Introduction to Machine Learning

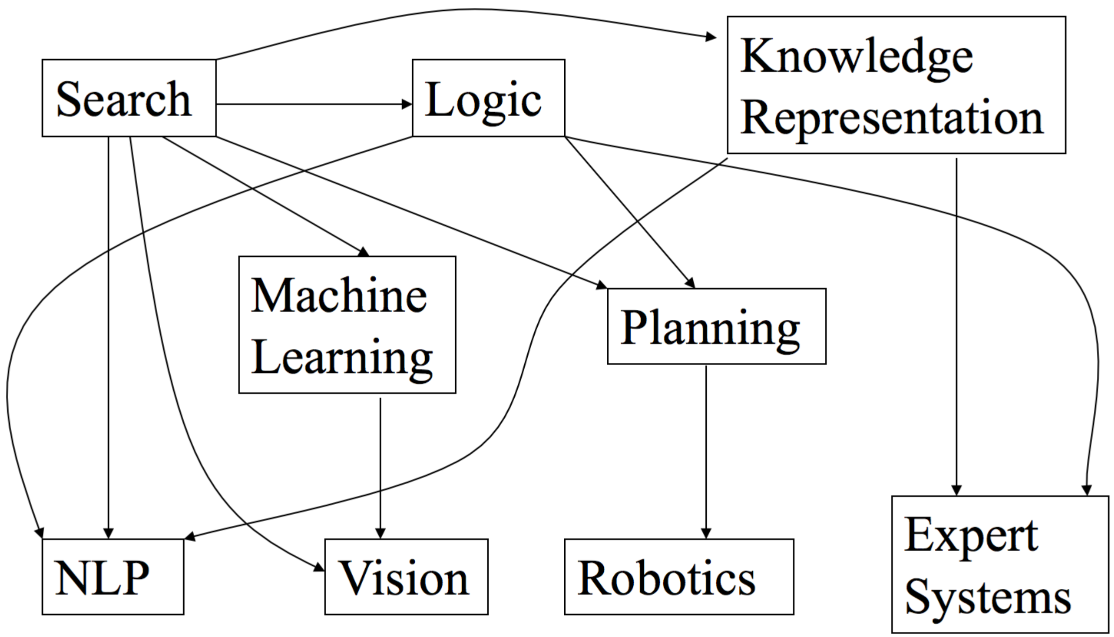
Raphael Cobe

December 11th, 2018

Artificial Inteligence - A.I.

Inside *AI*

- Many related areas;



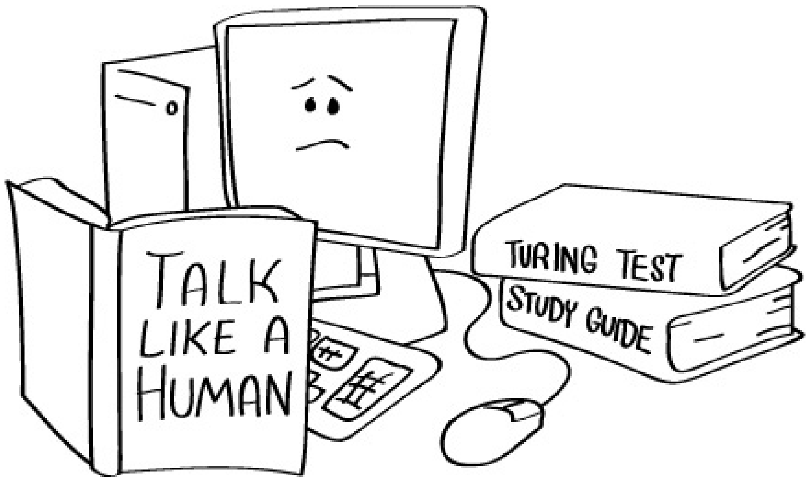
What is AI?

- Making computers that think?
- The automation of activities we associate with human thinking, like decision making, learning … ?
- The art of creating machines that perform functions that require intelligence when performed by people ?
- The study of mental faculties through the use of computational models ?
- Anything in Computing Science that we don’t yet know how to do properly ? (!)

What is AI?

Systems that think like humans?	Systems that think rationally?
Systems that act like humans?	Systems that act rationally?

