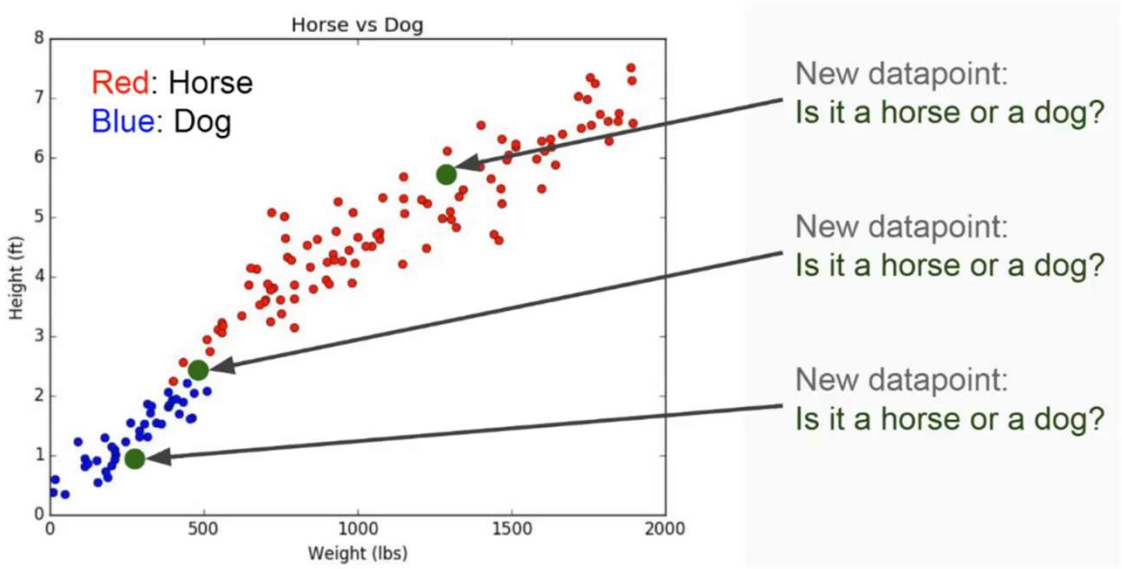

k - Nearest
Neighbours

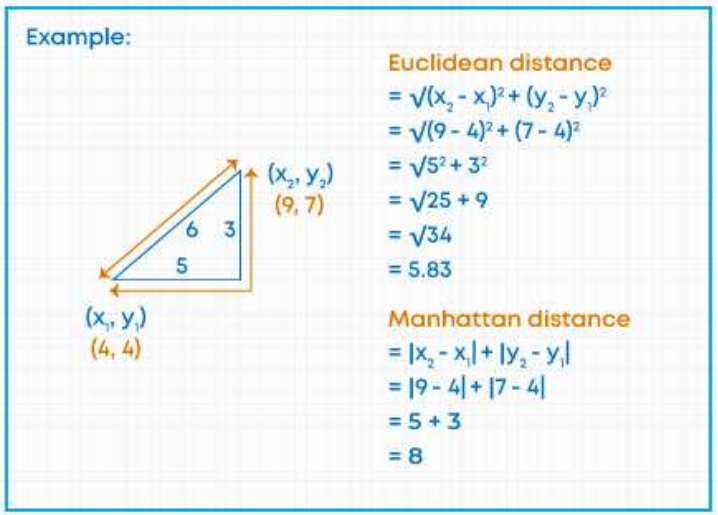
K-Nearest Neighbours

- KNN is a classification algorithm that operates on a very simple principle, best illustrated with an example:
- Imagine we had some data on dogs and horses with heights and weights



