Classification

Logistic Regression

Logistic Regression

- Logistic regression is kind of like linear regression but is used when the dependent variable is not a number, but something else (like a Yes/No response)
- Its called Regression but performs classification as based on the regression it classifies the dependent variable into either of the classes

Logistic Regression

- Logistic regression is used for prediction of output which is binary
- For example, if a credit card company is going to build a model to decide whether to issue a credit card to a customer or not, it will model for whether the customer is going to “Default” or “Not Default” on this credit card.

Logistic Regression

K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN)

- it is used to identify the data points that are separated into several classes to predict the classification of a new sample point
- K-NN is a **non-parametric,** lazy learning algorithm
- It classifies new cases based on a similarity measure (e.g. distance functions)

K-Nearest Neighbors (K-NN)

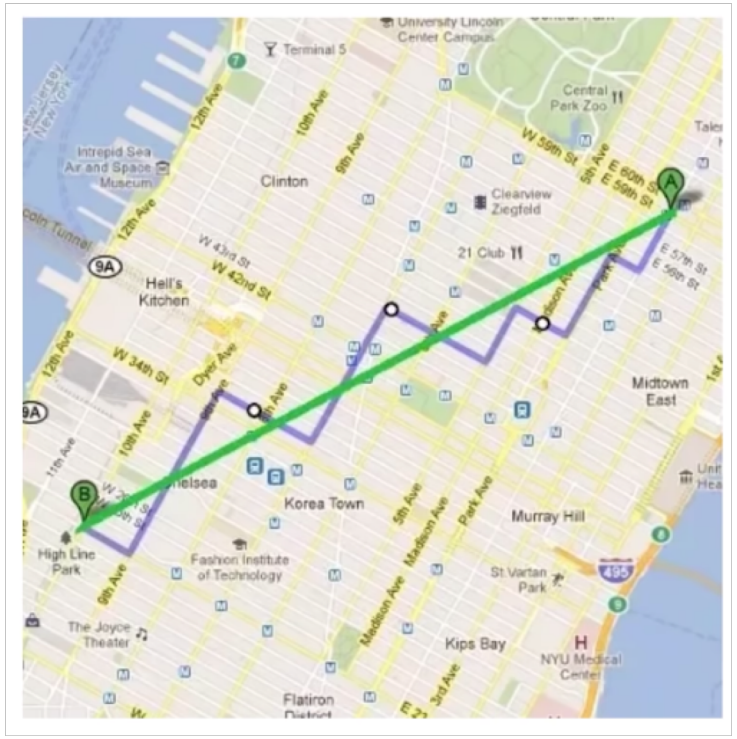
- K is the number of neightbors to the consider
- Scaling is important
- K should be odd in order to avoid the ties
- Voting can be weighted by the distance to each neighbor
- Does not scale to large data well

K-Nearest Neighbors (K-NN)

- K-Nearest Neighbor does not learn
- It is lazy and just memorizes the data
- Works well with a small number of input variables but struggles when the number of inputs is very large

K-Nearest Neighbors (K-NN)

- Distance Algorithms
  - Euclidean distance
  - Manhattan distance



Support Vector Machine

Support Vector Machine

- used for both regression and Classification
- It is based on the concept of decision planes that define decision boundaries
- A decision plane(hyperplane) is one that separates between a set of objects having different class memberships
- It performs classification by finding the hyperplane that maximizes the margin between the two classes with the help of support vectors.

SVM - Advantages

- **High Dimensionality**
  - SVM is an effective tool in high-dimensional spaces, which is particularly applicable to document classification and sentiment analysis where the dimensionality can be extremely large.

SVM - Advantages

- **Memory Efficiency**
  - Since only a subset of the training points are used in the actual decision process of assigning new members, just these points need to be stored in memory (and calculated upon) when making decisions.

SVM - Advantages

- **Versatility**
  - Class separation is often highly non-linear. The ability to apply new kernels allows substantial flexibility for the decision boundaries, leading to greater classification performance.

SVM - Disadvantages

- **Kernel Parameters Selection**
  - SVMs are very sensitive to the choice of the kernel parameters
  - In situations where the number of features for each object exceeds the number of training data samples, SVMs can perform poorly