Turing Test



Turing Test

- Uses the “Imitation Game”
- Usual method:
  - Three people play (man, woman, and interrogator)
  - Interrogator determines which of the other two is a woman by asking questions
    - Example: How long is your hair?
  - Typewritten or repeated by an intermediary

Turing Test

- Requires success in:
  - Natural language processing: communicate with the interrogator;
  - Knowledge representation: store and retrieve what it knows;
  - Automated reasoning: use the stored information to answer questions and to draw new conclusions;
  - Machine learning: adapt to new circumstances and to detect and extrapolate patterns

Turing Test

- Not a big effort to try to pass the Turing test;
- Acting like a human:
  - When AI programs have to interact with people
  - e.g. when an expert system explains how it came to its diagnosis;
  - e.g. natural language processing system has a dialogue with a user.
- When programs must behave according to certain normal conventions of human interaction?
- Underlying representation and reasoning may or may not be based on a human model.

What is inside A.I.?

- Search (includes Game Playing).
- Representing Knowledge and Reasoning with it.
- Planning;
- Learning;
- Natural language processing.
- Interacting with the Environment (e.g. Vision, Speech recognition, Robotics)

Machine Learning

Types of Learning

- Supervised learning
  - Training data includes desired outputs
- Unsupervised learning
  - Training data does not include desired outputs
- Semi-supervised learning
  - Training data includes a few desired outputs
- Reinforcement learning
  - Rewards from sequence of actions

Supervised Learning

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud
- Regression, Classification or a Mix.

Unsupervised Learning

- Learning **what normally happens**
  - Based on the underlying (unknown) data structure;
- Not based on examples;

“we have a bunch of data and we want to know how to separate it into meaningful groups”

- Clustering: Grouping similar instances
- Other applications: Summarization, Association Analysis

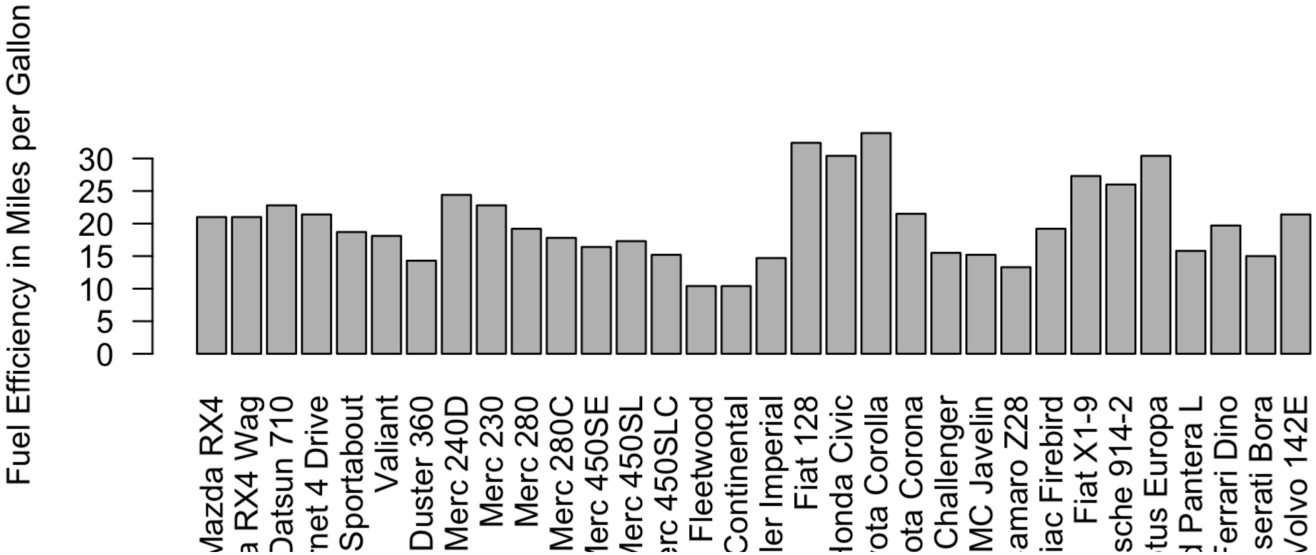
R and Machine Learning

- lot of good machine learning packages;
- Each package in R is like its own mini-ecosystem that requires a little bit of understanding first before going all out with it;
- Sometimes, you might need to select a package you're less familiar with for its specific functionality and leave your favorite one behind;

Why Build Models?

- Fundamental aspect of machine learning;
- Offer a static picture of what the data shows;
- A report is a static entity that doesn't offer an intuition as to how it evolves over time.
    - E.g.: "A distribution of vehicle fuel efficiency based on the built-in mtcars dataset, found in R".

