KNN

1399/10/07

Amir Shokri

KNN

Introduction::

K-nearest neighbors is a lazy learning instance based classification algorithm which is widely implemented in both supervised and unsupervised learning techniques.

Lazy Learning as it doesn’t learn from discriminative function from training data but memorizes training dataset.

this technique implements classification by considering majority of vote among the “k” closest points to the unlabeled data point.

It works on unseen data and will search through the training dataset for the k-most similar instances.

Euclidean distance / Hamming distance is used as metric for calculating the distance between points.

Amir Shokri

KNN

Euclidean Distance::

$$\boldsymbol{dist}(\,(\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{a},\boldsymbol{b})\,) = \sqrt{(\,(\boldsymbol{x}\,-\boldsymbol{a})^{\boldsymbol{2}} + (\boldsymbol{y}-\boldsymbol{b})^{\boldsymbol{2}})}$$

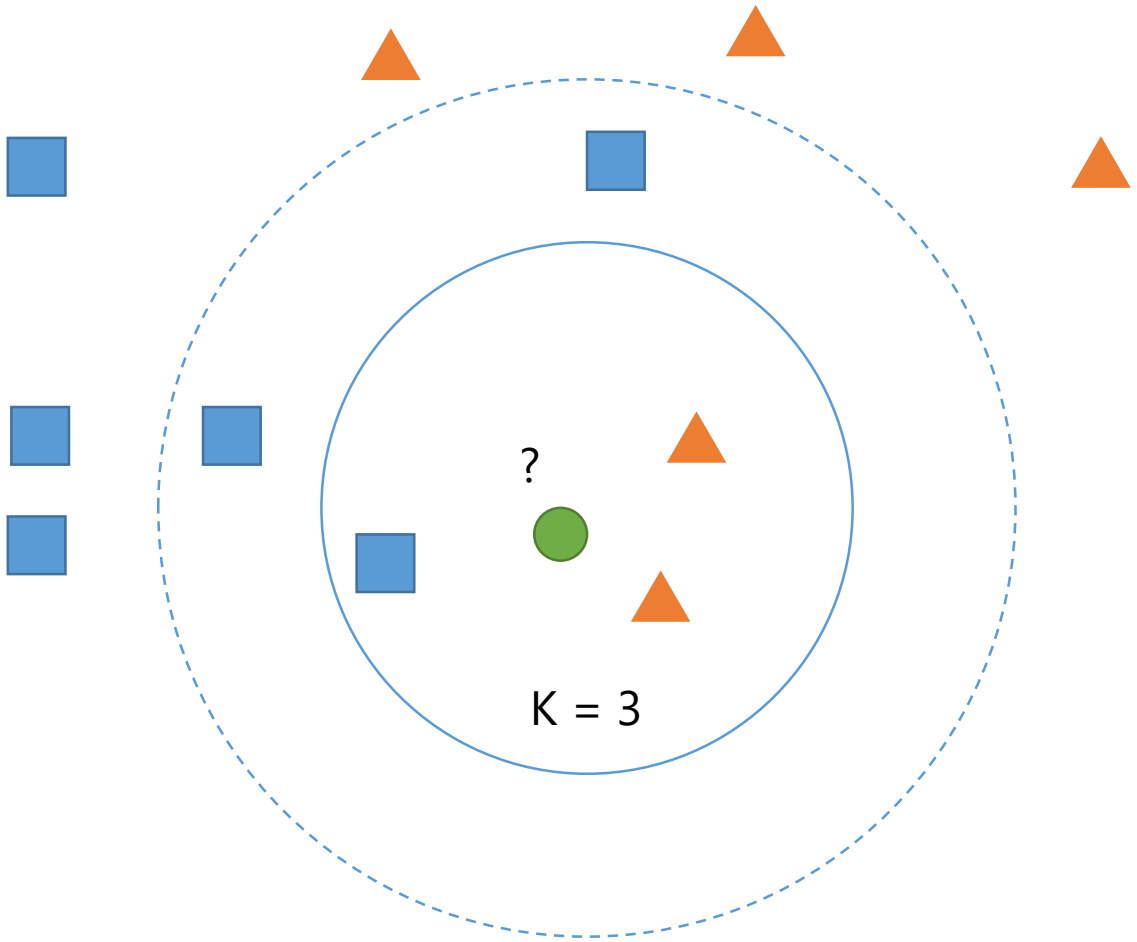
KNN

Hamming Distance::

Onforone	**oneandone**	**->**	**3**
11010110110	**11001111110**	**->**	**2**

Amir Shokri

KNN



Green circle = ?

