Supervised Learning : K-Nearest Neighbors

- Define a distance metric
  - Euclidean
  - Manhatten
  - Any vector norm
- Choose the number of K neighbors
- Find the K nearest neighbors of the new observation that we want to classify

Prediction problem : what group is an observation associated with?

- Assign class label by majority vote
- Important to find the right K
  - Commonly use $k = \frac{\sqrt{N}}{2}$ where N = number of samples

K - number of neighbors that are closest to the new observation

More dims, need more data ↑

Cuse of dimensionality
- the more dimensions, the sparser in space the data points

K small : observation is local

K large : observation is an average of the neighborhood observations in the training data

Non-parametric, instance-based, lazy

- Non-parametric ; Model is not defined by fixed set of parameters
- Instance-based **or** lazy learning : Model is the result of effectively memorizing training data
- Requires keeping the original data set
- Space and time complexity grw with size of training data
- Suffers from cosse of dimensionality : points become increasingly isolated with more dims, for a fixed size training set
- sklearn.neighbors.KNeighborsClassifier