Why Build Models



- A model is any sort of function that has predictive power!

Why Build Models

How do we turn this boring R eport into something more useful?

How do we bridge the gap between reporting and machine learning?

Regression

Linear Regression

- Regression analysis is used to describe the relationship between:
- A single response variable: Y ; and
- One or more predictor variables: $X_1, X_2, \ldots, X_n$
  - n = 1: Simple Regression
  - n > 1: Multivariate Regression

Linear Regression

- Model a continuous variable Y as a mathematical function of one or more X variable(s);

$$\widehat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 (Intercept): point in which the line intercepts the y-axis;
- β_1 (Slope): increase in Y per unit change in X.

Simple Linear Regression

We want to find the equation of the line that *best* fits the data. It means finding b_0 and b_1 such that the fitted values of y_i, given by

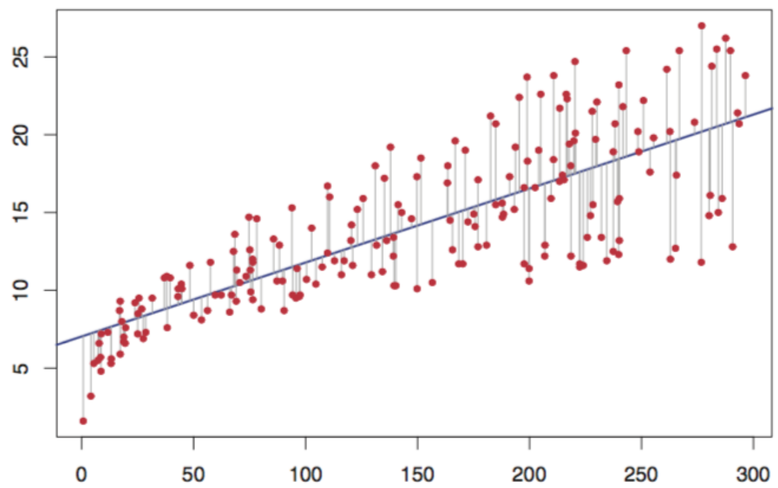
$$\hat{y}_i = b_0 + b_1 x_i$$

are as *close* as possible to the observed values y_i.

