Foundations of Data Science
Lecture 3

Rumi Chunara, PhD

CS3943/9223

So Far…

- What is Data Science?
- Data Handling
- Doing Data Science
- Intro to R
- Types of Data
- Data cleaning, sampling, processing

Today

- Getting Data + APIs
- Intro to ML – what is it
- Two Basic Algorithms
  - kNN
  - Linear Regression

What is machine learning?

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

What is machine learning?

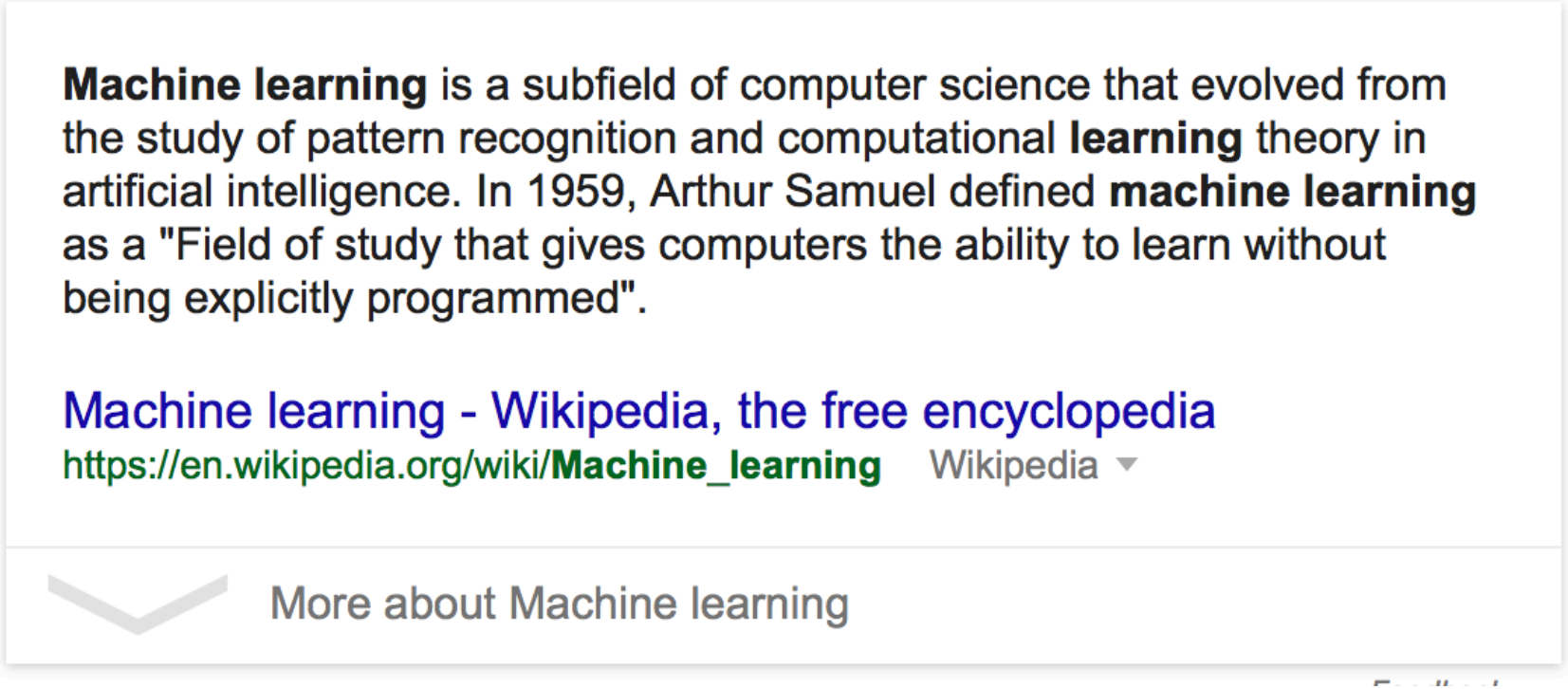
- Supervised Learning (Starting Today)
- Unsupervised Learning (Later)

Machine Learning

- **Supervised:** We are given input samples (X) and output samples (y) of a function y = f(X). We would like to "learn" f, and evaluate it on new data. Types:
  - **Classification:** y is discrete (class labels).
  - **Regression:** y is continuous, e.g. linear regression.

- **Unsupervised:** Given only samples X of the data, we compute a function f such that y = f(X) is "simpler".
  - **Clustering:** y is discrete
  - Y is continuous: **Matrix factorization, Kalman filtering, unsupervised neural networks**.

What is Machine Learning?



Source: http://www.reddit.com/r/MachineLearning/comments/25lnbt/ama_yann_lecun

What is Machine Learning?

- One definition: “Machine learning is the semi-automatic extraction of knowledge from data.”

- **Automatic extraction:** A computer provides the insight
- **Semi-automatic:** Requires many smart decisions by a human

Supervised Machine Learning

Supervised learning
(aka “predictive modeling”):

- Predict an outcome based on input data
- Example: predict whether an email is spam
- Goal is “generalization”

ML Terminology

150 observations *(n = 150)*

Feature matrix “X” has n rows and p columns

Response “y” is a vector with length n

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

response

4 features (p = 4)

ML Terminology

Observations are also known as: samples, examples, instances, records

Features are also known as: predictors, independent variables, inputs, regressors, covariates, attributes