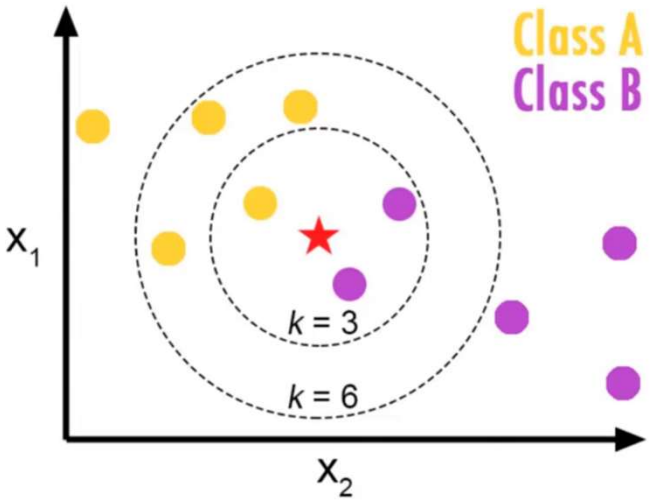
K – Nearest Neighbours

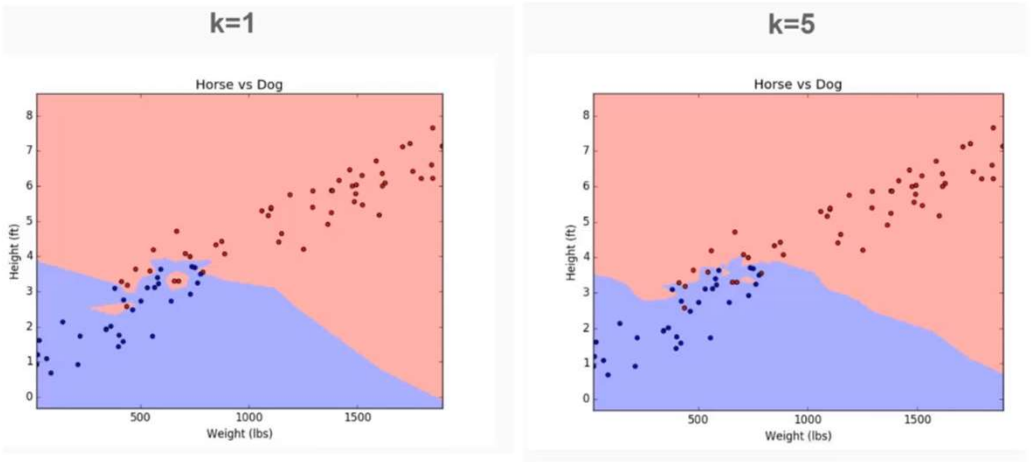
- As KNN is a distance based classifier, the closer two points are, the greater the similarities in behavior and therefore selection choice
- The different methods used to measure the distance are
  - Manhattan
  - Euclidean



Choosing K

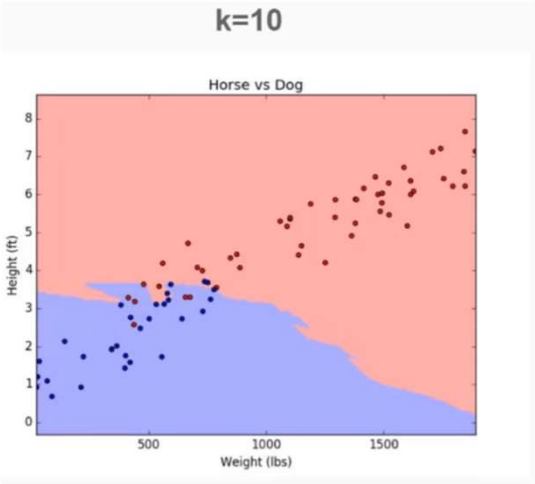
- Choosing the value for K will affect what class a new point is assigned to:

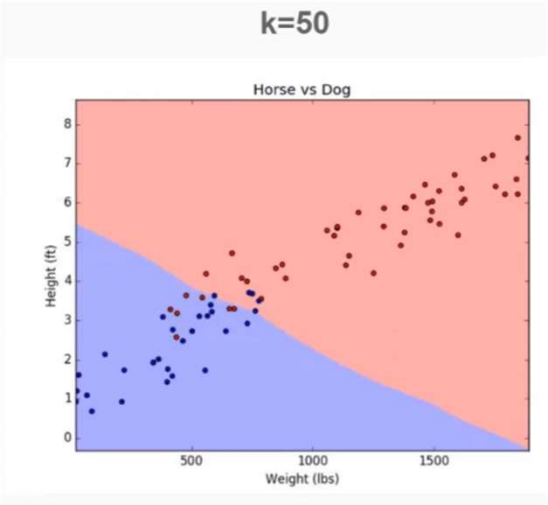




A smooth separation appears as k becomes higher in value

Choosing K


k=10
Horse vs Dog
Height (ft)
Weight (lbs)
0
1
2
3
4
5
6
7
8
500
1000
1500


k=50
Horse vs Dog
Height (ft)
Weight (lbs)
0
1
2
3
4
5
6
7
8
500
1000
1500

How to find ideal K?

- Using a number range, fit a KNN classifier for each number.
- Create predictions.
- Further evaluate the performance using the predictions produced in step 2.
- Compare results across each model and decide on the one with the least error
- This is also called the **Elbow Method**

Advantages

- Very simple
- Training is trivial
- Works with any number of classes
- Easy to add more data
- Fewer parameters: K, Distance Metric

Disadvantages

- High prediction cost (worse for large data sets)
- Not good with high dimensional data
- Categorical features won’t work well

Applications of KNN

- Text mining
- Agriculture
- Finance
- Medical
- Facial recognition
- Recommendation systems (Amazon, Hulu, Netflix, etc)