Simple Linear Regression

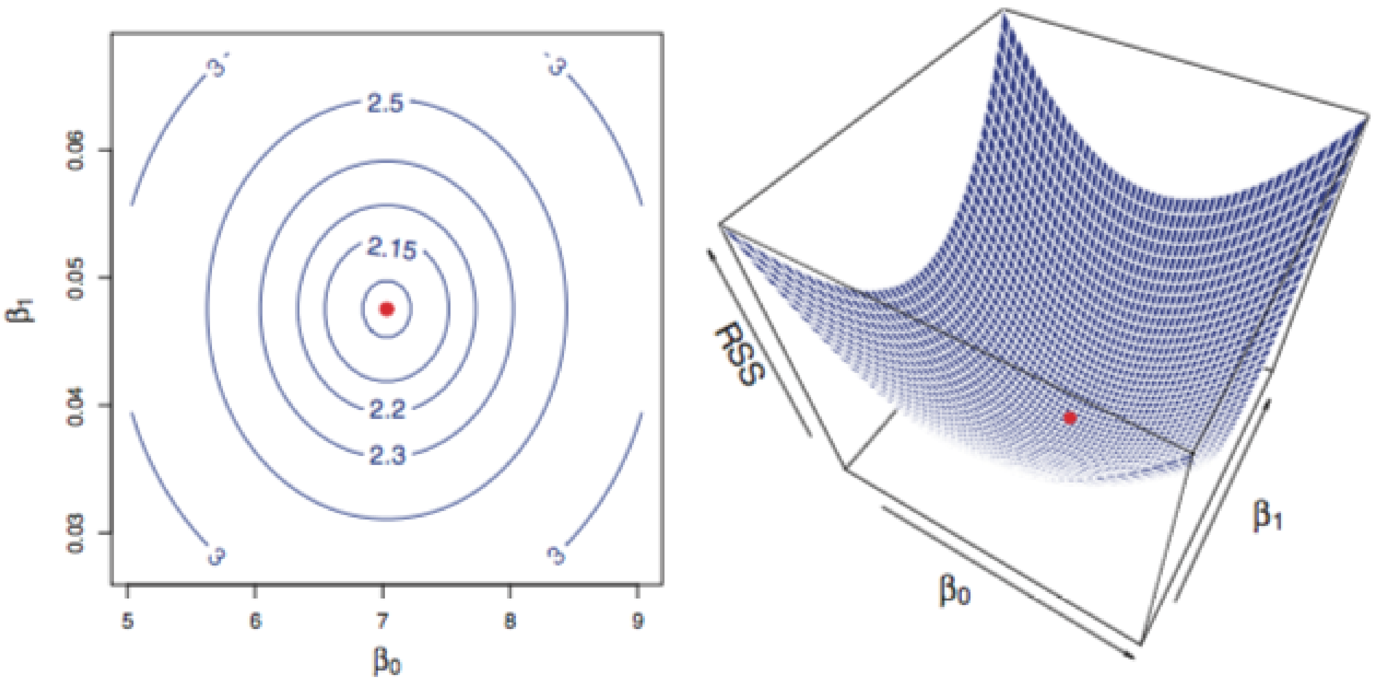
Residuals

- The difference between the observed value y_i and the fitted value $\hat{y}_i$: $e_i = y_i - \hat{y}_i$



Simple Linear Regression

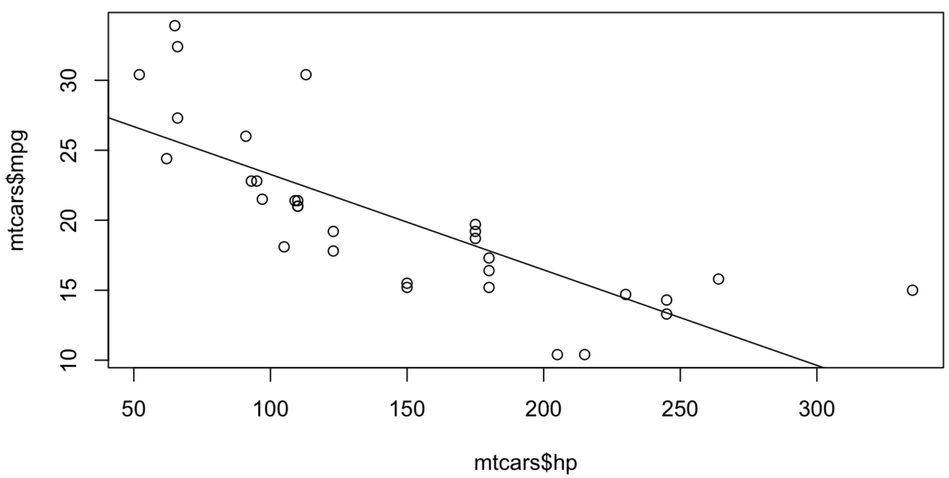
A usual way of calculating b_0 and b_1 is based on the minimization of the sum of the squared residuals;



Simple Linear Regression

Regression in R with `lm()` function:

```
cars.lm1 <- lm(mpg ~ hp, data = mtcars);
plot(x=mtcars$hp, y=mtcars$mpg);
abline(cars.lm1);
```



Simple Linear Regression

Check with `summary()` some details of the model:

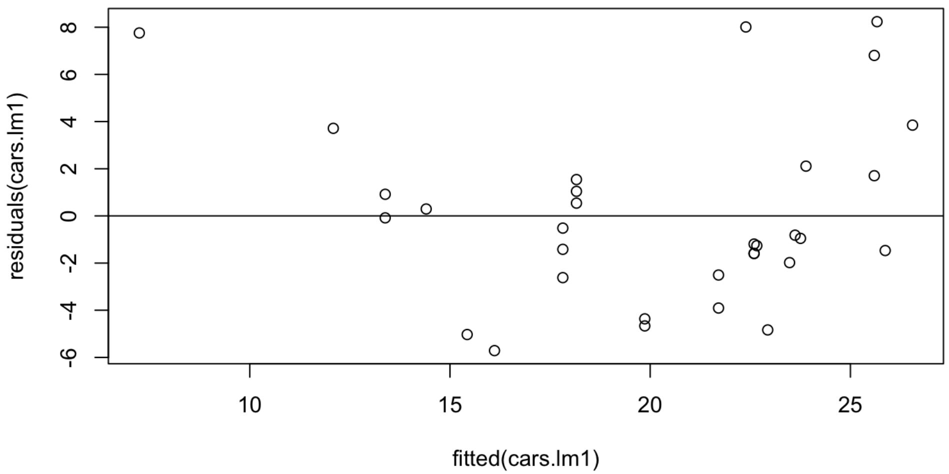
```
summary(cars.lm1);
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Simple Linear Regression

Obtain fitted values with `fitted()`:

