Lecture 8: Regression Trees

Instructor: Saravanan Thirumuruganathan

Outline

- Regression
- 2 Linear Regression
- Regression Trees

Regression and Linear Regression

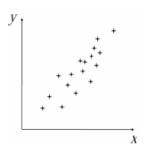
Supervised Learning

- Dataset:
 - Training (labeled) data: $D = \{(x_i, y_i)\}$
 - $x_i \in \mathbb{R}^d$
 - Test (unlabeled) data: $x_0 \in \mathbb{R}^d$
- Tasks:
 - Classification: $y_i \in \{1, 2, ..., C\}$
 - Regression: $y_i \in \mathbb{R}$
- **Objective:** Given x_0 , predict y_0
- Supervised learning as y_i was given during training

Regression

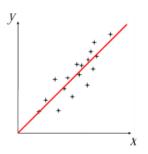
- Predict cost of house from details
- Predict job salary from job description
- Predict SAT, GRE scores
- Predict future price of Petrol from past prices
- Predict future GDP of a country, valuation of a company

Linear Regression: One-dimensional Case



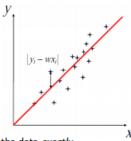
- Given: a set of N input-response pairs
- The inputs (x) and the responses (y) are one dimensional scalars
- Goal: Model the relationship between x and y

Linear Regression: One-dimensional Case



- Let's assume the relationship between x and y is linear
- Linear relationship can be defined by a straight line with parameter w
- Equation of the straight line: y = wx

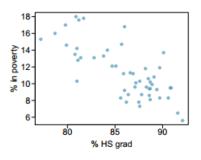
Linear Regression : One-dimensional Case



- The line may not fit the data exactly
- But we can try making the line a reasonable approximation
- Error for the pair (x_i, y_i) pair: e_i = y_i wx_i
- The total squared error: $E = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i wx_i)^2$
- The best fitting line is defined by w minimizing the total error E
- Just requires a little bit of calculus to find it (take derivative, equate to zero..)

Linear Regression: Poverty vs HS Graduation Rate

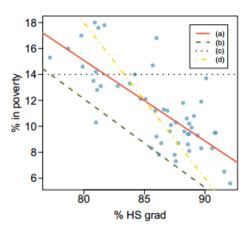
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below \$23,050 for a family of 4 in 2012).



Response variable?
% in poverty
Explanatory variable?
% HS grad
Relationship?
linear, negative, moderately strong

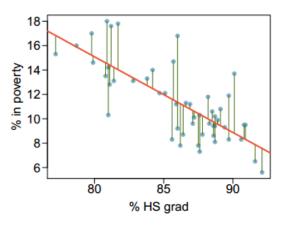
Linear Regression: Poverty vs HS Graduation Rate

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.



Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

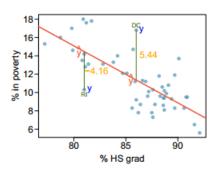


Residuals

Residual

Residual is the difference between the observed (y_i) and predicted \hat{y}_i .

$$e_i = y_i - \hat{y}_i$$



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
 - Option 1: Minimize the sum of magnitudes (absolute values) of residuals

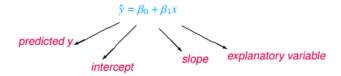
$$|e_1| + |e_2| + \cdots + |e_n|$$

2. Option 2: Minimize the sum of squared residuals - least squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

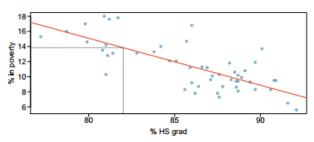
- Why least squares?
 - 1. Most commonly used
 - 2. Easier to compute by hand and using software
 - In many applications, a residual twice as large as another is usually more than twice as bad

Least Squares Line



Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value.



Linear Regression in Higher Dimensions

- Analogy to line fitting: In higher dimensions, we will fit hyperplanes
- For 2-dim. inputs, linear regression fits a 2-dim. plane to the data



- Many planes are possible. Which one is the best?
- Intuition: Choose the one which is (on average) closest to the responses Y
 - Linear regression uses the sum-of-squared error notion of closeness
- Similar intuition carries over to higher dimensions too
 - Fitting a D-dimensional hyperplane to the data
 - Hard to visualize in pictures though..
- The hyperplane is defined by parameters w (a D × 1 weight vector)

Linear Regression in Higher Dimensions

- Given training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$
- Inputs x_i : D-dimensional vectors (\mathbb{R}^D), responses y_i : scalars (\mathbb{R})
- The linear model: response is a linear function of the model parameters

$$y = f(\mathbf{x}, \mathbf{w}) = b + \sum_{j=1}^{M} w_j \phi_j(\mathbf{x})$$

- w_i's and b are the model parameters (b is an offset)
 - Parameters define the mapping from the inputs to responses
- Each φ_j is called a basis function
 - Allows change of representation of the input x (often desired)

