Machine learning!

By me

What is it?

- **Subset of Artificial Intelligence**
- Form of data analysis that teaches a computer how to learn about data
- Automates analytical model building
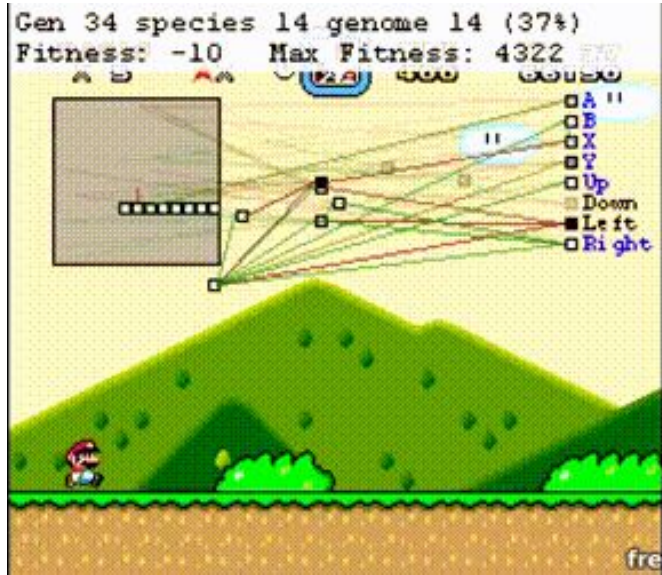- Goal of machine learning is to understand the structure of the data

Machine learning is a type of AI, they're not mutually exclusive!

Life of an ML algorithm

1. Ask a question about data
2. Collect data
3. Train the algorithm on the data
4. Test algorithm on new data
5. Collect feedback
6. Use feedback in algorithm

Types of ML Algorithms

- **Supervised**: Human provides labelled data for the algorithm to train on.
  - Ex. linear regression, random forest, SVM
  - Used in Siri/Cortana, Fingerprint ID, and email spam filters
- **Unsupervised**: Algorithm only has access to unlabelled data
  - Ex. k-means clustering, singular value decomposition, nearest neighbor maps
  - Used in many neural networks
- **(Semisupervised)**
- **Reinforcement Learning**: Algorithm discovers through trial and error which actions yield the greatest rewards
  - Used in algorithms that play video games, robotics, and lots of other stuff
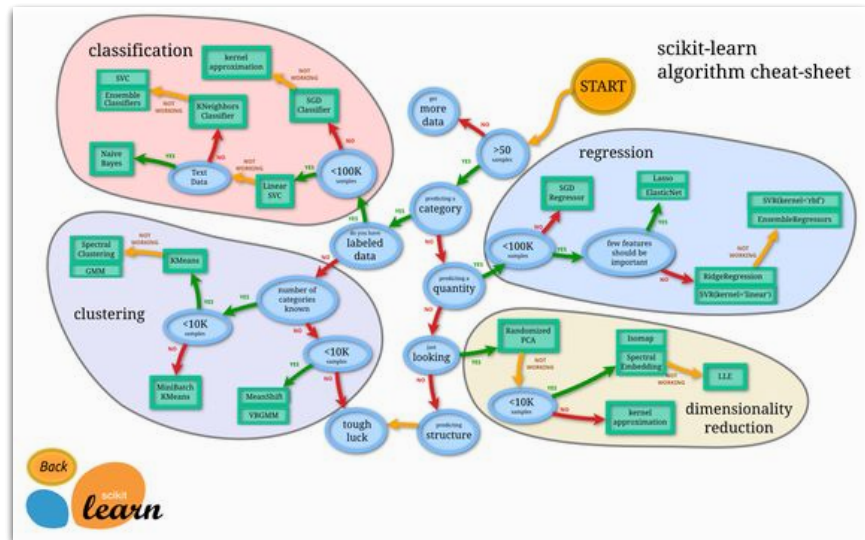
Supervised Machine Learning Problem Types