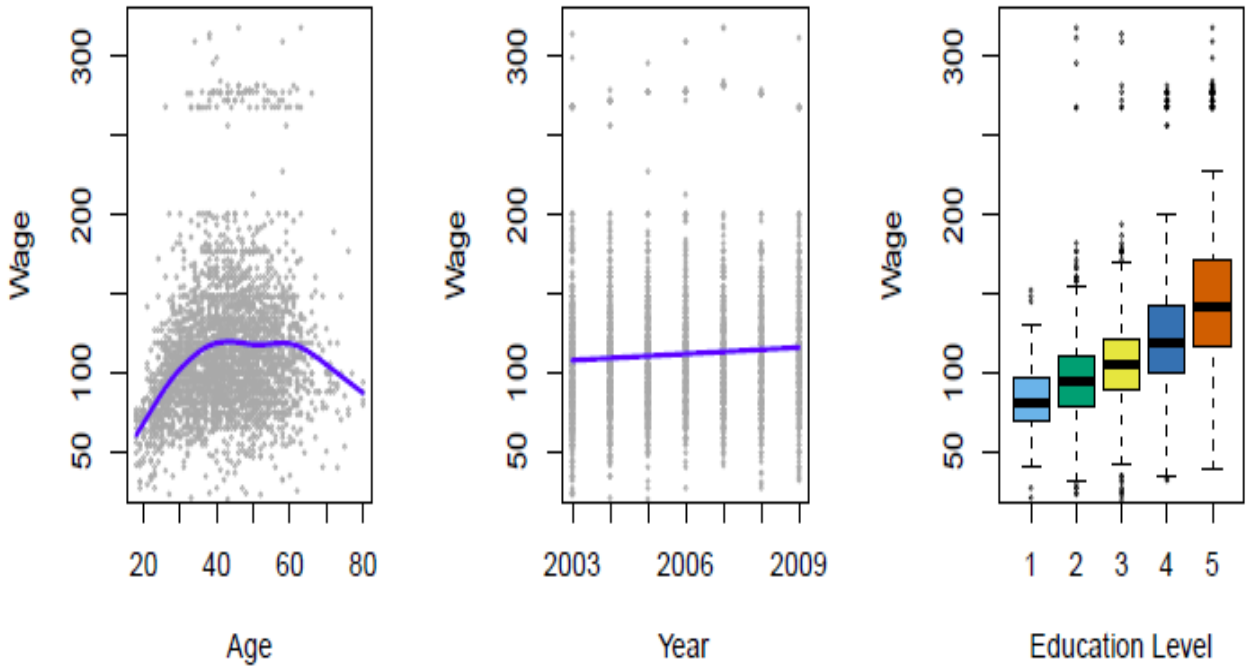
Response is also known as: outcome, label, target, dependent variable

Regression problems have a continuous response.
Classification problems have a categorical response.

The type of supervised learning problem has nothing to do with the features!

Supervised Learning Example

Predict salary using demographic data

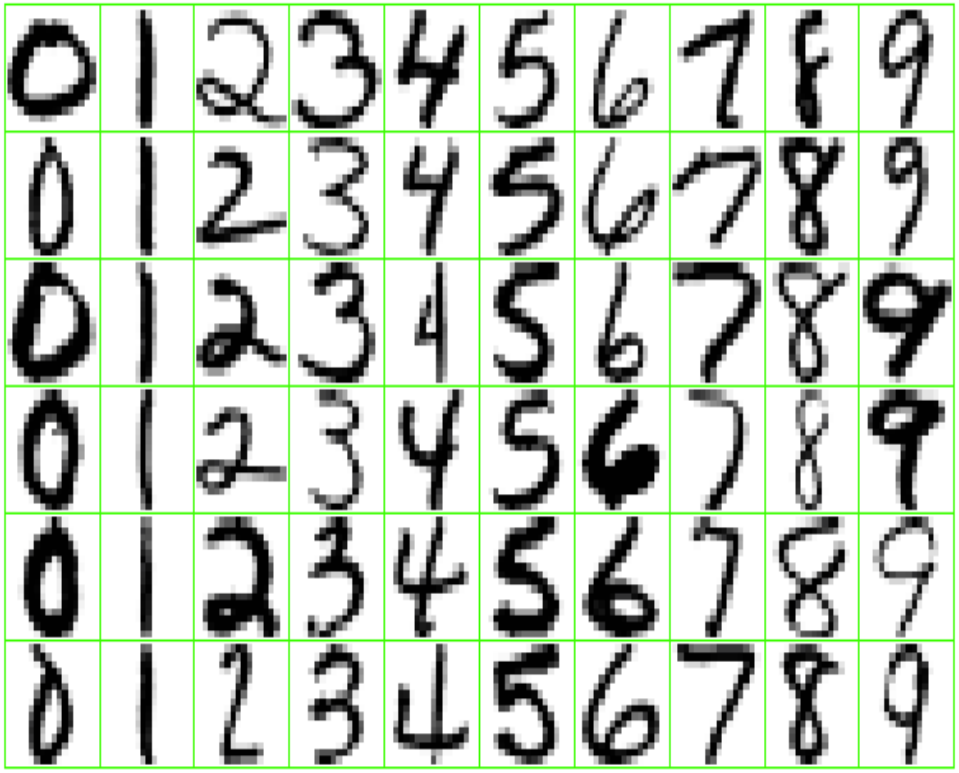


Income survey data for males from the central Atlantic region of the USA in 2009

Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

Supervised Learning Example

Identify the numbers in a handwritten zip code



Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

Categories of Supervised Learning

There are two categories of supervised learning:

Regression

- Outcome we are trying to predict is continuous
- Examples: price, blood pressure

Classification

- Outcome we are trying to predict is categorical (values in a finite, unordered set)
- Examples: spam/ham, cancer class of tissue sample

Regression or Classification?

