LINEAR MODELS

REGRESSION

Linear Regression
Least Squares criterion
Regularization

Shai Fine
Agenda

• The Regression Problem
  • Classification vs. Regression

• Linear Regression
  • The Least Squares Criterion and Solution
    • Residual Analysis

• Regularized Least Squares
  • Lasso vs. Ridge Regression
  • Elastic Net
The Regression Problem

• Regression is a functional approximation of the relations between _explanatory_ variables and a _response_ variable
  • Explanatory (independent, input) variables – the variables used to explain the dependent variable
  • Response (dependent, output) variable – the variable we wish to explain

• Regression analysis is used to
  • Predict the value of a dependent variable based on the values of the independent variables
  • Explain the impact of changes in an independent variable on the dependent variable

• Key issues to consider
  • Which function?
    • In the sequel we will focus on Linear
  • How to measure approximation?
    • Most commonly - conditional expectation of the dependent variable given the independent variables
      • The average value of the dependent variable when the independent variables are fixed
Classification vs. Regression

- The difference is in the output space and the approximation measurement
  - **Classification** is a mapping from a feature space to categories (class memberships, labels), so as to minimize the probability of being wrong
    \[
    Pr_{(X,Y) \sim D}(c(x) \neq y)
    \]
  - **Regression** is a mapping from a feature space to a numeric output space (e.g. real numbers), so as to minimize the approximation error, e.g. minimize the squared error
    \[
    E_{(X,Y) \sim D}[(f(x) - y)^2]
    \]

- These two problems are highly related and one can be reduced to the other
Example – House Price Model

• Assume we want to know what the price of a house would be for a given size?
  • Given house of size $x$ what is the price $y = f(x)$ of the house?

• A random sample of 10 houses is selected as “training set”:
  • **Dependent** variable $(x) = \text{house price in } $1000s
  • **Independent** variable $(y) = \text{square feet}

<table>
<thead>
<tr>
<th>Square Feet $(x)$</th>
<th>House Price in $1000s $(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400</td>
<td>245</td>
</tr>
<tr>
<td>1600</td>
<td>312</td>
</tr>
<tr>
<td>1700</td>
<td>279</td>
</tr>
<tr>
<td>1875</td>
<td>308</td>
</tr>
<tr>
<td>1100</td>
<td>199</td>
</tr>
<tr>
<td>1550</td>
<td>219</td>
</tr>
<tr>
<td>2350</td>
<td>405</td>
</tr>
<tr>
<td>2450</td>
<td>324</td>
</tr>
<tr>
<td>1425</td>
<td>319</td>
</tr>
<tr>
<td>1700</td>
<td>255</td>
</tr>
</tbody>
</table>
Graphical Representation

- The House Price scatter plot

- Which function $y = f(x)$ to choose?

- Assuming a *linear* connection between dependent and independent variables limits the search space
A Linear Model for the House Prices

- One independent variable (house size) which “explains” the dependent variable (house price)

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Intercept
- Slope
- Coefficient
- Independent Variable
- Random Error term, Residual
- Linear component
- Random Error component
Graphical Representation

Observed Value of $y$ for $x_i$

Predicted Value of $y$ for $x_i$

Intercept = $\beta_0$

Random Error for this $x$ value

$y = \beta_0 + \beta_1 x + \epsilon$

Slope = $\beta_1$
Linear Regression Assumptions

• The underlying relationship between the $x$ variable and the $y$ variable is linear

• Error values ($\varepsilon$) are statistically independent

• The probability distribution of the errors is normal and independent of $x$
  • With mean 0 and an equal but unknown variance for all values of $x$

• $E[\varepsilon] = 0$, hence $E[y|x] = \beta_0 + \beta_1 x$ is a linear function
  • Given a training set, the goal of linear regression is to estimate this function (the regression model)
Estimated Regression Model

- The sample regression line provides an estimate of the population regression line

\[ \hat{y} = w_0 + w_1 x \]