```
plot(fitted(cars.lm1), residuals(cars.lm1));
abline(a=0,b=0); # Intercept and Slope
```



Assessing the Model quality

Residual Standard Error - *RSE*:

- Derived from the Residual Sum of Squares - RSS;
- Associated with each observation is an error term ϵ:

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

- Even if we knew the true regression line, we would not be able to perfectly predict Y from X;
- The RSE is an estimate of the standard deviation of ϵ;
- The average amount that the response will deviate from the true regression line

Assessing the Model quality

Residual Standard Error - *RSE*:

- a measure of the lack of fit of the model to the data;
- If the predictions obtained using the model are very close to the true outcome:
  - RSE will be small, and we can conclude that the model fits the data very well;
- If $\hat{y}_i$ is very far from y_i for one or more observations, then:
  - The RSE may be quite large, indicating that the model doesn???t fit the data well;

Assessing the Model quality

R^2:

- Provides an alternative measure to RSE;
- "*Unitless*";
- The proportion of variance explained;
- Always takes on a value between 0 and 1;
- Independent of the scale of Y;

Assessing the Model quality

R^2:

- Statistic close to 1:
  - A large proportion of the variability in the response has been explained by the regression.
- A value near 0:
  - Indicates that the regression did not explain much of the variability in the response;
- it can still be challenging to determine what is a good R^2 value;
  - depend on the application;

Linear Regression

Quick Challenge: Investigate if the Car Weight has some impact on its Fuel Efficiency. How good is your model?

Multiple Linear Regression

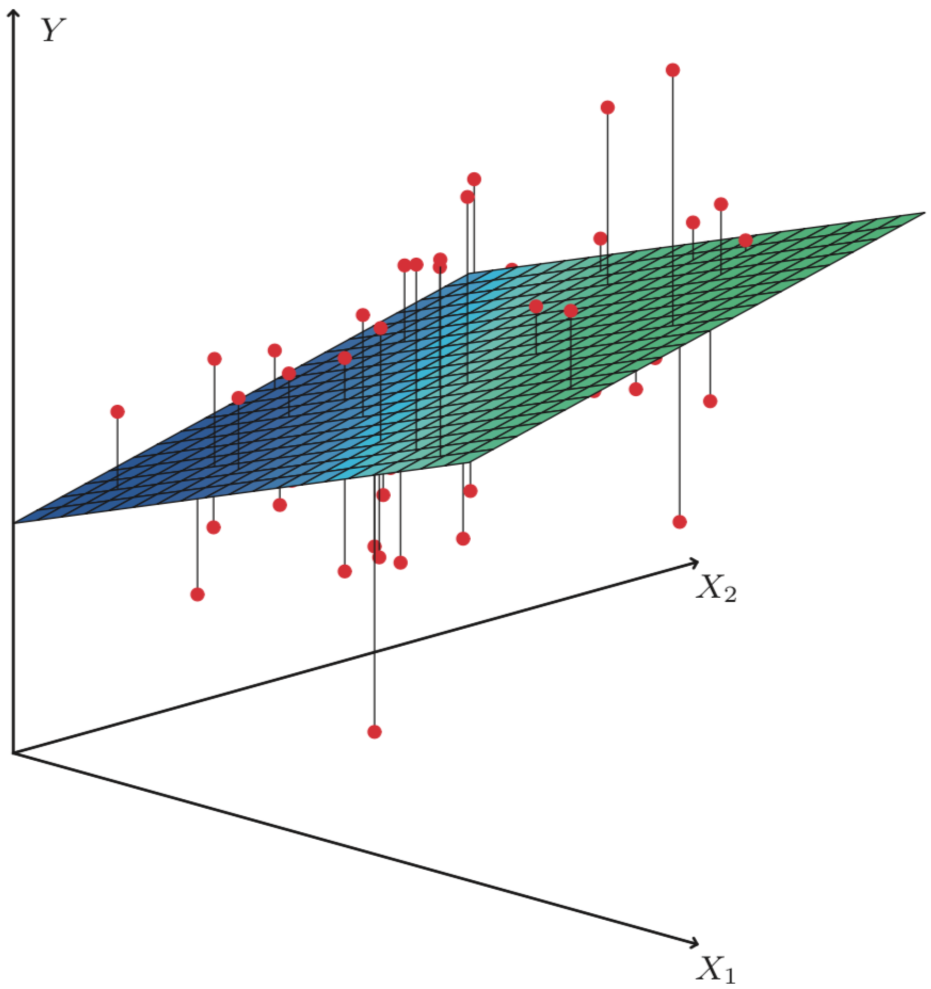
- Extend the simple linear regression model to directly accommodate multiple predictors:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Multiple Linear Regression

- The values of `radio` and `TV` better explain the variance;
- Fitting a separate simple linear regression model for each predictor is bad;
- Each of the three regression equations ignores the other two media;

Multiple Linear Regression



Multiple Linear Regression

In R (define a more complex formula):

