



Advanced Regression

CASA0006: Spatial Data Capture and Analysis
CASA0009: Data Science for Spatial Systems

Huanfa Chen

Some slides courtesy of Ed Manley & Thomas Evans

CASA0006

1 Introduction to Databases

2 Introduction to SQL

3 Advanced SQL

4 Data Munging

5 Advanced Clustering

6 Advanced Regression

7 Classification

8 Dimension Reduction

9 Unstructured Data

10 Analysis Workflow

CASA0009

1 Introduction to Databases

2 Introduction to SQL

3 Advanced SQL

4 Data Munging

5 Advanced Clustering

6 Advanced Regression

7 Interactive Viz 1: HTML + CSS

8 Interactive Viz 2: Javascript

9 Server Side Coding: Node.JS

10 Real-time data visualisation

Recap

What we already know

We can handle and clean data

Database, SQL, Python Pandas/Sklearn

We can do clustering analysis

Kmeans, DBSCAN, hierarchical clustering

Today we extend our skills in data analysis, exploring the use of **regression methods** for analysing data

Data Analysis

Picking an Approach

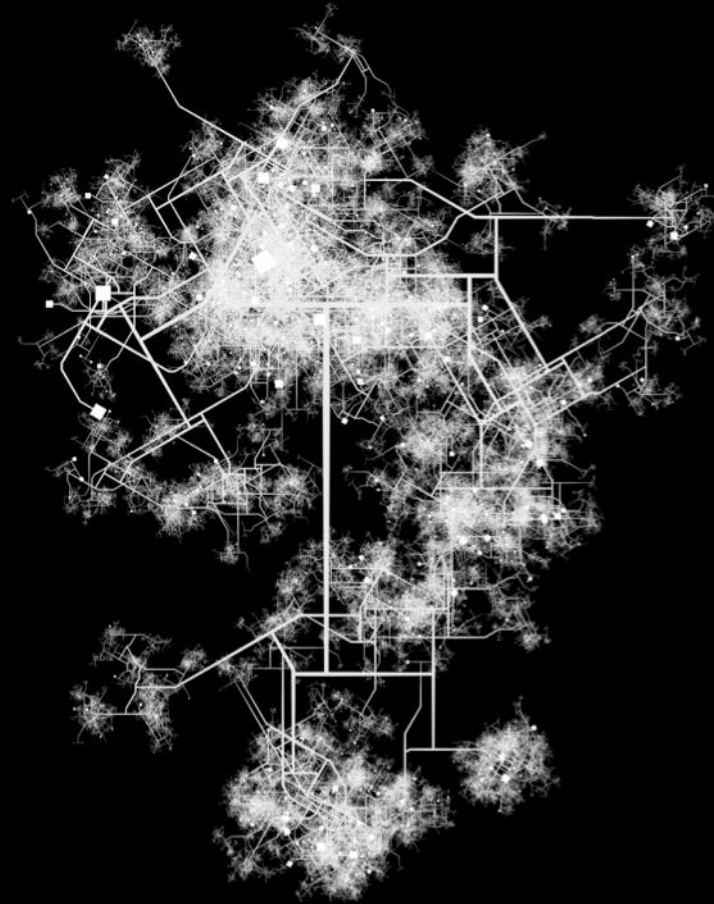
The approach to take towards analyzing your data depends on what you want to understand from it

| Input Dataset | → | Method | → | Output |
|---------------|---|--------------------------|---|--------------------------------|
| | | Clustering | → | Creation of Groupings |
| | | Regression | → | Identify Data Relationships |
| | | Classification | → | Identify Discrete Class |
| | | Dimensionality Reduction | → | Understand Influential Factors |
| | | Association Rule Mining | → | Identify Dependencies |
| | | Anomaly Detection | → | Identify Outliers |