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

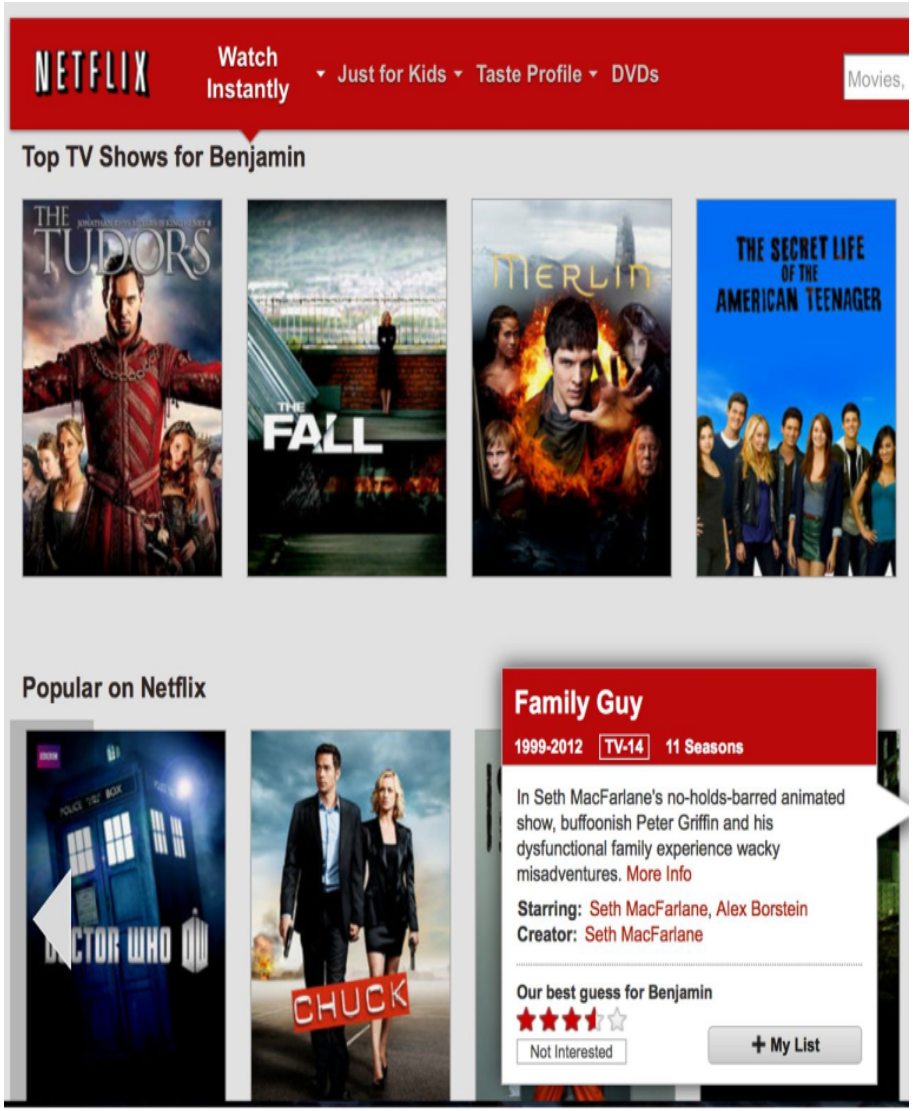
Goal: Detect subtle patterns in the data that predicts infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

Regression or Classification?



Regression or Classification?

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Supervised Learning

How does supervised learning "work"?

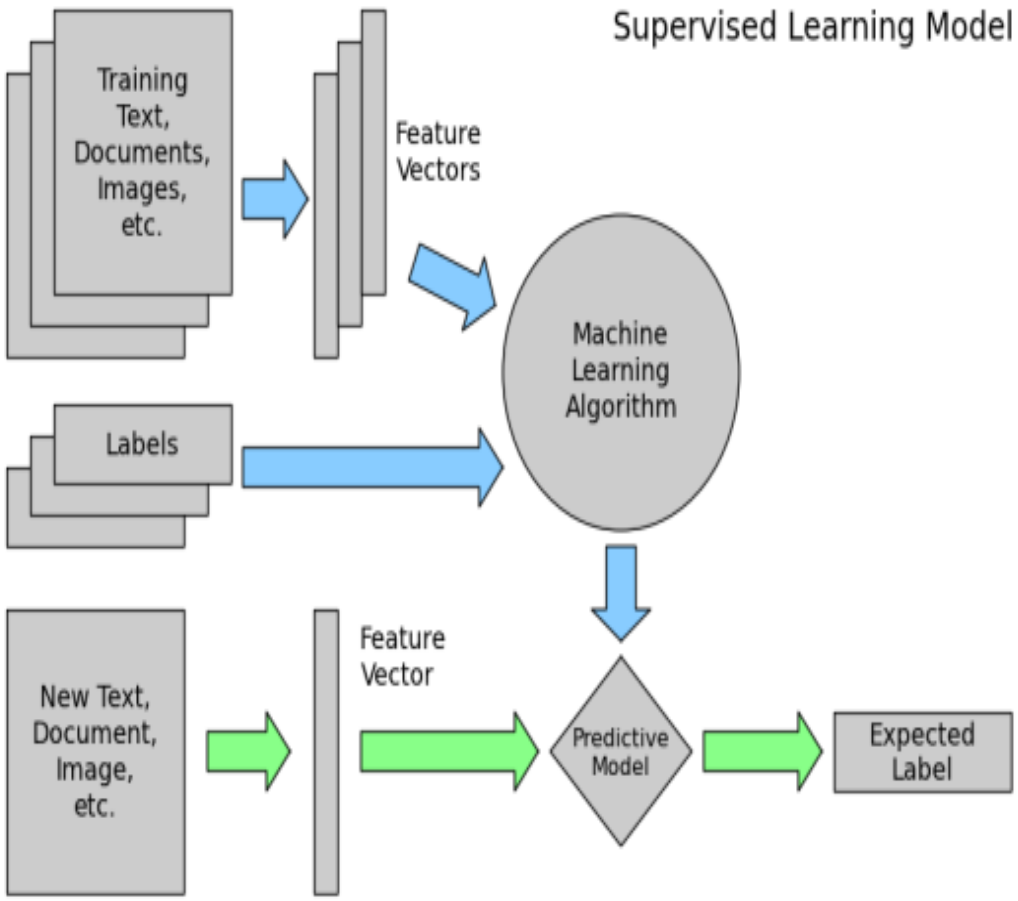
1. Train a **machine learning model** using **labeled data**
   - "Labeled data" is data with a response variable
   - "Machine learning model" learns the relationship between the features and the response

2. Make predictions on **new data** for which the response is unknown

The primary goal of supervised learning is to build a model that "generalizes": It accurately predicts the **future** rather than the **past**!