Linear Regression in Higher Dimensions

The linear model:

$$y = b + \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) = b + \mathbf{w}^T \phi(\mathbf{x})$$

- $\phi = [\phi_1, \dots, \phi_M]$
- $\mathbf{w} = [w_1, \dots, w_M]$, the weight vector (to learn using the training data)
- We consider the simplest case: φ(x) = x
 φ_i(x) is the i-th feature of the data (total D features, so M = D)
- The linear model becomes

$$y = b + \sum_{j=1}^{D} w_j x_j = b + \mathbf{w}^T \mathbf{x}$$

 Note: Nonlinear relationships between x and y can be modeled using suitably chosen φ_i's (more when we cover Kernel Methods)

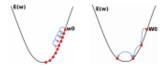
Linear Regression: Objective Function

- Parameter w that satisfies $y_i = \mathbf{w}^T \mathbf{x}_i$ exactly for each i may not exist
- . So we look for the closest approximation
- Specifically, w that minimizes the following sum-of-squared-differences between the truth (y_i) and the predictions (w^Tx_i), just as we did for the one-dimensional case:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Linear Regression: Gradient Descent based Solution

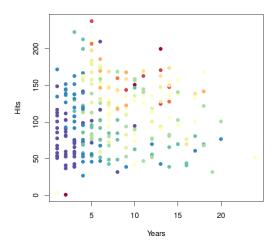
- The least-squares linear regression objective is a convex function
 - It has a unique minimum
 - Gradient descent will find the unique minimum (or get very close it to, depending in the learning rate α)
 - · For general functions, GD can only find a local minimum
- Effect of the learning rate α (left: small α , right: large α)



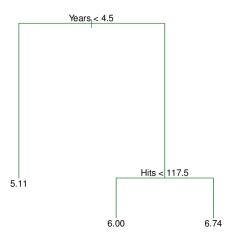
Regression Trees

Predicting Baseball salary data

Salary is color-coded from low (blue, green) to high (yellow,red)

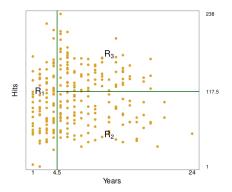


Decision tree for Baseball Salary Prediction

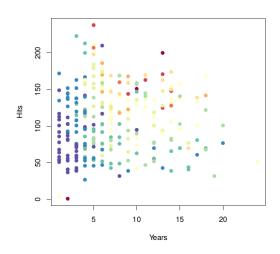


Decision tree for Baseball Salary Prediction

• Overall, the tree stratifies or segments the players into three regions of predictor space: $R_1 = \{X \mid Years < 4.5\}$, $R_2 = \{X \mid Years > =4.5, Hits < 117.5\}$, and $R_3 = \{X \mid Years > =4.5, Hits > =117.5\}$.



Interpreting the Decision Tree



Interpreting the Decision Tree

- Years is the most important factor in determining Salary, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of Hits that he made in the previous year seems to play little role in his Salary.
- But among players who have been in the major leagues for five or more years, the number of Hits made in the previous year does affect Salary, and players who made more Hits last year tend to have higher salaries.
- Surely an over-simplification, but compared to a regression model, it is easy to display, interpret and explain

High Level Idea

- Classification Tree: Quality of split measured by general "Impurity measure"
- Regression Tree: Quality of split measured by "Squared error"

High Level Idea

- We divide the feature space into J distinct and non-overlapping regions R_1, R_2, \ldots, R_J
- For every observation that falls into the region R_i , we make same prediction, which is simply the mean of the response values for the training observations in R_i
- **Objective:** Find boxes $R_1, R_2, ..., R_J$ that minimizes Residual Sum of Square (RSS)

$$RSS = \sum_{i=1}^{J} \sum_{j \in R_i} (y_j - \widehat{y_{R_i}})^2$$

where $\widehat{y_{R_i}}$ is the mean response for the training in the *i*-th box.

Building Regression Trees

- We first select the feature X_i and the cutpoint s such that splitting the feature space into the regions $\{X|X_i < s\}$ and $\{X|X_i \geq s\}$ leads to the greatest possible reduction in RSS.
- Next, we repeat the process, looking for the best attribute and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions.
- The process continues until a stopping criterion is reached; for instance, we may continue until no region contains more than five observations.

Summary

Major Concepts:

- Geometric interpretation of Classification
- Decision trees

Slide Material References

- Slides from ISLR book
- Slides by Piyush Rai
- Slides from OpenIntro Statistics book (http://www.webpages.uidaho.edu/~stevel/251/ slides/os2_slides_07.pdf)
- See also the footnotes