- Estimated (or predicted) y value
- Estimate of the regression intercept
- Estimate of the regression slope
- Independent variable

The individual random error terms \( e_i \) have a mean of zero
Least Squares Criterion

- $w_0$ and $w_1$ are obtained by minimizing the sum of the squared residuals
  \[
  \sum \epsilon^2 = \sum (y - \hat{y})^2 = \sum (y - (w_0 + w_1 x))^2
  \]

- Actually this is just the same as the error we had for classification
  \[
  E(w) = \frac{1}{2} \sum (c(x) - w^T x)^2
  \]

- Hence, the same algorithms applies
  - Iterative gradient descent
Analytic Solution

\[ E(w_0, w_1) = \sum(y - (w_0 + w_1 x))^2 = \sum(y - \hat{y})^2 = \sum\varepsilon^2 \]

\[ \frac{\partial}{\partial w_0} E = -\sum 2(y - (w_0 + w_1 x)) = 0 \]
\[ \sum y - nw_0 - w_1 \sum x = 0 \]
\[ w_0 = \frac{\sum y - w_1 \sum x}{n} = \bar{y} - w_1 \bar{x} \]

\[ \frac{\partial}{\partial w_1} E = \sum 2(y - (w_0 + w_1 x))(-x) = 0 \]
\[ \sum xy - w_1 \sum x^2 - (\bar{y} - w_1 \bar{x})\sum x = 0 \]
\[ \sum xy - \bar{y}\sum x = w_1 \sum x^2 - w_1 \bar{x}\sum x \]
\[ w_1 = \frac{\sum xy - \bar{y}\sum x}{\sum x^2 - \bar{x}\sum x} = \frac{\sum(y - \bar{y})(x - \bar{x})}{\sum(x - \bar{x})^2} = \frac{\text{Cov}(y, x)}{\text{Var}(x)} \]

\[ \sum z_i = n \bar{z} = \sum \bar{z} \]
House Price Model

*Scatter Plot and Regression Line*

\[
\text{house price} = 98.24833 + 0.10977 \times \text{(square feet)}
\]

- **Intercept** = 98.248
- **Slope** = 0.10977
Interpretation of the Parameters

• The Intercept
  • $w_0$ is the estimated average value of $Y$ when the value of $X$ is zero
    • *For the house price model, no houses had 0 square feet, so $w_0 = 98.24833$ just indicates that for houses within the range of sizes observed, $98,248.33$ is the portion of the house price not explained by square feet*

• The Slope
  • $w_1$ measures the estimated change in the average value of $Y$ as a result of a one-unit change in $X$
    • *For the house price model, $w_1 = 0.10977$ tells us that the average value of a house increases by $0.10977(\$1000) = \$109.77$, on average, for each additional one square foot of size*

\[
\text{house price} = 98.24833 + 0.10977 \text{ (square feet)}
\]
Least Squares Regression Properties

• The regression line always passes through \((\bar{x}, \bar{y})\)
  
  \[ w_0 + w_1 \bar{x} = \bar{y} \]

• The sum of residuals \(\Sigma \epsilon\) from the least squares regression line is 0
  
  \[
  \frac{1}{n} \sum \epsilon = \frac{1}{n} \sum (y - (w_0 + w_1 x)) = \\
  \frac{1}{n} \sum (y - (\bar{y} - w_1 \bar{x}) - w_1 x) = \\
  \frac{1}{n} \sum (y - \bar{y}) - \frac{w_1}{n} \sum (x - \bar{x}) = 0 - 0 = 0
  \]

• The residuals \(\epsilon\) are uncorrelated with the \(x\) values
  
  \[
  \sum \epsilon x = \sum \epsilon x - \sum \epsilon \bar{x} = since \ \sum \epsilon = 0 \\
  \sum \epsilon (x - \bar{x}) = \\
  \sum (y - \bar{y})(x - \bar{x}) - w_1 \sum (x - \bar{x})^2 = 0 \\
  since \ w_1 = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (x - \bar{x})^2}
  \]