Supervised Learning

How does supervised learning "work"?



Supervised Learning Example

Supervised learning example: Dog detector

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: Dog (yes or no), hand-labeled

1. Train a **machine learning model** using **labeled data**
   - Model learns the relationship between the image data and the "dog status"

2. Make predictions on **new data** for which the response is unknown
   - Give it a new image, predicts the "dog status" automatically

Machine Learning

- **Supervised:**
  - Is this image a cat, dog, car, house?
  - How would this user score that restaurant?
  - Is this email spam?
  - Is this blob a supernova?
- **Unsupervised**
  - Cluster some hand-written digit data into 10 classes.
  - What are the top 20 topics in Twitter right now?
  - Find and cluster distinct accents of people at Berkeley.

Semi-supervised Learning

- Suppose that we have a set of n observations.
- For m of the observations, where $m < n$, we have both predictor measurements and a response measurement.
- For the remaining $n - m$ observations, we have predictor measurements but no response measurement.
- Such a scenario can arise if the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect.

Techniques

- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Linear Regression
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Topic Models

k-Nearest Neighbors

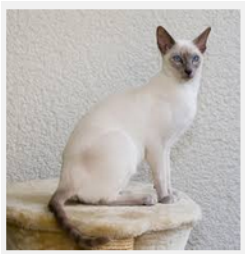
- No training needed.
- Accuracy generally improves with more data.
- Matching is simple and fast (and single pass).
- Usually need data in memory, but can be run off disk.

k-Nearest Neighbors

Given a query item:
Find k closest matches
in a labeled dataset ↓

k-Nearest Neighbors

Given a query item:
Find k closest matches



Return the most
Frequent label

k-Nearest Neighbors

k = 3 votes for “cat”