Unsupervised: no ground truth

Supervised: with ground truth

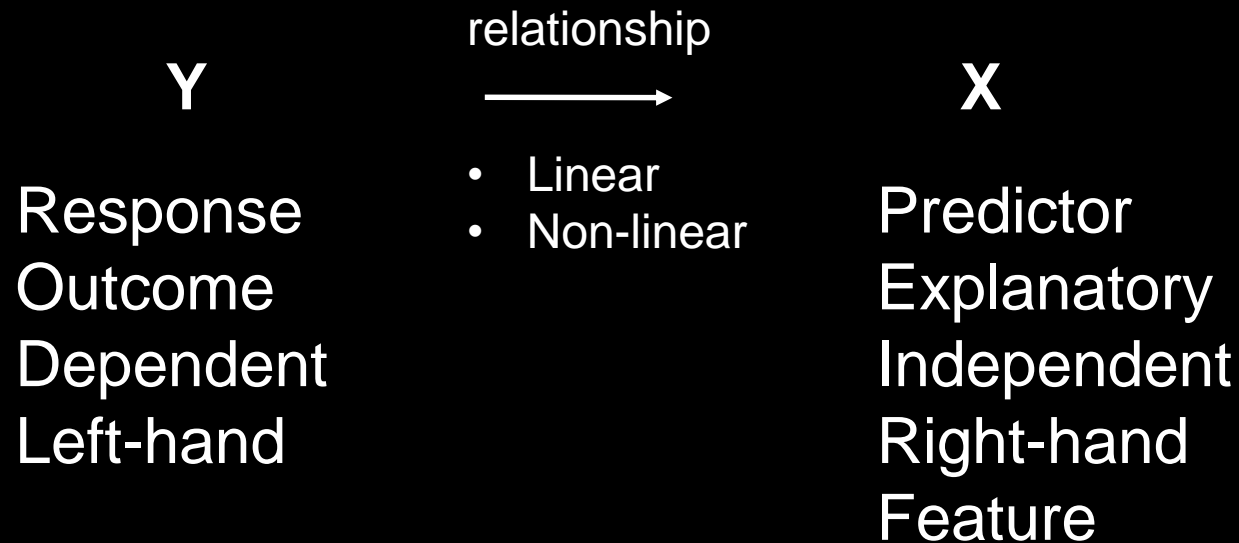
Outline



1. Introduction to regression
2. Linear regression
3. VIF and Lasso
4. Regression tree
5. Random forest

Regression

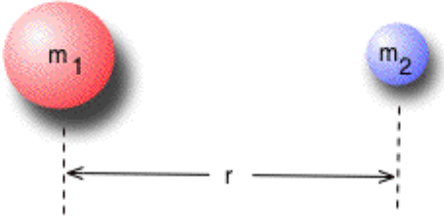
Finding relationship between y and x variables



Types of relationship

- Deterministic (or functional) relationship

Law of gravity



$$F = \frac{G m_1 m_2}{r^2}$$

F = The pull of gravity [N]
 m_1 = One object's mass [kg]
 m_2 = Other object's mass [kg]
 r = distance between the object's
 G = the universal constant(
do not memorize it
 $G = 6.67 \times 10^{-11} \text{ (N}\cdot\text{m}^2)/(\text{kg}^2)$

Ideal gas equation

The Ideal Gas Equation

$$pV = nRT$$

- p = pressure (Pa)
- V = volume (m^3)
- n = number of moles
- R = the gas constant = $8.31 \text{ J K}^{-1} \text{ mol}^{-1}$
- T = temperature (K)

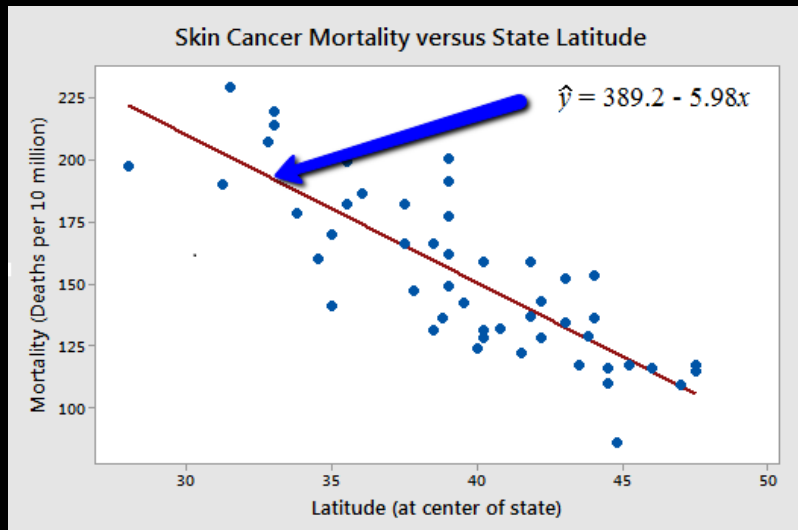
The equation exactly describes the relationship between the variables.

Not the interest of this lecture. Instead, we are interested in statistical relationships, which is not perfect.

Types of relationship

- Statistical relationship (often imperfect)

Showing some 'trend',
exhibiting some 'scatter'



