TIM245 Lecture 8 (4/26/17)

Agenda

1) Roadmap and schedule for the quarter

2) Review Lecture 7

3) Multiple Linear Regression

4) Model training

5) Extension to linear regression: ridge, lasso, elastic nets

1 Roadmap for the course

Schedule for the next seven weeks

Week	Lecture	Homework	Project	Exams
Week 4 4/24	Supervised Prediction	HW 1 Due	Phase II Due	
Week 5 5/1/17	Supervised Classification	HW 2 Assigned	Phase II Assigned	
Week 6 5/8	Unsupervised Clustering	HW 2 due	Phase III due	Midterm Assigned
Week 7 5/15	Unsupervised Association			Midterm Due
Week 8 5/22	Text Mining	HW 3 Assigned	Phase IV Assigned	
Week 9 5/29	Time Series Analysis	HW 3 Due	Phase IV due	
Week 10 6/5	Graph Mining		Phase V assigned	Final assigned
Finals 6/12			Phase V due	Final Due

③ Multiple Linear Regression

<u>Multiple linear regression</u>: linear regression with multiple independent variables or attributes and a single dependent variable or target

(Multivariate linear regression refers to multiple dependent variables)

$$\hat{y} = \beta_0 + \beta_1 X_{i1} + \beta X_{i2} + \ldots + \beta_m X_{im}$$

The values $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ are the linear regression model.

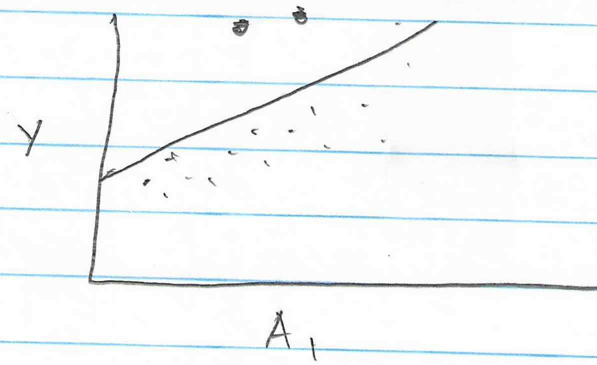
The general process or learning algorithm for creating the model

1) Model Training: estimate coefficients that minimize squared error (least squared error) on the training data

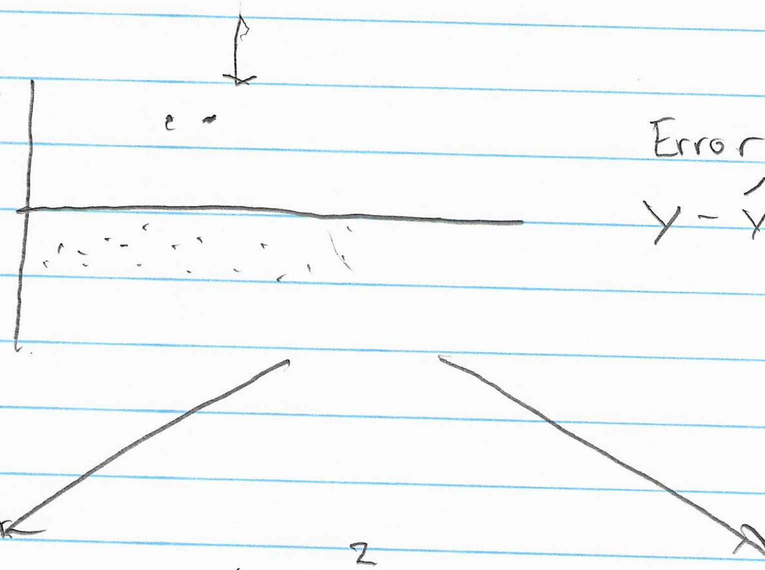
2) Model Testing: evaluate coefficients on the test data set