• The residuals \(\epsilon\) are uncorrelated with the fitted values \(\hat{y} = w_0 + w_1 x\)
  
  • Follow from the previous one, since \(\hat{y}\) is a linear function of the \(x\)
Correlation and Linear Regression

- At the solution
  \[ y = w_0 + w_1 x = (\bar{y} - w_1 \bar{x}) + w_1 x = \bar{y} + w_1 (x - \bar{x}) = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x}) = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \]

- The Pearson correlation coefficient \( r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \) indicates the strength of the linear relationship
  - Does not completely characterize their relationship
  - Doesn’t imply causation

- When \((X, Y) \sim N_2(\mu, \Sigma)\), i.e. bivariate normal, the conditional mean of \(Y\) is a linear function of \(X\)
  \[ \mu_{Y|X} = \mu_Y + r \frac{\sigma_Y}{\sigma_X} (X - \mu_X) \]
  - However, in general the conditional mean of \(Y\) given \(X\), is not necessarily linear in \(X\), and the correlation coefficient does not fully determine the form of \(E[Y|X]\)
Correlation and Linear Relationships

- **Perfect**: $r = -1$
- **Strong**: $r = -0.6$
- **No Relationship**: $r = 0$
- **Weak**: $r = +0.3$
- **Curvilinear**: $r = +0.8$
Anscombe's quartet

- Four datasets that have nearly identical simple statistical properties, yet appear very different when graphed
  - Each dataset consists of eleven (x,y) points

<table>
<thead>
<tr>
<th>Property (in each case)</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x</td>
<td>9</td>
</tr>
<tr>
<td>Variance of x</td>
<td>11</td>
</tr>
<tr>
<td>Mean of y</td>
<td>7.50</td>
</tr>
<tr>
<td>Variance of y</td>
<td>4.122 or 4.127</td>
</tr>
<tr>
<td>Correlation between x and y</td>
<td>0.816</td>
</tr>
<tr>
<td>Linear regression line</td>
<td>$y = 3.00 + 0.500x$</td>
</tr>
</tbody>
</table>

- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties
Residual Analysis

• Purposes
  • Examine for linearity assumption
  • Examine for constant variance for all levels of $x$
  • Evaluate normal distribution assumption

• Graphical Analysis of Residuals
  • Can plot residuals $\epsilon$ vs. $x$
  • Can create histogram of residuals $\epsilon$ to check for normality
Residual Analysis for Linearity

Not Linear

Linear
Residual Analysis for Constant Variance

Non-constant variance

Constant variance
Explained and Unexplained Variation

The *Total Variation* is made up of two parts

\[ \sum (y - \bar{y})^2 = \]

\[ \sum (y - \hat{y} + \hat{y} - \bar{y})^2 = \]
\[ \sum (y - \hat{y})^2 + 2\sum (y - \hat{y})(\hat{y} - \bar{y}) + \sum (\hat{y} - \bar{y})^2 = \]
\[ \sum (y - \hat{y})^2 + 2\sum \epsilon \hat{y} - 2\bar{y}\sum \epsilon + \sum (\hat{y} - \bar{y})^2 = \]
\[ \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 \]

**SST**
Total Sum of Squares
\[ \sum (y - \bar{y})^2 \]

**SSE**
Sum of Squares Error
\[ \sum (y - \hat{y})^2 \]

**SSR**
Sum of Squares Regression
\[ \sum (\hat{y} - \bar{y})^2 \]

**Total Variation** of the \( y \) values around their mean \( \bar{y} \)

**Unexplained Variation** attributed to factors other than the relationship between \( x \) and \( y \)

**Explained variation** attributed to the relationship between \( x \) and \( y \)
Explained and Unexplained Variation

\[ \text{SST} = \sum (y_i - \bar{y})^2 \]