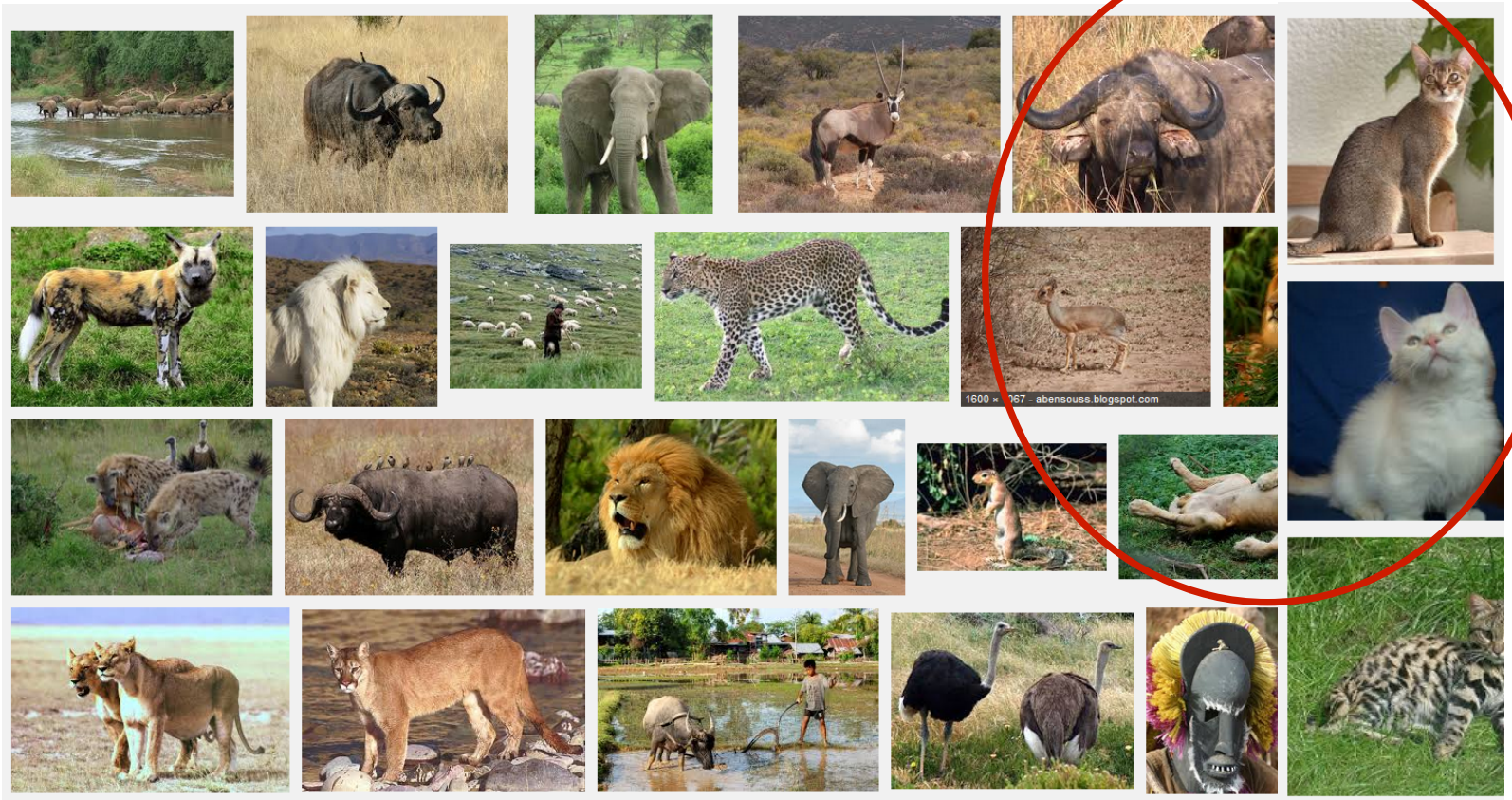
k-Nearest Neighbors

2 votes for cat,

1 each for Buffalo,

Deer, Lion

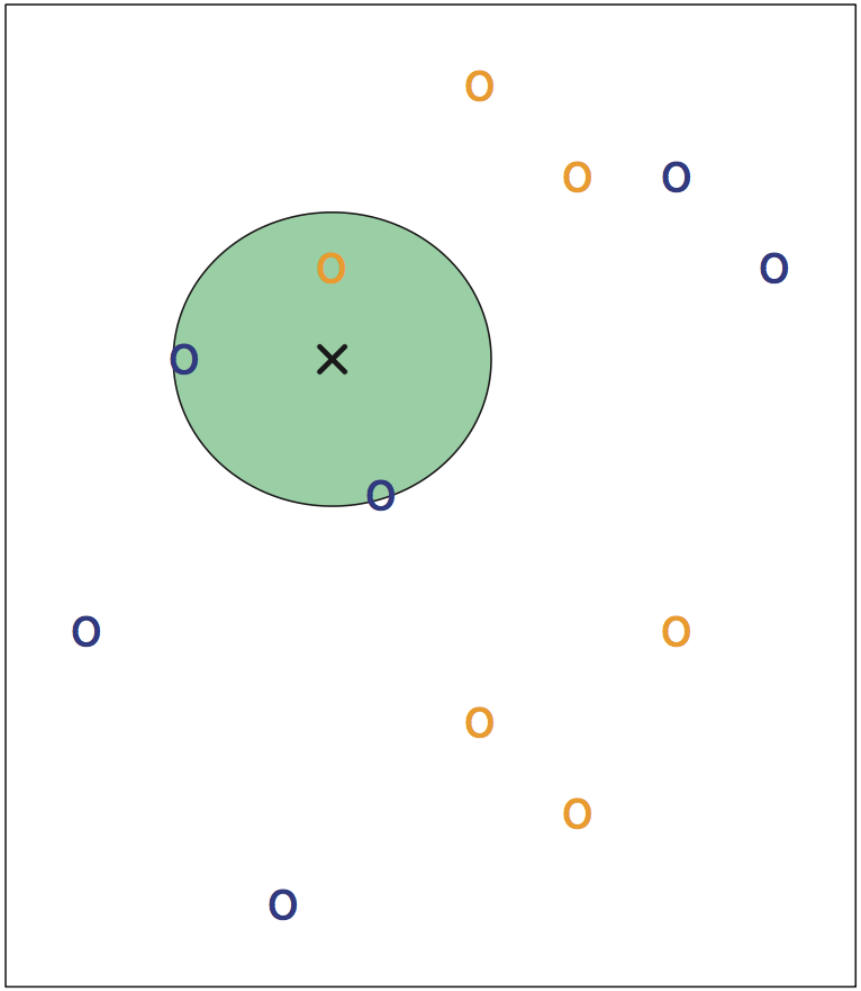
Cat wins…

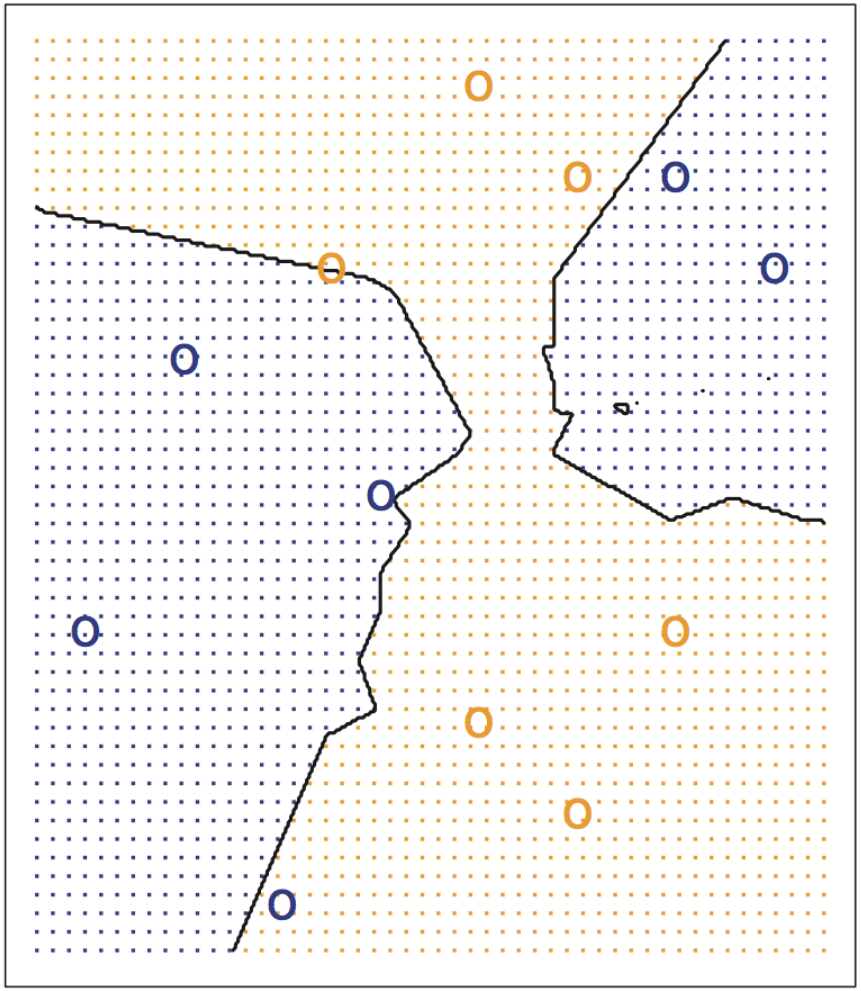


KNN – Method Motivation and Overview

- In theory we would always like to predict qualitative responses using the Bayes classifier (known probabilities)
- For real data, we do not know the conditional distribution of Y given X
- Some approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability (e.g. KNN classifier).
- Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class *j* as the fraction of points in N_0 whose response values equal *j*
- Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