```
formula = sales ~ TV + radio + newspaper;
```

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

Quick Challenge

- Use the `Prestige` dataset in `cars` dataset;
- Answer the questions:
  1. Is at least one of the predictors $X1$, $X2$, . . . , Xp useful in predicting the response?
  2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
  3. How well does the model fit the data?

Polynomial Regression

- Fitting a higher degree function to the data;
- Differs from the simple linear cases by having multiple degrees for each feature in the dataset;

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \ldots + \beta_n X_n^n$$

Polynomial Regression in R

Use the `poly()` function while defining the formula:

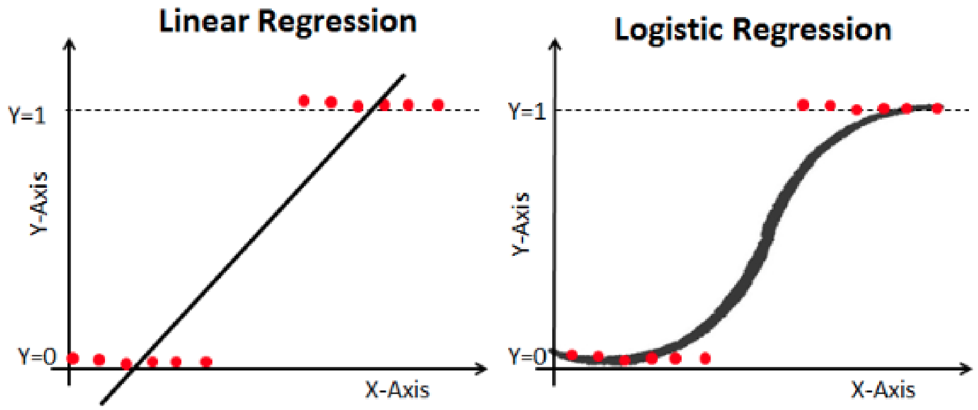
```
lm2 <- lm(pop$uspop ~ poly(pop$year, 2))
```

Classification

Binary Classification

- You want to see if a given data point is of a categorical nature instead of numeric;