- Regression: Predicting continuous valued output
  - Random Forest, Support Vector Regression
  - Ex. "Predict average house price based on average income in that area"
- Classification: Predicting discrete valued output
  - Decision trees, Naive Bayes, K nearest neighbors, etc.
  - Binary classification: only 2 categories, usually labeled 0 and 1
 - Multi-class classification: more than 2 categories
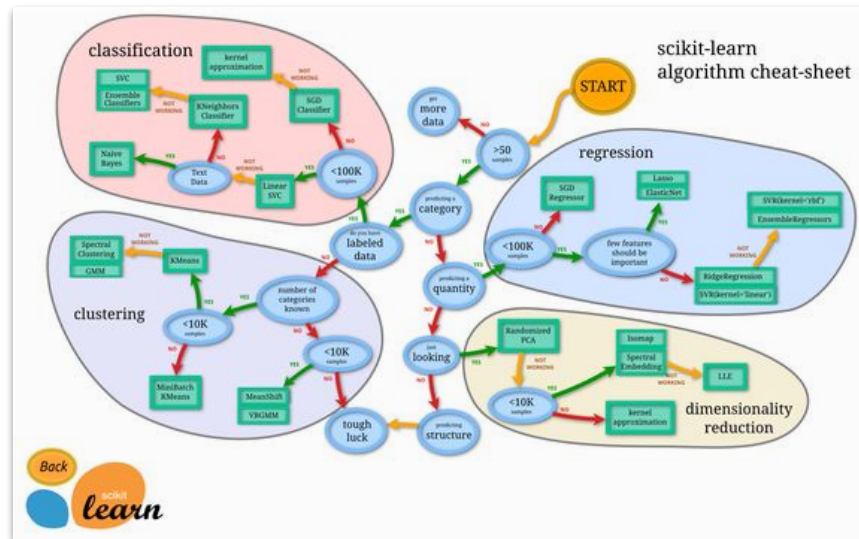
Unsupervised Machine Learning Problem Types

- Clustering: Discovering inherent groupings present in data
  - K-means, t-SNE, Spectral clustering
  - Ex. "What do customers tend to buy?"
- Association: Discovering rules that explain associations between the variables
  - Apriori, FP-growth
  - Ex. "What do customers that buy product X also tend to buy?"

What is Linear Regression?

- Simple way to predict output values given input data
- One of the simplest algorithms for solving <u>regression</u> problems
- Supervised machine learning algorithm
- Assumes there's a linear relation between the input value(s), X, and the output value(s), Y
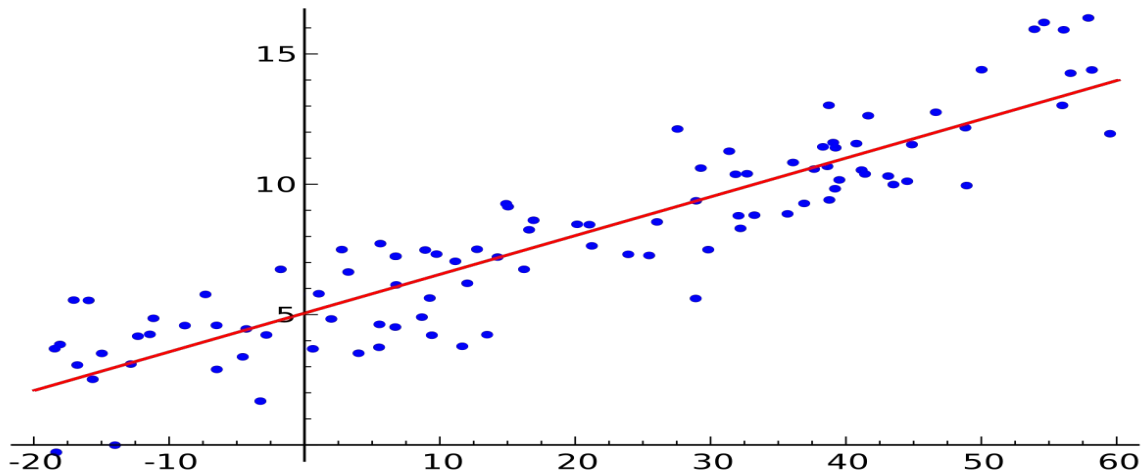
How does it work?