KNN Visual Explanation





k-NN issues

The Data is the Model

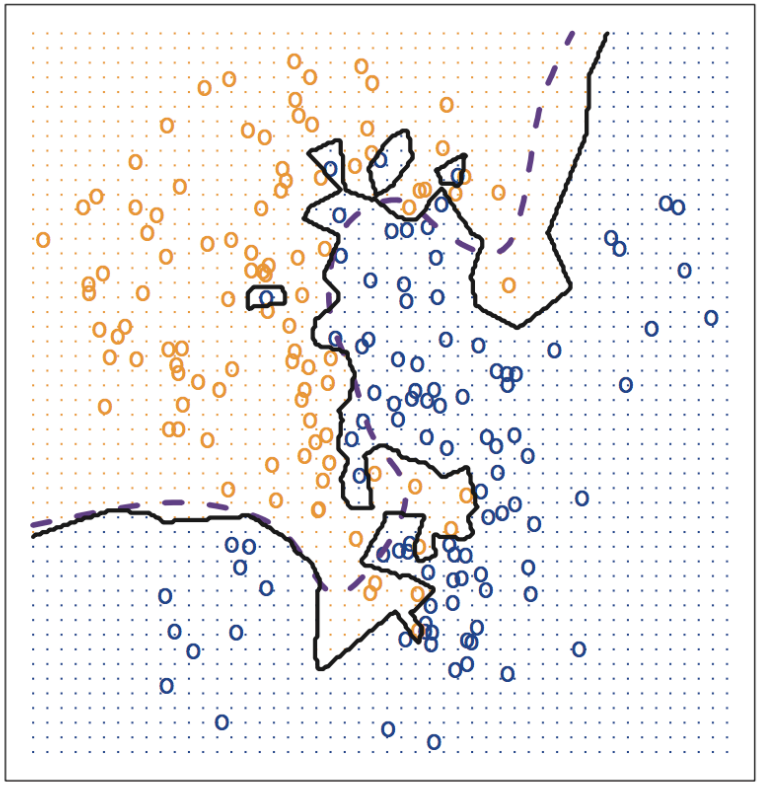
- No training needed.
- Accuracy generally improves with more data.
- Matching is simple and fast (and single pass).
- Usually need data in memory, but can be run off disk.

Minimal Configuration:

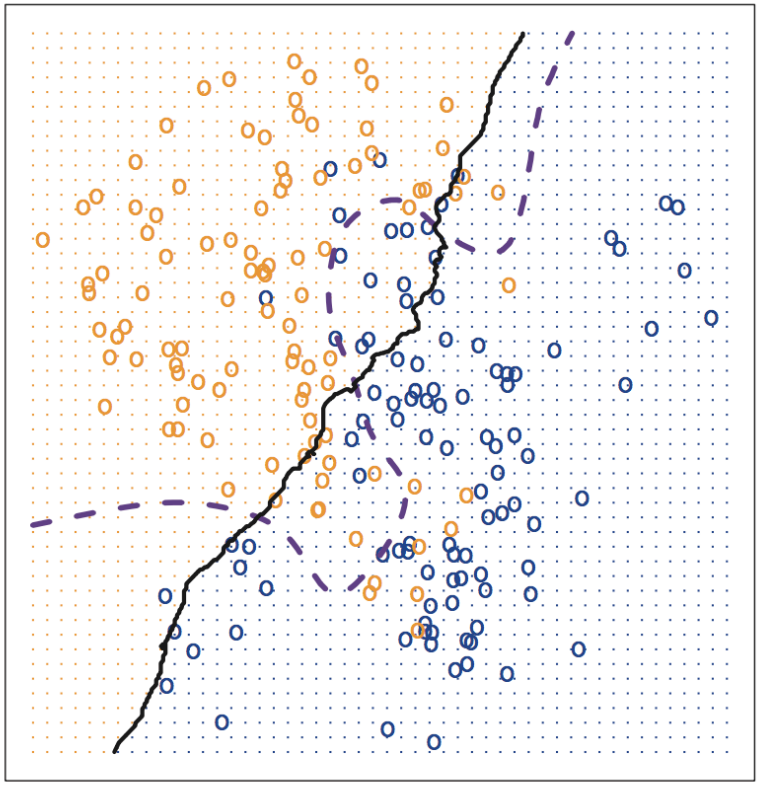
- Only parameter is k (number of neighbors)
- Two other choices are important:
  - Weighting of neighbors (e.g. inverse distance)
  - Similarity metric

Choice of “K”

KNN: K=1



KNN: K=100



K-NN metrics

- **Euclidean Distance:** Simplest, fast to compute:

$$= \left|\vec{a} - \vec{b}\right|^2 = (\vec{a} - \vec{b})^T (\vec{a} - \vec{b}) = 2(1 - \cos(\vec{a}, \vec{b}))$$

- **Cosine Distance:** Good for documents, images, etc.

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} + \vec{b}}{\left|\vec{a}\right| + \left|\vec{b}\right|}$$

- **Jaccard Distance:** For set data:

 # matching / # attributes not involved in 00 metric

- **Hamming Distance:** For string data: gives the result of how many attributes where different.

$$D_H = \sum_{i=1}^{k} | x_i - y_i |$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

K-NN metrics

- **Manhattan Distance:** Coordinate-wise distance:

$$\sum_{i=1}^{k} | x_i - y_i |$$

- **Edit Distance:** for strings, especially genetic data.

Given two strings *a* and *b* on an alphabet Σ (e.g. the set of ASCII characters, the set of bytes [0..255], etc.), the edit distance d(*a*,*b*) is the minimum-weight series of edit operations that transform *a* into *b*.

Selecting Distance Measures

- Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure: Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms.
- Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.

Selecting Distance Measures

- If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance?
- Note that two human beings share > 99.9% of the same genes.