A picture of regression



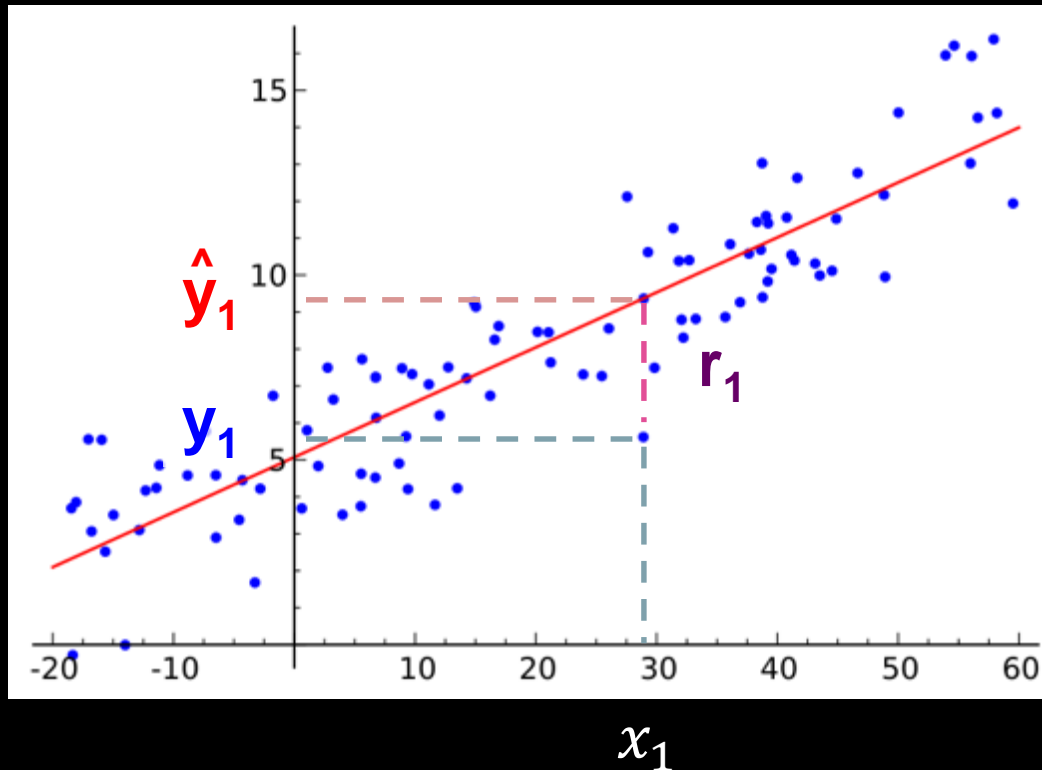
Two goals of regression

- Extracting information on the 'nature'
- Predicting the y in response to some x

These two goals can be conflicting.

Some models are easy to understand but not good at predicting, or vice versa.

Linear relationship



The linear relationship:

$$\hat{y} = mx + c$$

Residual: $r_1 = \hat{y}_1 - y_1$

Regression minimises: $SSE = r_1^2 + r_2^2 + \dots + r_n^2$

Linear relationship

How well does the model represent the data

Coefficient of Determination: R^2 Value

Variation around the line: $\sum_i r_i^2 = \sum_i (\hat{y}_i - y_i)^2$

Also called SSE (sum of squared error)

Total variation: $\sum_i (y_i - \bar{y})^2$ (\bar{y} is mean of original y)

Also called SST (total sum of squares)

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

“The higher R-squared, the better the model explains the data”

Linear relationship

What to report and visualise

Fitted Equation: $\hat{y} = mx + c$

p-value: Is the result significant?

R-Squared Value: Goodness of fit

Data Scatter Plot: With fitted line (only for simple linear regression)

Residuals vs Fits Plot: LINE conditions

Python

- The sklearn package can conduct linear regression analysis, but it does not provide a summary function.
- If you want a summary of the linear regression, use the statsmodels package.

Linear relationship

Interpretations

Example: predicting the daily number of rented bikes, given weather and calendar information

| Predictor | Type | Weight |
|---------------------------|-------------|---------|
| (Intercept) | | 2399.4 |
| seasonSUMMER | Categorical | 899.3 |
| seasonFALL | | 138.2 |
| seasonWINTER | | 425.6 |
| holidayHOLIDAY | | -686.1 |
| workingdayWORKING DAY | | 124.9 |
| weathersitMISTY | | -379.4 |
| weathersitRAIN/SNOW/STORM | | -1901.5 |
| temp | Numerical | 110.7 |
| hum | | -17.4 |
| windspeed | | -42.5 |
| days_since_2011 | | 4.9 |

Using categorical data in regression

| Indicator | One-hot encoding → | isWorkingDay |
|-----------|--------------------------|--------------|
| 'Workday' | | 1 |
| 'Weekend' | | 0 |
| 'Weekend' | | 0 |
| 'Workday' | | 1 |

- 'Weekend' as reference category
- This technique applies for multiple categories

Linear relationship

Interpretations

Example: predicting the daily number of rented bikes, given weather and calendar information

| | Type | Weight |
|---------------------------|-------------|---------|
| (Intercept) | | 2399.4 |
| seasonSUMMER | Categorical | 899.3 |
| seasonFALL | | 138.2 |
| seasonWINTER | | 425.6 |
| holidayHOLIDAY | | -686.1 |
| workingdayWORKING DAY | | 124.9 |
| weathersitMISTY | | -379.4 |
| weathersitRAIN/SNOW/STORM | | -1901.5 |
| temp | Numerical | 110.7 |
| hum | | -17.4 |
| windspeed | | -42.5 |
| days_since_2011 | | 4.9 |