\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 \]

\[ \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \]
Coefficient of Determination, $R^2$

- The *coefficient of determination* (also called *R-squared*, $R^2$)
  - The portion of the total variation in the dependent variable that is explained by variation in the independent variable
    \[
    R^2 = \frac{SSR}{SST} = \frac{\Sigma (\hat{y} - \bar{y})^2}{\Sigma (y - \bar{y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\Sigma (y - \hat{y})^2}{\Sigma (y - \bar{y})^2}
    \]
  - Where $0 \leq R^2 \leq 1$, and $R^2 = 1$ if there is a perfect linear relationship
  - In linear least squares, $R^2 = r^2$ (square Pearson correlation coefficient)

None of the variation in $y$ is explained by variation in $x$

Some but not all of the variation in $y$ is explained by variation in $x$

100% of the variation in $y$ is explained by variation in $x$
Multivariate Linear Regression Model

Examine the linear relationship between one dependent variable (y) and two or more independent variables (x_i)

**Population model:**

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \epsilon \]

**Estimated regression model:**

\[ \hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d \]
Closed-form Solution

- The parameters \( w \) are obtained by minimizing the sum of the squared residuals
  \[
  E(w) = \|Xw - y\|^2 = \sum_{(x,y)} (w^T x - y)^2
  \]

- Forming the derivative yields
  \[
  \nabla E(w) = \sum_{(x,y)} 2(w^T x - y)x = 2X^T (Xw - y)
  \]

- Setting the derivative to zero yields the necessary condition for minimum
  \[
  X^T Xw = X^T y
  \]

- Now, \( X^T X \) is square and often nonsingular and so we can solve for \( w \) uniquely as
  \[
  w = (X^T X)^{-1} X^T y
  \]

- The \( d \times d \) matrix \( (X^T X)^{-1} X^T \) is the **Pseudo Inverse of \( X \)**
Regularized Least Squares

- Consider the function

\[ E_D(w) + \lambda E_W(w) \]

Data term + Regularization term

- **Ridge Regression**
  - *sum-of-squares* error function and a *quadratic regularization term*
    \[
    \frac{1}{2} \sum_{(x,y)} (w^T x - y)^2 + \lambda w^T w
    \]
  - Which is minimized by
    \[
    w = (\lambda I + X^T X)^{-1} X^T y
    \]
  - \( \lambda \) is called the *regularization coefficient*
    - It is selected either by cross validation
      - Sometimes it can be inferred from theory by writing a generalization bound

- **The Shrinking Effect**
  - When \( \lambda \to 0 \) we resort back to the (unconstrained) Least Squares solution
  - As \( \lambda \) increased, the num. of degrees of freedom of \((\lambda I + X^T X)^{-1} X^T\) will be reduced
Other Regularization Terms

- A more generic regularized least squares expression:

\[ \frac{1}{2} \sum_{(x,y)} (w^T x - y)^2 + \lambda \sum_{j=0}^{d} |w_j|^q \]

- Lasso
- Ridge
  - Quadratic
Ridge vs. Lasso

- Lasso tends to generate sparser solutions than a Ridge (quadratic) regularization.
Elastic Net

• *Elastic Net* penalize the size of the regression coefficients based on both their $l^1$ norm and their $l^2$ norm :

$$\arg\min_w \sum_{(x,y)} (w^T x - y)^2 + \lambda_1 \sum_{j=0}^{d} |w_j| + \lambda_2 \sum_{j=0}^{d} w_j^2$$

• The $l^1$ norm penalty generates a sparse model
• The $l^2$ norm penalty:
  • Removes the limitation on the number of selected variables
  • Encourages grouping effect
  • Stabilizes the $l^1$ regularization path

Geometric Illustration of Elastic Net, Ridge regression, and LASSO

- Singularities at the vertexes (necessary for *sparsity*)
- Strict convex edges (necessary for *grouping*)
THANKS