Problem 5: Linear Least Square Problem and the Bias-Variance Trade-off in Machine Learning

Presenter: Professor Montaz Ali, University of the Witwatersrand

Problem Statement:

1 Linear Least Square

Given an m by n matrix A and an m by 1 vector $\boldsymbol{b}$, the linear least squares problem is to find an n by 1 vector $\boldsymbol{x}$ minimizing $||\boldsymbol{Ax}-\boldsymbol{b}||_2$. If $m=n$ and A is nonsingular, the answer is simply $\boldsymbol{x}=A^{-1}\boldsymbol{b}$. But if $m>n$ so that we have more equations than unknowns, the problem is called over-determined, and generally no $\boldsymbol{x}$ satisfies $A\boldsymbol{x}=\boldsymbol{b}$ exactly. One occasionally encounters the under-determined problem, where $m<n$, but we will concentrate on the more common over-determined case. Let us begin with two types of linear regression problems: the polynomial regression & statistical modeling via regression (there is also the K-Nearest Neighbor Regression).

1.1 Polynomial Regression

Suppose that we have m pairs of numbers $(y_1, b_1), \cdots, (y_m, b_m)$ and that we want to find the best cubic polynomial fit to b_i as a function of y. This means finding polynomial coefficients $x_l, \cdots, x_4$ so that the polynomial

$$f(y)=\sum_{j=1}^{4} x_j y^{j-1} \tag{1}$$

minimizes the residual $r_i = f(y_i) - b_i$ for $i = 1 \cdots m$. We can also write this as minimizing

$$\boldsymbol{r}=\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix}=\begin{pmatrix} f(y_1) \\ f(y_2) \\ \vdots \\ f(y_m) \end{pmatrix}-\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

This implies

$$\boldsymbol{r}=\begin{pmatrix} 1 & y_1 & y_1^2 & y_1^3 \\ 1 & y_2 & y_2^2 & y_2^3 \\ \vdots & \vdots & \vdots\vdots & \\ 1 & y_m & y_m^2 & y_m^3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}-\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$\boldsymbol{r}=\boldsymbol{Ax}-\boldsymbol{b}$$

One frequently uses $||\boldsymbol{Ax}-\boldsymbol{b}||_2$ which corresponds to minimizing the sum of the squared residuals $\sum_{i=1}^{m} r_i^2$, is a linear least squares problem.

1.2 Regression in Statistical Modeling

Suppose that we are doing medical research on the effect of a certain drug on blood sugar level. We collect data from each patient (numbered from $i = 1, 2 \cdots, m$) by recording his or her initial blood sugar level ($a_{i,1}$), final blood sugar level (b_i), the amount of drug administered ($a_{i,2}$), and other medical quantities, including body weights on each day of a week-long treatment ($a_{i,3}$ through $a_{i,n}$). In total, there are $n < m$ medical quantities measured for each patient. Our goal is to predict b_i given $a_{i,1}$ through $a_{i,n}$ and we formulate this as the least squares problem: minimize $||\boldsymbol{Ax} - \boldsymbol{b}||_2$. We plan to use $\boldsymbol{x}$ to predict the final blood sugar level b_j of future patient j by computing

$$b_j = \sum_{k=1}^{n} a_{jk} x_k$$

1.3 Normal Equation

To derive the normal equations, we look for the $\boldsymbol{x}$ where the gradient of $||\boldsymbol{Ax} - \boldsymbol{b}||_2^2$ vanishes. Prove that the normal equation is given by $A^T A\boldsymbol{x} = A^T\boldsymbol{b}$. Why is $\boldsymbol{x} = A^T A^{-1} A^T \boldsymbol{b}$ is the minimizer of $||\boldsymbol{Ax} - \boldsymbol{b}||_2$?

2 Bias and Variance

Understanding how different sources of error lead to bias and variance helps us improve the data fitting process resulting in more accurate models. We first define bias & variance conceptually.

2.1 Conceptual Definition

Error due to Bias: The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Of course you only have one model so talking about expected or average prediction values might seem a little strange. However, imagine you could repeat the whole model building process more than once: each time you gather new data and run a new analysis creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions. Bias measures how far off in general these models' predictions are from the correct value.

Error due to Variance: The error due to variance is taken as the variability of a model prediction for a given data point. Again, imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.

2.2 The Intuition on Bias–Variance Trade-off

Suppose that we have a (training) data set

$$S = \{(y_1, b_1), \cdots, (y_m, b_m)\}$$

The goal in learning is not to learn (developed a model) an exact representation of the training data itself, but to build a statistical model of the process which generates the data. We will see in the case of polynomial regression that:

- models with too few parameters can perform poorly
- models with too many parameters can perform poorly

Need to optimize the complexity of the model to achieve the best performance. One way to get insight into this tradeoff is the decomposition of generalization error into bias^2 + variance:

- a model which is too simple, or too inflexible, will have a large bias (high bias means a poor match).
- a model which has too much flexibility will have high variance (a high variance means a weak match).

We would like to minimize each of these. Unfortunately, we cannot do this independently, there is a trade-off.

2.3 Bias-Variance Analysis in Regression

Consider you have a data set S with 500 pairs of (y_i, b_i). You can generate these pairs from the function $b = f(y) + \epsilon$ i.e.

$$b = y + 2\sin(1.5 \times y) + \epsilon, \qquad \epsilon \in N(0, 0.02), \quad y \in [0, 10]$$

where $N(0, 0.02)$ is a normal distribution with zero means and $\sigma = 0.02$. Here the true function is $f(y)$. We would like to develop or estimate a model $h(y)$ of $f(y)$.

Construct five sample $S_1, S_2 \cdots, S_5$ each S_i has 50 pairs generated randomly (with replacement) from S. For each data set S_i construct a polynomial model of degree p i.e.

$$h(y) = \sum_{j=1}^{p} x_j y^{j-1}. \tag{2}$$

The optimal $\boldsymbol{x}$ is what minimizes $||\boldsymbol{Ax} - \boldsymbol{b}||_2$.

Now that we have a model $h(y)$, given any new data point y^* (with observed value $b^* = f(y^*) + \epsilon$, the pairs (y^*, b^*) are in S and not in any of S_i), we would like to understand the expected prediction error

$$E\left[(b^* - h(y^*))^2\right] \tag{3}$$

Equation (3) can be decomposed into Bias^2, Variance and Irreducible error. We now consider the bias-variance calculation.

Select a set $\hat{S}$ of pairs from S that do not belong to any of the S_i. Hence for each pair $(y_i, b_i) \in \hat{S}$ there are predictions from five models. Hence calculates $E\left[\left(h(y) - h(\bar{y})\right)^2\right]$ or $E\left[(h(y) - E\left[h(y)\right])^2\right]$ (Variance) and $\left(h(\bar{y}) - f(y)\right)^2$ or $\left(E\left[h(y)\right] - f(y)\right)^2$ (Bias^2). Repeat the process for $p = 1, 2$ and 3 and compare the bias and Variance for each p.