K = 3

Closest 3 points taken

2 -> orange

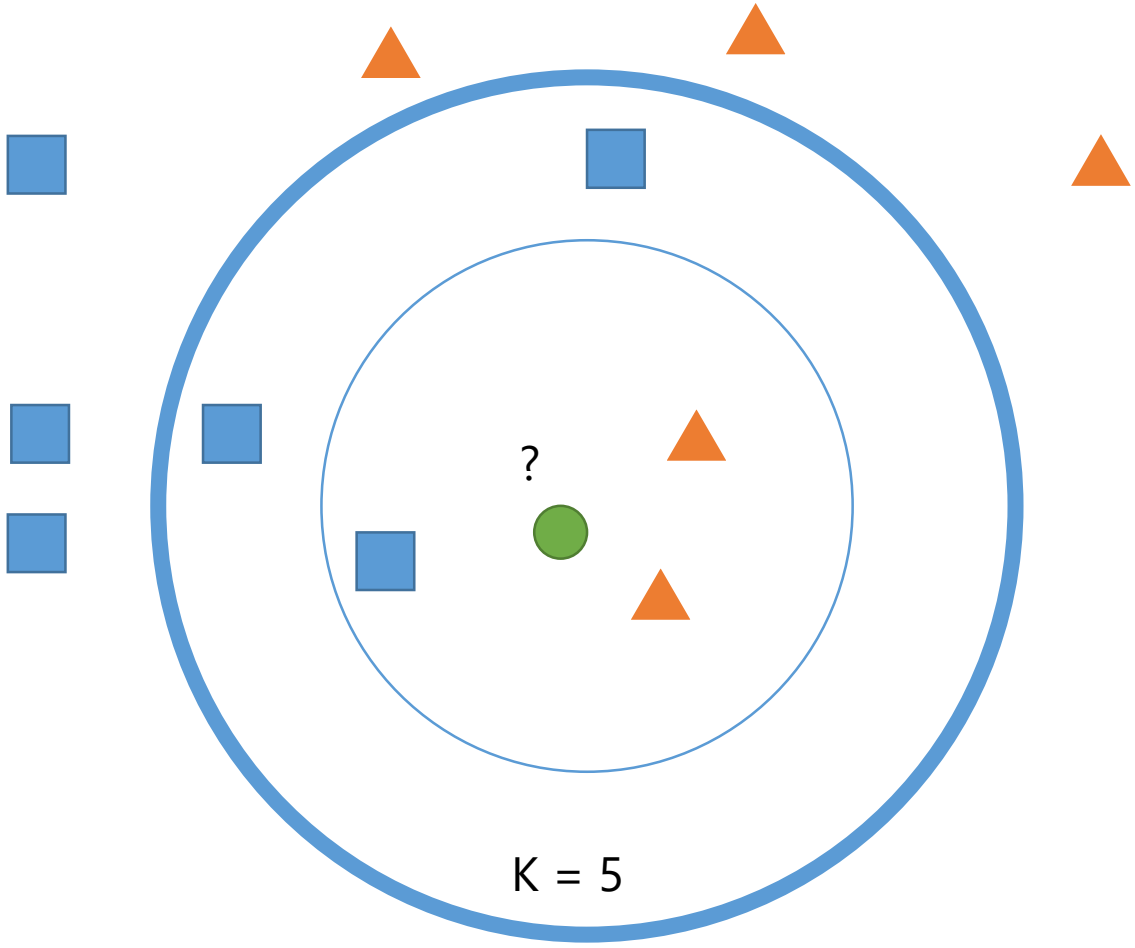
1 -> blue

2 orange > 1 blue

Green circle is a orange triangle

Amir Shokri

KNN



Green circle = ?

K = 5

Closest 5 points taken

2 -> orange

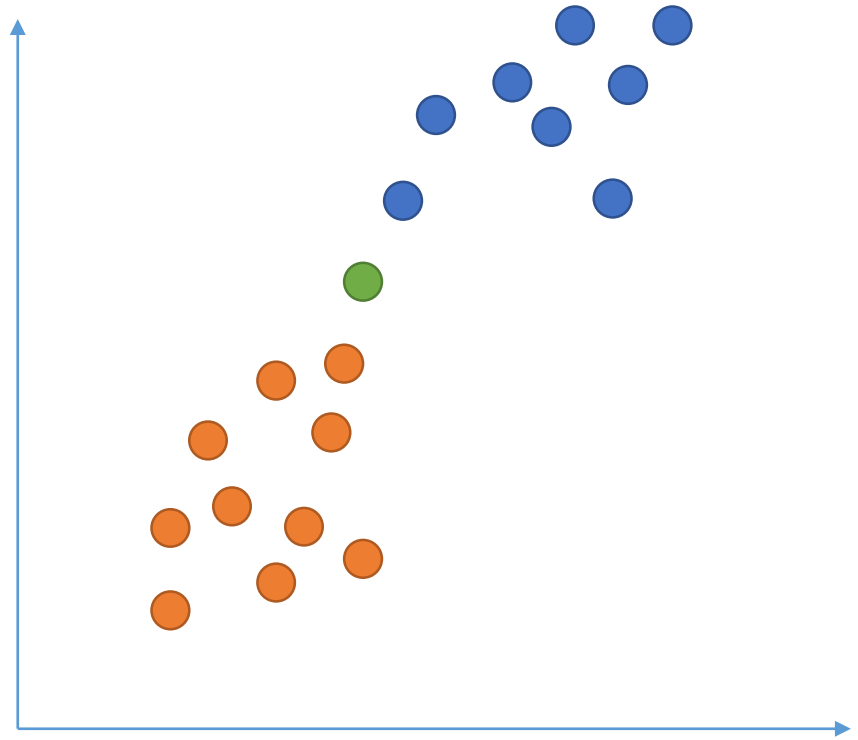
1 -> blue

2 orange < 3 blue

Green circle is a blue square

Amir Shokri

KNN



Green circle = ?

