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

<table>
<thead>
<tr>
<th>Week</th>
<th>Lecture</th>
<th>Homework</th>
<th>Project</th>
<th>Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 4</td>
<td>Supervised</td>
<td>HW 1</td>
<td>Phase II</td>
<td></td>
</tr>
<tr>
<td>4/24</td>
<td>Prediction</td>
<td>Due</td>
<td></td>
<td>Due</td>
</tr>
<tr>
<td>Week 5</td>
<td>Supervised</td>
<td>HW 2</td>
<td>Phase II</td>
<td></td>
</tr>
<tr>
<td>5/1</td>
<td>Classification</td>
<td>Assigned</td>
<td>Assigned</td>
<td></td>
</tr>
<tr>
<td>Week 6</td>
<td>Unsupervised</td>
<td>HW 2</td>
<td>Phase III</td>
<td>Midterm</td>
</tr>
<tr>
<td>5/8</td>
<td>Clustering</td>
<td>due</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 7</td>
<td>Unsupervised</td>
<td></td>
<td></td>
<td>Assigned</td>
</tr>
<tr>
<td>5/15</td>
<td>Association</td>
<td></td>
<td></td>
<td>Midterm</td>
</tr>
<tr>
<td>Week 8</td>
<td>Text</td>
<td>HW 3</td>
<td>Phase IV</td>
<td></td>
</tr>
<tr>
<td>5/22</td>
<td>Mining</td>
<td>Assigned</td>
<td>Assigned</td>
<td></td>
</tr>
<tr>
<td>Week 9</td>
<td>Time Series</td>
<td>HW 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/24</td>
<td>Analysis</td>
<td>due</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 10</td>
<td>Graph</td>
<td></td>
<td>Phase V</td>
<td>Final</td>
</tr>
<tr>
<td>6/5</td>
<td>Mining</td>
<td></td>
<td>assigned</td>
<td></td>
</tr>
<tr>
<td>Finals</td>
<td></td>
<td></td>
<td>Phase V</td>
<td>Final</td>
</tr>
<tr>
<td>6/12</td>
<td></td>
<td></td>
<td>due</td>
<td></td>
</tr>
</tbody>
</table>
(3) Multiple Linear Regression

Multiple linear regression: linear regression with multiple independent variables or attributes and a single dependent variable or target.

(Multivariate linear regression refers to multiple dependent variables)

\[ \hat{y} = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} \]

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.

4) Model Training

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

![Diagram showing model training process]

- Square 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)

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

2) Minimize absolute error (L1 Norm)

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

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 solutions

L2 Norm loss function typically used because of computational efficiency

Two general approaches to minimize $\text{RSS}(\beta)$

- Normal Equations
  - closed form analytical solution for $\beta$
  - (slow for lots of attributes)
- Stochastic Gradient Descent (SGD)
  - approximate solution through gradient descent
  - (useful for large data sets)
Normal Equations:

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

\[ = (y - \beta x_0)^T (y - \beta x_0) \]

where

Training Data Set

\[
\begin{array}{ccc|c}
A_1 & A_2 & \cdots & \cdots & Y \\
\hline
\times & \times & \cdots & \cdots & \times \\
\end{array}
\]

\[ \beta_0 \text{ placeholder} \]

\[
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_m
\end{bmatrix}
\]

\[
X = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1m} \\
1 & x_{21} & x_{22} & \cdots & x_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\]

\[
Y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
\]
Differentiate with respect to $\beta$, set equal to 0, and solve for $\beta$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
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 $\text{RSS}(\mathbf{B})$

Subject to $\sum_{j=1}^{m} B_j^2 \leq t$ (L2 Norm)

Where $t$ is a user-defined value called the ridge parameter.

This leads to a penalized loss function:

$$\text{RSS}(\mathbf{B}) + \lambda \sum_{j=1}^{m} B_j^2$$

Which has the closed form solution:

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

$\lambda$ controls the regularization, i.e., the size of the coefficients.

$\lambda \to 0$: Least Squares

$A \to \infty$: $B_0$ (intercept only)
Least Absolute Shrinkage and Selection Operator (Lasso) allows $\beta_0 = 0$, i.e., feature selection

Minimize $\text{RSS}(\beta)$

\[ \text{S.t. } \sum_{j=1}^{m} |\beta_j| < \tau \quad (L1 \text{ Norm}) \]

This leads to the penalized loss function

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

No closed form solution

Large value for $\lambda$ will set some coefficients to 0, i.e., feature selection