- **‘workingdayWORKING DAY’**
When it is working day, the predicted number of bicycles is 124.9 higher compared to weekend, *given all other features remain fixed*

- **‘temp’**
An increase of the temperature by 1 degree Celsius increases the predicted number of bicycles by 110.7, *when all other features remain fixed*

Linear relationship

Interpretations

- Notes
 - **Linear regression is not causality analysis.** You can't say a bad weather is the reason for the decrease of rented bicycle use.
 - **Weight does not imply importance.** You can't say windspeed (weight of '-42.5') is more important than humidity (-17.4) in the model.
 - Be careful to interpret a regression model when some variables are log-transformed.

Linear relationship

Advantages

- **Transparent and intuitive:** the prediction being a weighted sum of predictors.
- **Widely accepted for inference and predictive modelling:** in many subjects and fields. Used as a starting point for modelling.
- **Mathematically, it is guaranteed to find optimal weights:** given that the assumptions have been met
- **A large toolbox:** solid statistical theory, confidence intervals, tests, extensions (e.g. generalised linear model). One important extension is Lasso, with which we can ensure that the number of features used remains small.

Linear relationship

Problems

- **Multicollinearity:** when some predictors are highly correlated, variance of the coefficient is large and the model becomes unstable and unreliable (solution: VIF, Lasso)
- **Difficulty to account for non-linearity or interaction:** these have to be hand-crafted and explicitly given to the model as an input feature. In the bike example, to account for high-temp and high-humidity weather, we need to create an interaction term $\text{temp} \times \text{hum}$
- **Low predictive performance:** the relationship that can be learned are restricted, which usually simplify the complex reality.

Variance Inflation Factors (VIF)

- Given y and x_1, x_2, \dots, x_p , the VIF for the x_k variable is

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 value obtain by regressing the x_k on the remaining x variables:

$$x_k = \sum_{i=1}^{k-1} b_i x_i + \sum_{i=k+1}^p b_i x_i$$

- The larger VIF_k , the higher multicollinearity, as x_k can be largely represented by a linear combination of the other variables and thus x_k is redundant.

VIF for variable selection

1. Initialise \underline{L} as the list of predictor variables. (*HINT*: the response variable is not needed for VIF)
2. Calculate the VIF for each variable in \underline{L} . (*HINT*: the order of computing VIF is irrelevant).
3. If the highest VIF is larger than the threshold, remove the corresponding variable from the list \underline{L} . A threshold of 5 is often used.
4. Repeat Step 2-3, until no VIF is larger than the threshold.
5. Output \underline{L} .

Optional: more about VIF:

<https://online.stat.psu.edu/stat501/lesson/12/12.4>

Lasso

“SSE: sum of squared errors”

Perform feature selection using a penalty function

Linear regression: $y = \beta x + \beta_0$

$$\text{Obj: } \min_{\beta, \beta_0} \text{SSE} = \min_{\beta, \beta_0} \left\{ \frac{1}{N} \sum_{i=1}^N (y_i - \beta_0 - \beta x_i)^2 \right\}$$

With penalty term:

$$\text{Lasso: } \min_{\beta, \beta_0} \{ \text{SSE} + \lambda \sum ||\beta||_1 \}$$

• $||\beta||_1$ is called L1-norm
e.g. given a linear model $y = -3x_1 + 4x_2 + 5$,
 $||\beta||_1 = |-3| + |4| = 7$

- λ is a parameter controlling the strength of penalty effect.
- As λ increases, the fewer predictors are present in the model, as their weights become zero.
- λ needs to be tuned (using Lasso path).

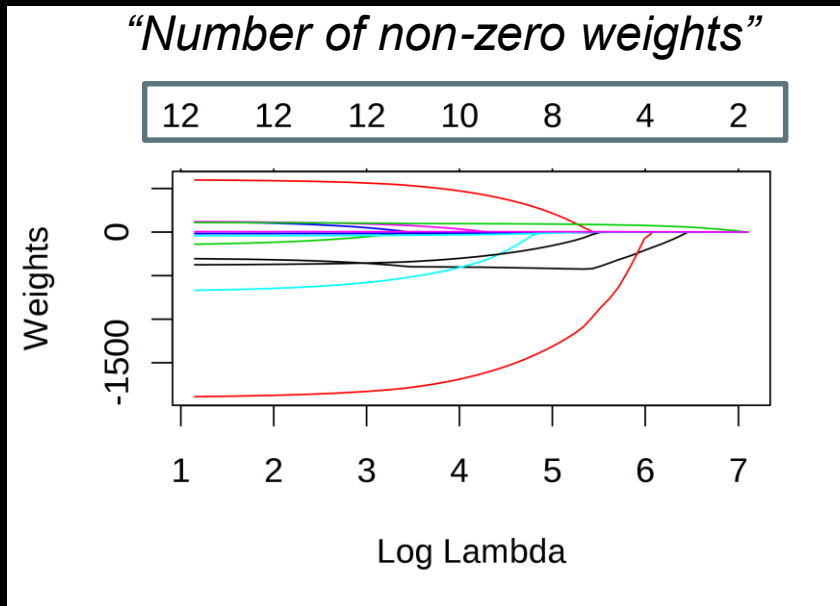
Some methods such as ridge and elastic net are similar with Lasso, but the penalty term is a bit different.

