic Concepts

1. **Understanding Error Surface:**

   According to www.geogebra.org, the relationship between human height (in inches) and weight (in pounds) is given by

   $$t = 3.86x - 110.42$$

   (a) Generate 25 meaningful data points from this relationship, mimicking a noisy sensor, where the noise follows a zero mean Gaussian with a variance of 20. Plot the scatter plot of the data.

   (b) Now, we need to estimate the above relationship from the noisy data generated in (a) by fitting a line, i.e., $\hat{t} = y(x, w) = w_0 + w_1 x$. Let us use least squares criterion discussed in the class to estimate the parameters w_0 and w_1. Generate and plot the error surface $J(w_0, w_1)$ associated with this approach. Locate the minimum on this error surface.

   (c) Estimate the parameters using least squares approach, and compare them with the desired values.

2. **Understanding Model Order and Over-fitting:**

   (a) Generate 20 data points from $t_n = \sin(2\pi x_n) + e_n$, where $x_n \in [01]$ and $e_n / \mathcal{N}(0, 0.1)$, and divide them into two sets, a training set and a testing test each containing 10 points.

   (b) Fit an M^{th} degree polynomial to the training data using least squares approach, i.e.,

   $$\hat{t}_n = w_0 + w_1 x + \cdots + w_m x^m + \cdots + w_M x^M$$

Use the estimated parameter vector $\mathbf{w}$, to predict the target values in training and testing datasets. Plot the root mean squared error associated with each dataset, for $M = 0, 1, \ldots, 9$. Explain your results.

(c) Increase the size of the training dataset o 100 points, and repeat (b).

(d) Add a l_2 regularization term to the objective function in (b) and repeat (b) and (c). Study the affect of Lagrange multiplier λ on the root mean squared error of the training and testing datasets.

(e) Modify the function in (a) to $t_n = 5 + \sin(2\pi x_n) + e_n$ to study the effect of regularizing the bias coefficient w_0.

3. **Understanding Choice of Kernel:**

(a) Generate 100 data points from $t_n = \sin(2\pi x_n) + e_n$, where $x_n \in [0\ 1]$ and $e_n / \mathcal{N}(0, 0.1)$, and divide them into two sets, a training set and a testing test each containing 10 points. Fit an M^{th} degree polynomial using polynomial, Gaussian and sigmoidal kernels, and study the goodness of fit in each case, for different model orders M.

(b) Repeat (a) by modifying the target function to

$$t_n = \begin{cases} \text{sinusoid} + e_n & x \in [0, 1) \\ \text{triangle} + e_n & x \in [1, 2) \\ \text{Gaussian pulse} + e_n & \in [2, 3] \end{cases}$$

Clearly discuss your observations/results for each of the three kernels.

4. **Understanding Online Training:**

(a) Repeat 3(a) and 3(b) using stochastic gradient descent for weight update. Study the effect of step size η on convergence of the weights, and compare them to those obtained using closed form expressions in 3. Plot the mse as a function of iterations. (b) Study the effect of batch size on the speed of convergence.

5. **Understanding Bias variance Tradeoff:**

Generate L=100 datasets of noisy sinusoidal data, each having N=25 data points. For each dataset, fit a $M = 25^{th}$ order linear regression model consisting of 24 Gaussian basis functions and one bias parameter. Use regularized least squares, governed by the parameter λ, to estimate the parameters $\mathbf{w}$. Illustrate the concept of bias and variance using these 100 different parameter fits.

6. **Understanding MAP estimate**

(a) Generate 100 noisy data points of a sinusoid. Fit a 20th order linear regression model with Gaussian basis functions. Starting from a standard normal prior, update the statistivs of the posterior density of the parameters using Bayesian sequential updates.

(b) Sample a parameter vector from the posterior distribution, and obtain the curve fit for this realization. Repeat this for several times, and estimate the average of these curve fits, and compare it with the original sinusoid.

(c) Use the posterior distribution of the parameters to evaluate the predictive distribution of target $p(t_0/x_0, X, t)$, and plot it for different number of training datapoints, as discussed in the class.