- e.g., Discover whether or not a car is automatic by examining its Fuel Efficiency;
- Fitting a linear regression model to this data would not work, because we cannot have half a transmission value;
- Logistic regression model;
    - **Logit Function**;

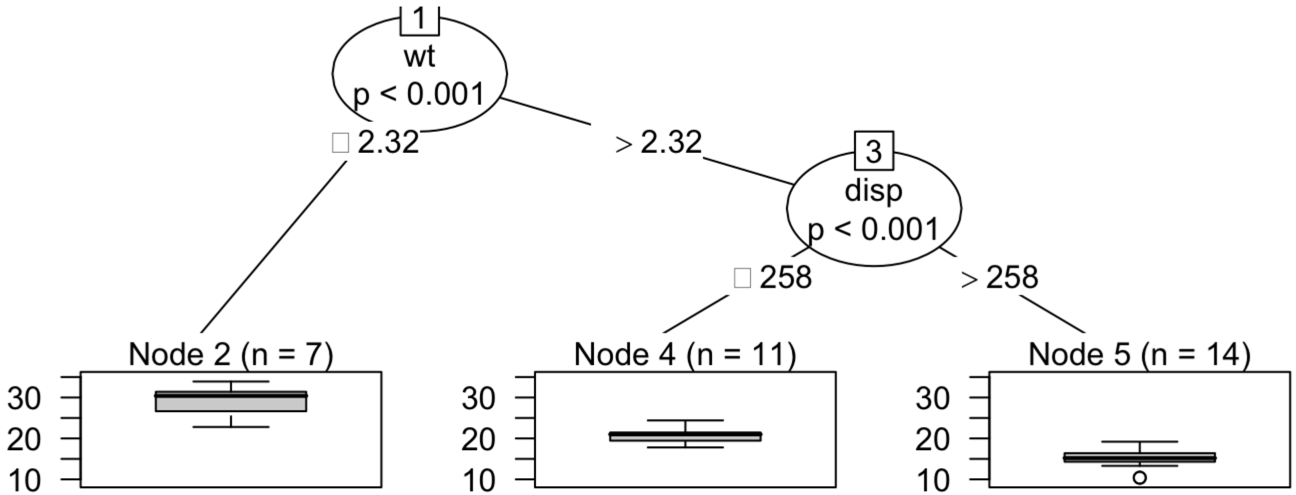
Logistic Regression



- Challenge: perform the same analysis using the `iris` dataset. Classify try to classify the *setosa* species in terms of Sepal Width and Length;

Decision Trees

- a tree is a structure that has nodes and edges;
- For a decision tree, at each node we might have a value against which we split in order to gain some insight from the data;



Decision Trees

Challenge: Use the Iris dataset to perform an analysis of theis species using Decision Trees.

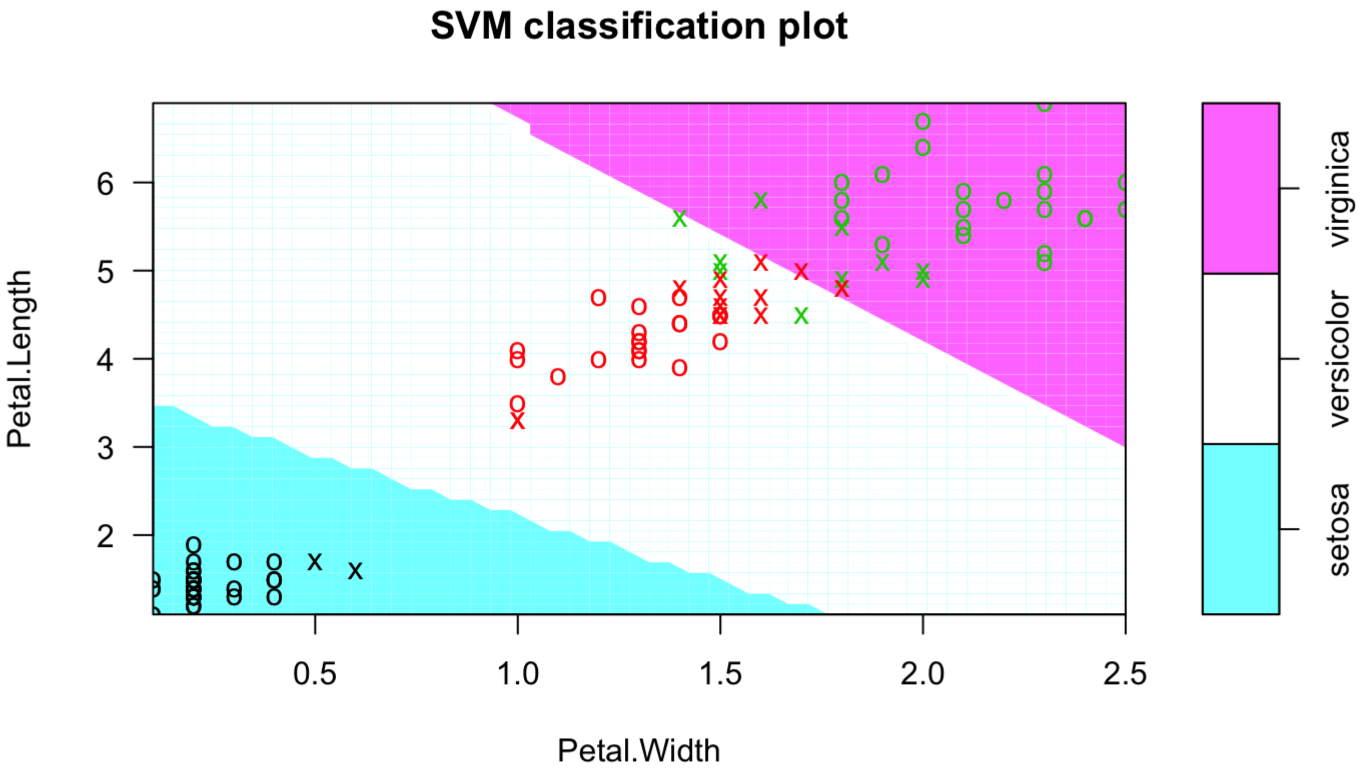
Support Vector Machines

- Optimal hyperplane for linearly separable patterns;