Distance Measure Examples

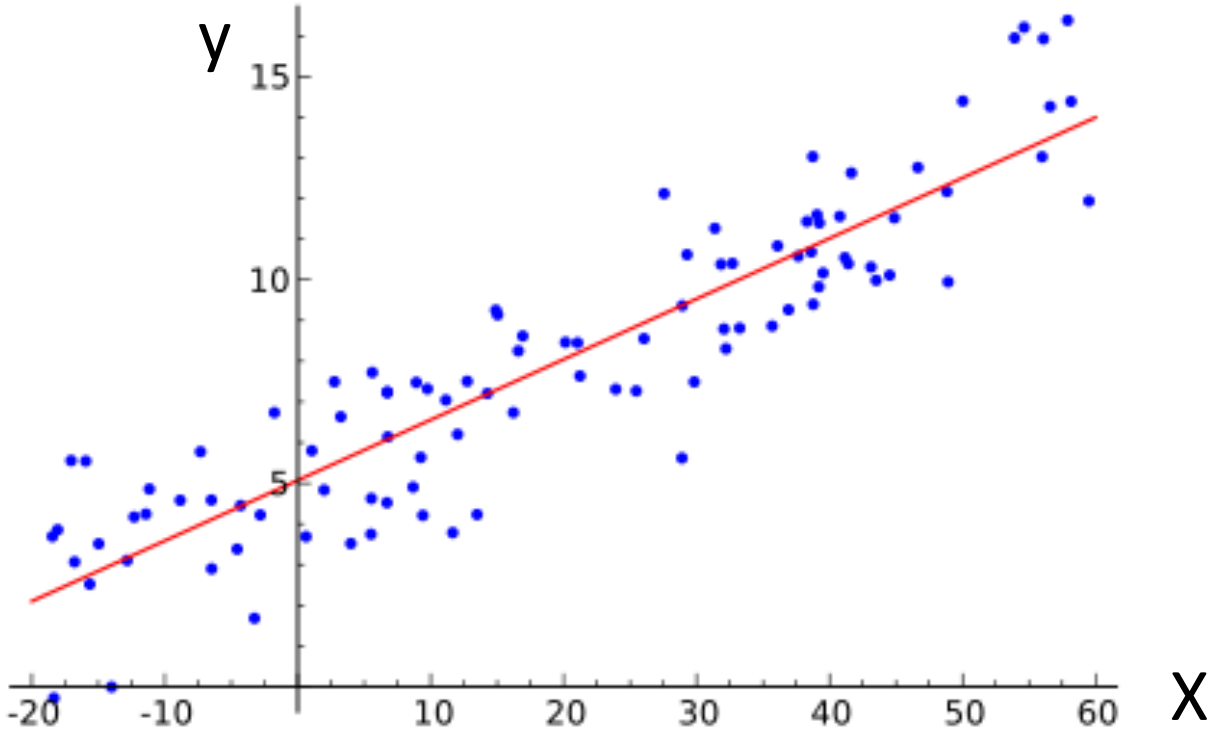
- For the following vectors, x and y, calculate the indicated similarity or distance measures.
- x = (0,1,0,1), y = (1,0,1,0)
- cosine, correlation, Euclidean, Jaccard

Linear Regression

- LR: a very simple approach for supervised learning.
- LR useful for predicting a quantitative response.
- Been around for a while, useful for large data sets also (quick)

Linear Regression

We want to find the best line (linear function y=f(X)) to explain the data.



Linear Regression

The predicted value of y is given by:

$$y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

The vector of coefficients β is the regression model.

If $X_0 = 1$, the formula becomes a matrix product:

$$y = \mathrm{X}\,\beta$$

Linear Regression in Matrix Form

- Consider writing an equation for each observation: $Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$

- Model in Matrix Form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \dots & \dots \\ 1 & X_3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Residual Sum-of-Squares

To determine the model parameters β from some data, we can write down the Residual Sum of Squares:

$$RSS = \sum_i (y_i - \beta x_i)^2$$

or symbolically $RSS(\beta) = (\vec{y} - \vec{X}\beta)^T (\vec{y} - \vec{X}\beta)$ To minimize it, take the derivative wrt β which gives:

$$\vec{X}^T (\vec{y} - \vec{X}\beta) = 0$$

And if $\vec{X}^T \vec{X}$ is non-singular, the unique solution is:

$$\beta = (\vec{X}^T \vec{X})^{-1} (\vec{X}^T \vec{y})$$

Example: Stochastic Gradient

A very important set of iterative algorithms use **stochastic gradient** updates.

They use a **small subset or mini-batch X** of the data, and use it to compute a gradient which is added to the model