Lasso

Perform feature selection using a penalty function

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_0$$

Lasso: $\min_{\beta, \beta_0} \{RSS + \lambda \sum ||\beta||_1\}$

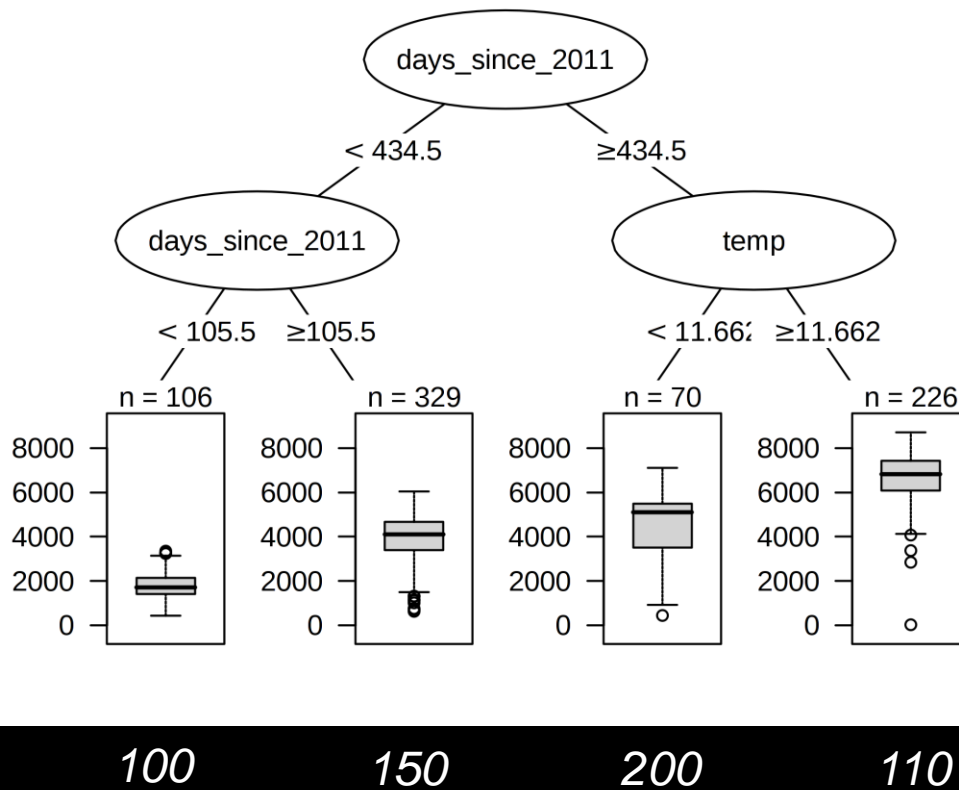


- Each β_i is a function of the lambda, and is represented as a curve in the figure.
- The weight decreases as lambda increases.
- Advantages of Lasso
 - It can be automated
 - It considers all predictors simultaneously
 - The penalty term can be controlled by lambda

Regression Trees

Split the data automatically and iteratively and create local models

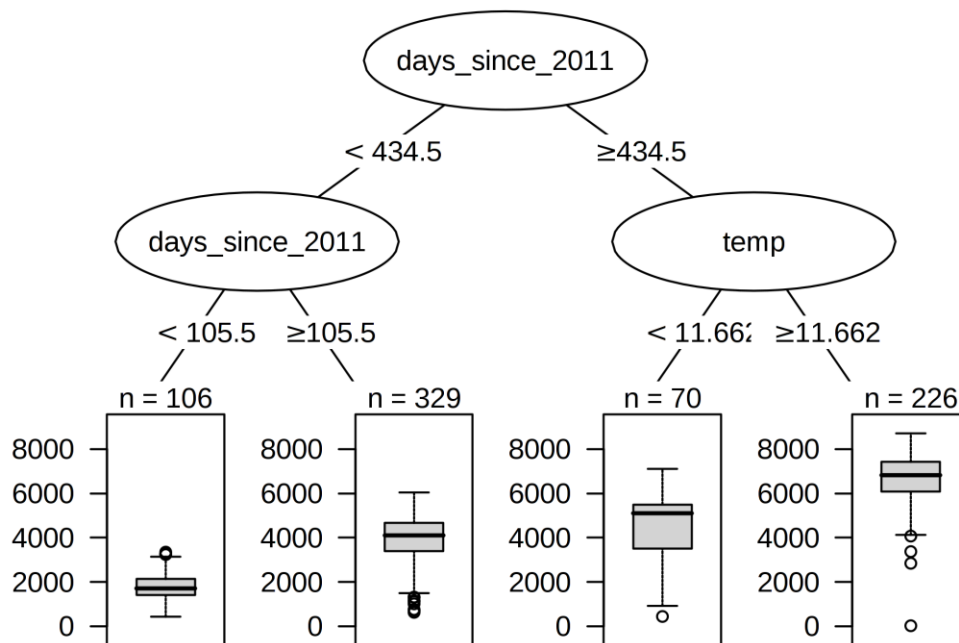
Example: predicting the bike rental using the days_since_2011 and temp