- Extend to patterns that are not linearly separable by transformations of original data to map into new space
    - the Kernel function

Support Vector Machines

Support Vector Machines

- In R use the `svm()` function;
- define the Kernel Family;
  - Linear, Radial, etc…



Unsupervised Methods

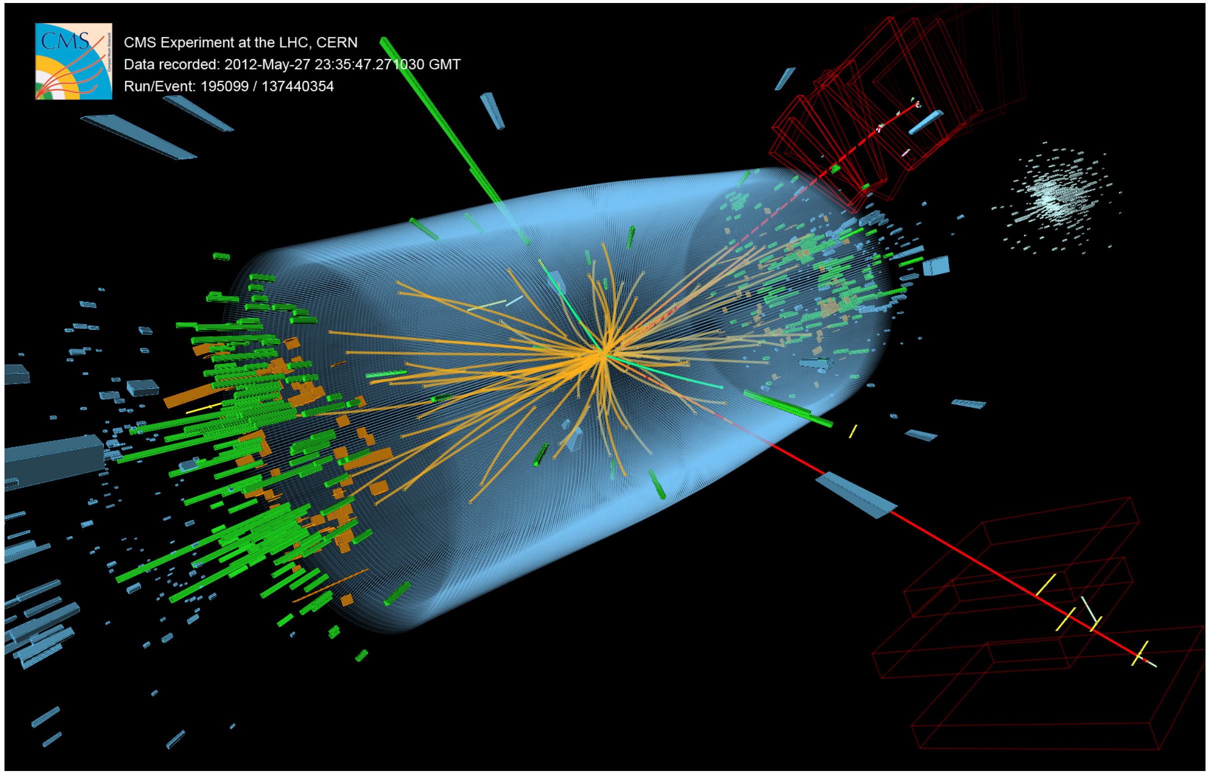
KMeans

- clusters are represented by a central vector or a centroid
  - This centroid might not necessarily be a member of the dataset;
- iterative clustering algorithm;
- notion of similarity
  - derived by how close a data point is to the centroid of the cluster;
- The initial number of centroids should be specified;

Challenge

CMS Calorimeter

- The CMS barrel calorimeter



The Challenge

- Use the variables inside the dataset and try to build the particle clusters. Be aware that the dataset contains the cluster to which the particle belongs the `jet` variable.
- Try classification algorithms. How we assess the quality of the models proposed?