$$\beta' = \beta + \alpha\nabla$$

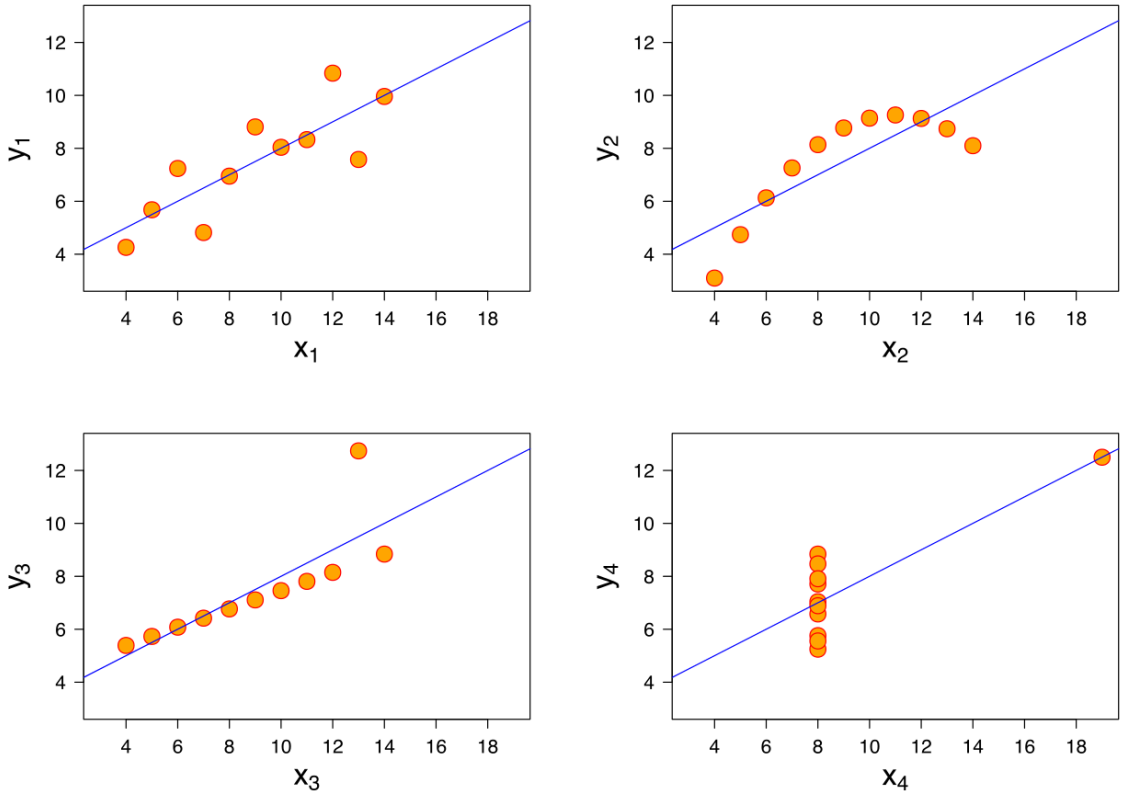
Where α is called the **learning rate**.

These updates happen **many times** in one pass over the dataset.

Its possible to compute high-quality models with very few passes, sometime with less than one pass over a large dataset.

R^2-values and P-values

We can **always** fit a linear model to any dataset, but how do we know if there is a **real linear relationship**?



R^2-values and P-values

Approach: Use a hypothesis test. The null hypothesis is that there is no linear relationship ($\beta = 0$).

Statistic: Some value which should be small under the null hypothesis, and large if the alternate hypothesis is true.

R-squared: a suitable statistic. Let $y = X\beta$ be a predicted value, and y be the sample mean. Then the R-squared statistic is

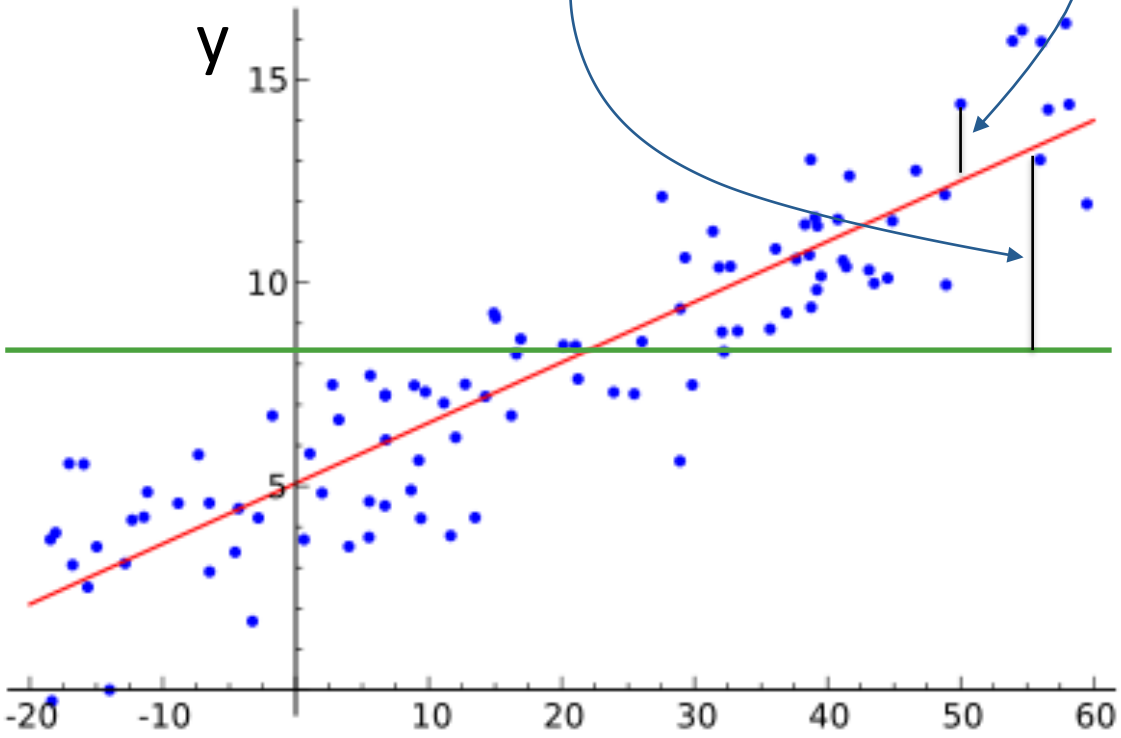
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

And can be described as the fraction of the total variance not explained by the model.

R-squared

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Small if good fit



Line of y

Line of y

R^2-values and P-values

Statistic: From R-squared we can derive another statistic (using degrees of freedom) that has a standard distribution called an **F-distribution**.

From the CDF for the F-distribution, we can derive a **P-value** for the data.

The P-value is, as usual, the probability of observing the data under the null hypothesis of no linear relationship.

If **p is small**, say less than 0.05, we conclude that **there is a linear relationship**.

Today

- Getting Data + APIs
- Intro to ML – what is it
- Two Basic Algorithms
  - kNN
  - Linear Regression

Next Time

- Spatio-temporal analyses
- Implementing KNN, Regression and Clustering in R