(4) Model Training

How do we find the best fit line for the training data set



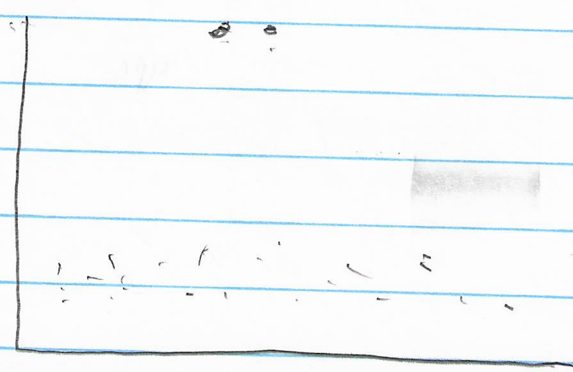
Training Data

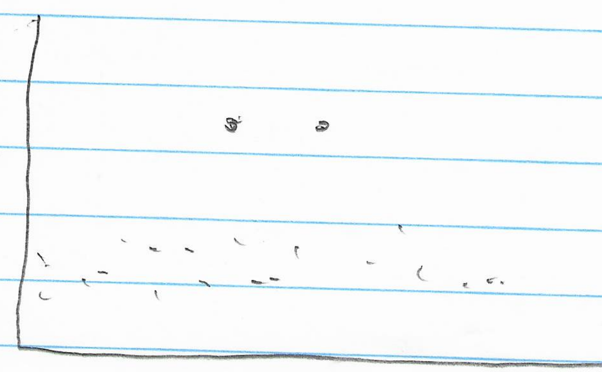


Error

$Y - \hat{Y}$

Squared Error $(y-\hat{y})^2$ | Absolute Error $|y-\hat{y}|$





Two general ways to find the best fit line:

1) Minimize squared error (L2 Norm)

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \beta X_i)^2$$

2) Minimize absolute error (L1 Norm)

$$AE(\beta) = \sum_{i=1}^{n} |y_i - \beta X_i|$$

where

$$\beta = [\beta_1, \beta_2, \ldots \beta_m]$$

$$X_i = [x_{i1}, x_{i2}, \ldots x_{im}]$$

$$\beta X_i = \hat{y}_i$$

L2 Norm loss function: not robust, stable, one solution

L1 Norm loss function: robust, not stable, multiple Solution

L2 Norm loss function typically used because of computational efficiency

Two general approaches to minimize $RSS(\beta)$

<u>Normal Equations</u>
Closed form analytical Solution for β
(Slow for lots of attributes)

<u>Stochastic Gradient Descent (SGD)</u>
Approximate Solution through gradient descent
(useful for large data sets)

Normal Equations:

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

$$= (y_i - \beta x_i)^T (y_i - \beta x_i)$$

where

Training Data Set

A_1	A_2	$\cdots$		Y
	X			Y

β_0 placeholder

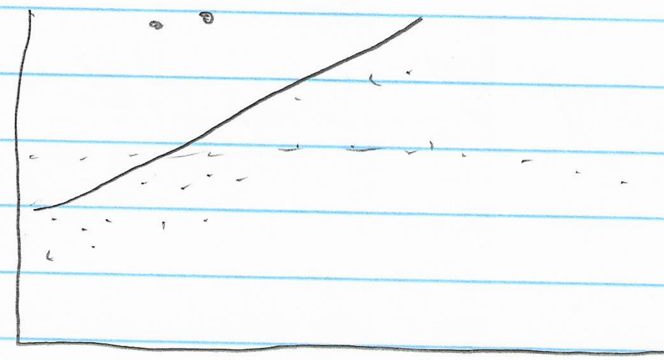
$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & \vdots & & & \\ 1 & \vdots & & \ddots & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Differientiate with respect to β, Set equal to 0, and Solve for β

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

⑤ Regularization: Ridge, Lasso, and Elastic Net



Least squares is an unbiased estimator, sometimes adding bias can help.

Regularization is an important tool that allows us to prevent overfitting

Three main types of regularization used with linear regression models

1) Ridge Regression: Coefficient Shrinkage

2) Lasso Regression: Coefficient Shrinkage + feature selection

3) Elastic Net: Combination of ridge and lasso

Ridge regression Controls the variance by imposing a constraint on the coefficients

Minimize $RSS(\beta)$

s.t. $\sum_{j=1}^{m} \beta^2 < t$ (L2 Norm)

Where t is a user defined value called the ridge parameter

The leads to a penalized loss function

$$RSS(\beta) + \lambda \sum_{j=1}^{m} \beta_j^2$$

Which has the closed form solution

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

λ controls the regularization, i.e. the size of the coefficients

$\lambda \to 0$: Least squares

$\lambda \to \infty$; β_0 (intercept only)

Least Absolute Shrinkage and Selection operator (Lasso) allows $\beta_j = 0$, i.e. feature selection

Minimize $RSS(\beta)$

s.t. $\sum_{j=1}^{m} |\beta_j| < t$ (L1 Norm)

This leads to the penalized loss function

$$RSS(\beta) + \lambda \sum_{j=1}^{m} |\beta_j|$$

No closed form solution

Large value for λ will set some coefficients to 0, i.e. feature selection