- The dependent variables are assumed to be linearly related to the independent variables
- Thus we can write each dependent variable as a linear combination of the independent variables
  - $y = \beta_0 + \beta_1 x + \beta_2 x + \ldots + \beta_n x_n$
  - Each of the β_k is a coefficient modifying the independent x_k variable
- Algorithm tries to minimize the error in the predictions vs. actual by changing the coefficient values

OH, HOW DOES IT WORK?
IT DOESN'T WORK.

What is Logistic Regression?

- Most basic method for binary <u>classification</u> problems
- Supervised machine learning algorithm
- Uses the logistic function, more commonly known as a sigmoid function
- Models the probability an input belongs to the default class
- Assumes all input belongs to one of two output classes
- Easy to implement and efficient to train
- Data must be linearly separable
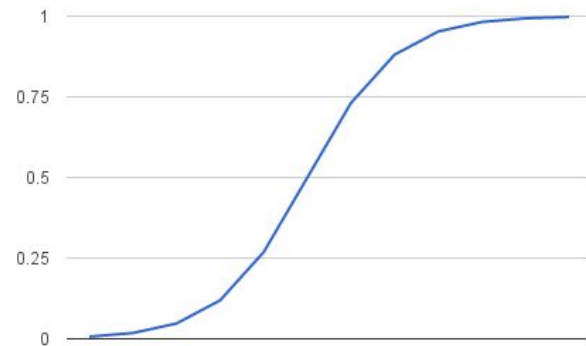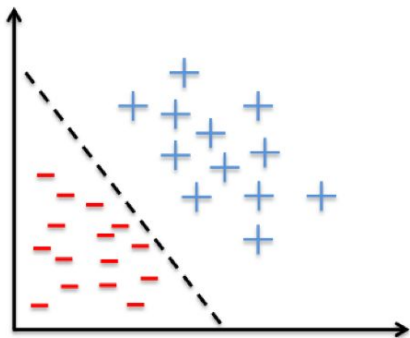
How do we do Logistic Regression?

- We use Maximum Likelihood Estimation (MLE) to maximize the parameters of the sigmoid function
  - The math here is pretty hairy, but the computer does it quickly and accurately
  - If you're interested, a good explanation is [here](#)
- Use labeled data to train the model with MLE, then test with unlabeled data

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

$$t = \beta_0 + \beta_1 x$$
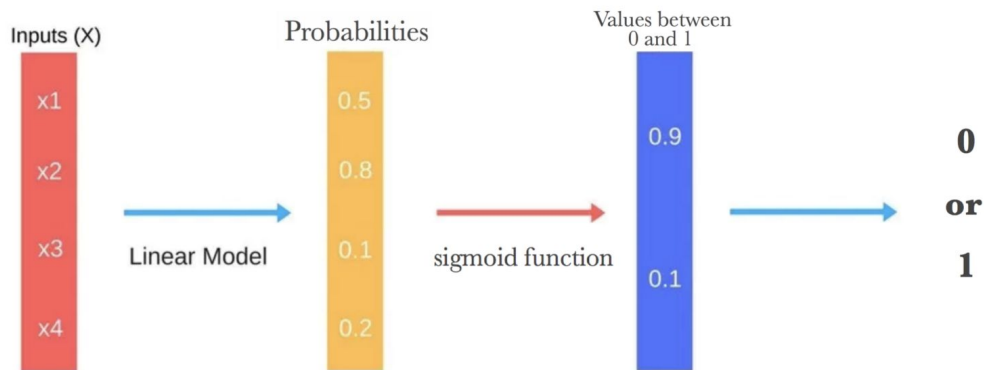
Multinomial Logistic Regression

- The jupyter notebook showcases multinomial logistic regression
- Works on categorical variables
- Principles are exactly the same as basic logistic regression
- Coefficient estimation uses a slightly more complicated algorithm, it essentially functions the same way