- Through splitting, different subsets of the data are created.
- Final subsets are called terminal or leaf nodes, and the average outcome is used for a leaf node.

Regression Trees

Using Regression Tree for prediction



100

150

200

110

Avg bike rental of each terminal

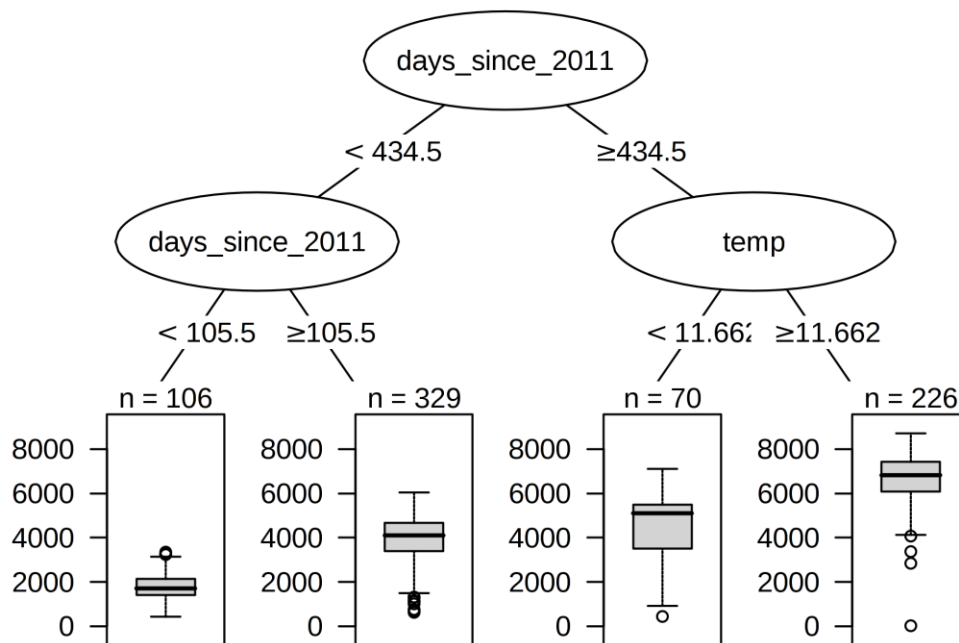
Instance (days_since_2011, temp)

(435, 12): predicted_bike_rental = ??

(434, 12): predicted_bike_rental = ??

Regression Trees

Illustrations



100

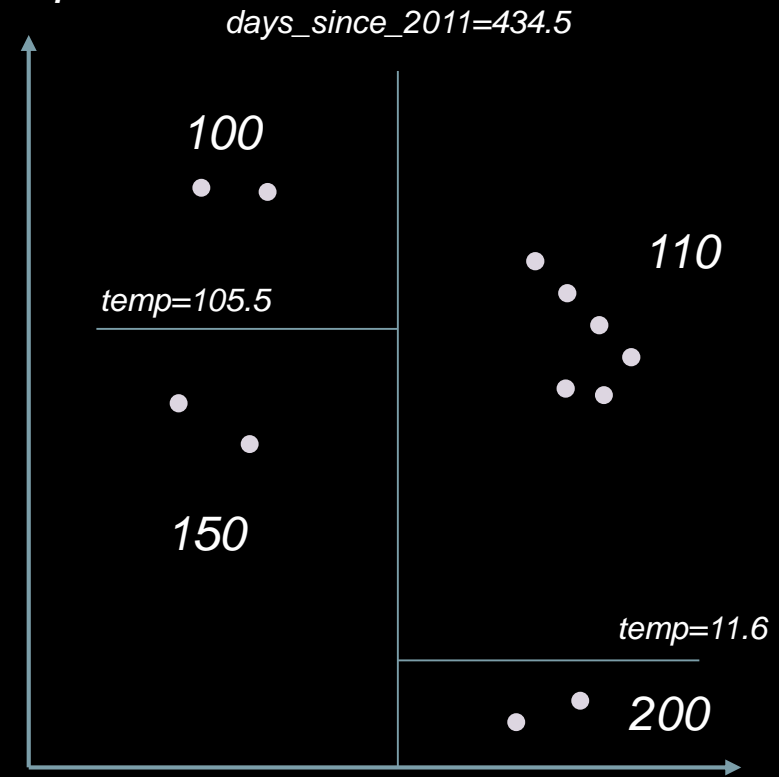
150

200

110

Avg bike rental of each terminal

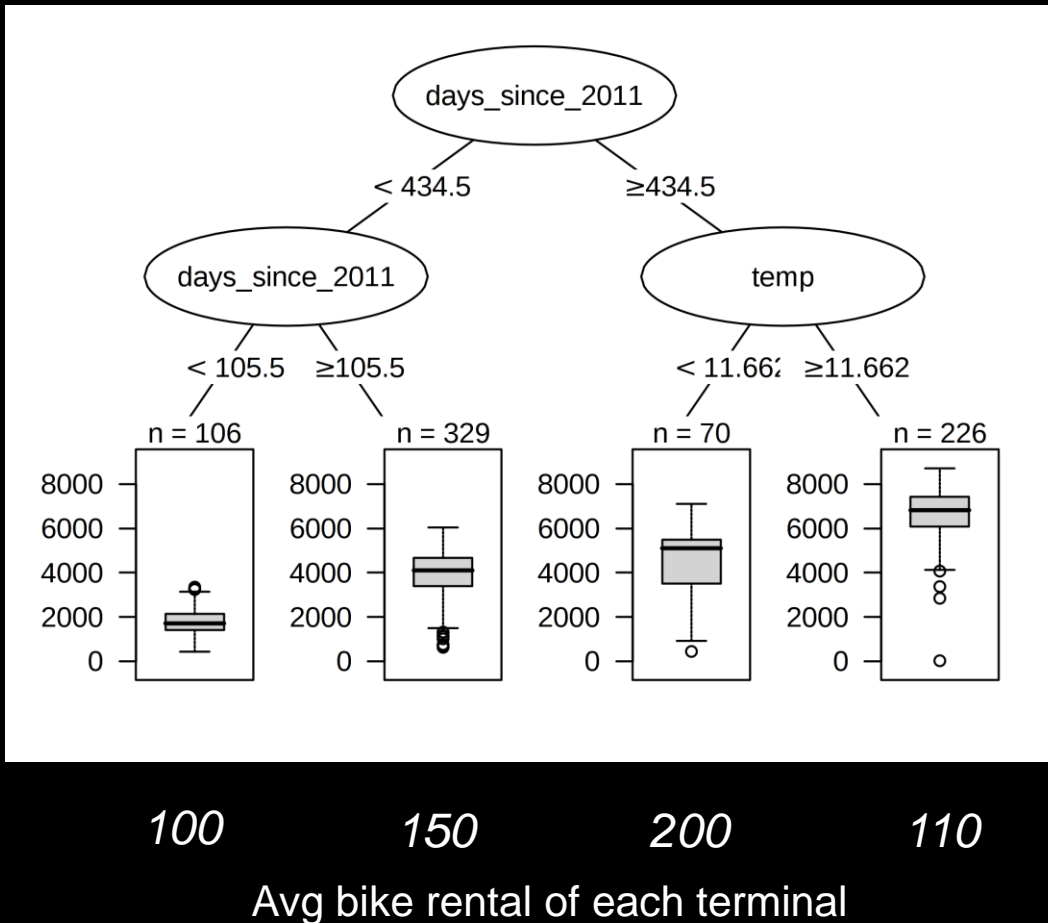
temp



days_since_2011

Regression Trees

Growing a tree by searching for optimal split



Defining stopping criteria (when to stop splitting a node)

- Max tree depth
- Minimal instances in a node

Splitting a node

- Is any stopping criteria met? If yes, don't split.
- If no, search all (feature, split_value) for the split that maximises the difference between two new nodes

Growing a tree

- Iteratively split new nodes until all nodes stop splitting.

Regression Trees

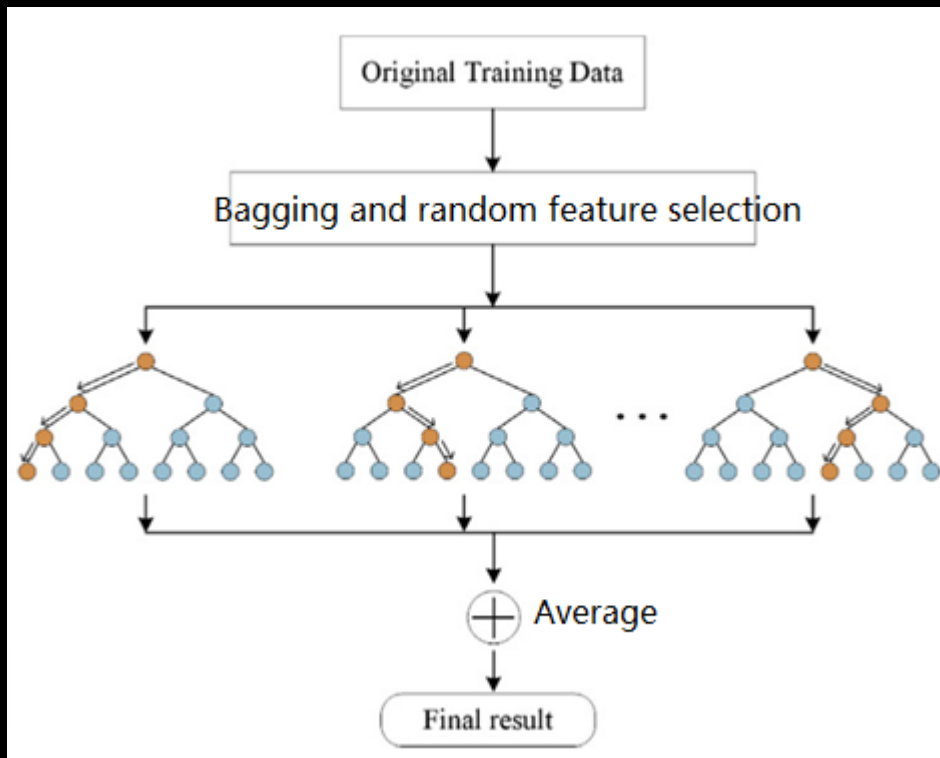
Split the data many times based on cutoff values

- Advantages of regression trees
 - Able to work with categorical and numeric data
 - Relatively easy to understand
 - No assumptions of data distribution and no transformations needed
- Disadvantages
 - Lack of smoothness. Slight changes in the predictors can have a big impact on the response
 - Tendency of overfitting, meaning that the tree is unable to generalize to new data (solution: random forest)
- Extra notes
 - Regression trees are a type of decision tree. Decision trees can be used for both regression and classification.

Random Forest

Creating many trees and combining their response

- A single tree may be overfitting and does not perform well on new datasets.
- A good solution is to randomly grow a bunch of random and different trees.



- Given an input, the response is a combination (e.g. average) of the output of all trees.

Amended from source image:

Cheng, L., Chen, X., De Vos, J., Lai, X., and Witlox, F. (2019)
Applying a random forest method approach to model travel
mode choice behavior. *Travel Behaviour and Society*

Growing a different tree

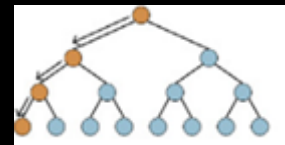
- Each tree is grown by
 - Resampling training data with replacement
 - Sampling a random selection of predictors
- This guarantees that each tree is different.

| Index | x1 | x2 | x3 |
|-------|----|-----|----|
| 1 | 2 | 1.0 | 2 |
| 2 | 3 | 1.5 | 3 |
| 3 | 5 | 2.0 | 4 |
| 4 | 4 | 2.6 | 6 |

Sampling:
2 features,
7 samples

| Index | x1 | x3 |
|-------|----|----|
| 1 | 2 | 2 |
| 2 | 3 | 3 |
| 3 | 5 | 4 |
| 4 | 4 | 6 |
| 2 | 3 | 3 |
| 4 | 4 | 6 |
| 2 | 3 | 3 |

Train a tree



Random Forest

Pros and cons

- Advantages of random forest
 - Able to work with categorical and numeric data
 - No assumptions on data distribution
 - Good predictive performance
 - Good generalisation
- Disadvantages
 - Difficult to interpret ('black-box' methods). There are some methods to interpret RF but they are not very intuitive.

Take-home message

Regression

Regression explores the relationship between predictors and a continuous response.

It is a form of supervised learning.

- **Linear regression:** fit linear relationship
- **Lasso:** add penalty terms so they can reduce number of predictors in linear models. (So is VIF)
- **Regression tree:** creates a tree structure by dividing predictors into subsets
- **Random forest:** creates many trees randomly and combines their predictions



**Thank You
Questions?**

Huanfa Chen

huanfa.chen@ucl.ac.uk

Workshop

Data Mining

- This workshop will focus on using regression methods to analyse a multivariate dataset
- You'll continue to use the scikit-learn Python library.
- Don't worry, you're not expected to understand all of the maths and computation, only the usefulness, distinctions and application of these approaches.
- **Download this week's Python Notebook from Moodle, open it in Anaconda and work through**