k = 1	:	orange
K = 3	:	orange
K = 5	:	orange
K = 7	:	blue
K = 9	:	orange

Amir Shokri

KNN

Choosing value of "K"

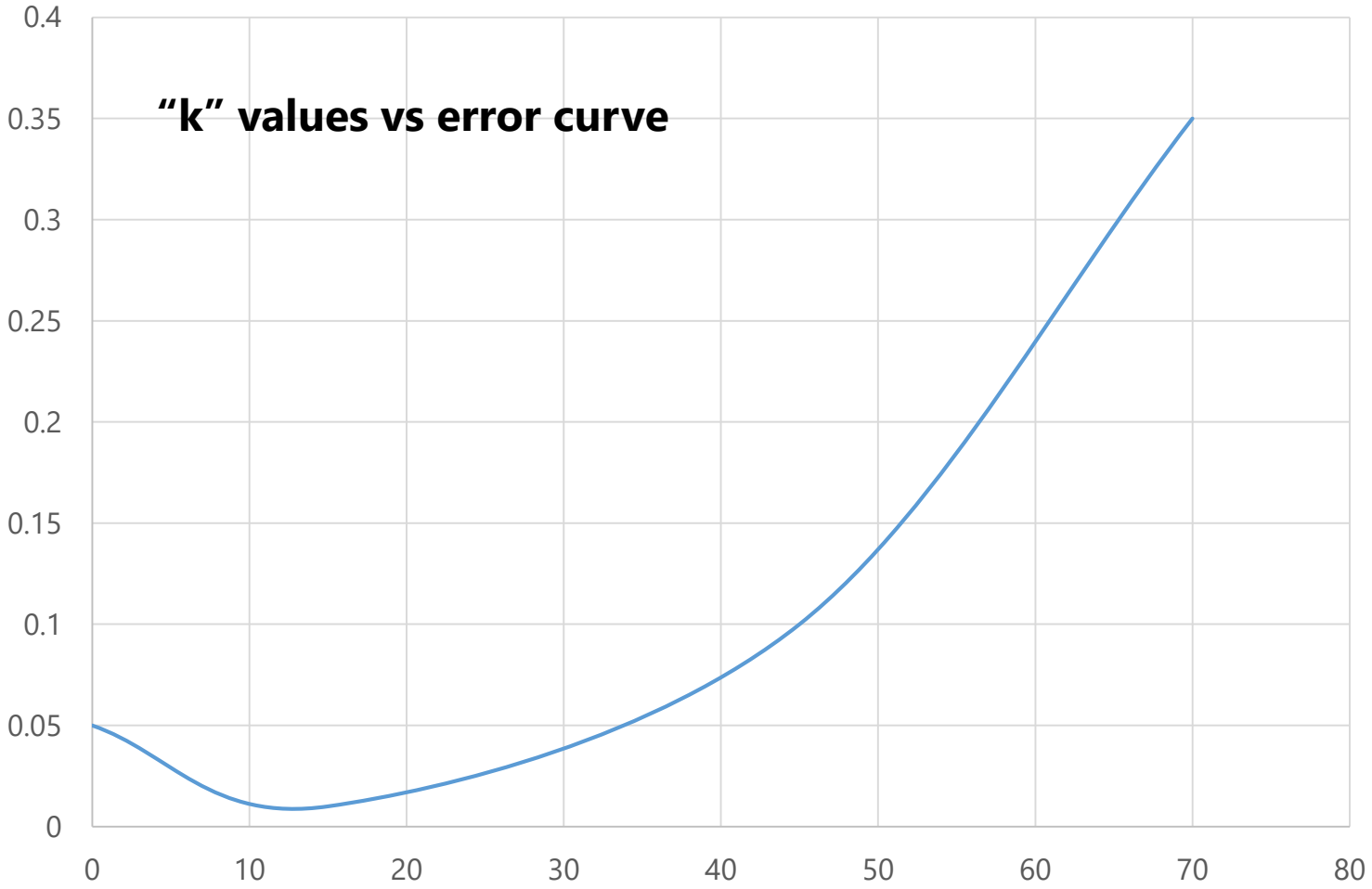
- "k" should be large so that error rate is minimized "k" too small will lead to noisy decision boundaries
- "k" should be small enough so that only nearby samples are included "k" too large will lead to over-smoothed boundaries
- Setting "k" to the square root of the number of training samples can lead to better results.

No Of features = 20

$k = \sqrt{20} = 4.4 \sim 4$

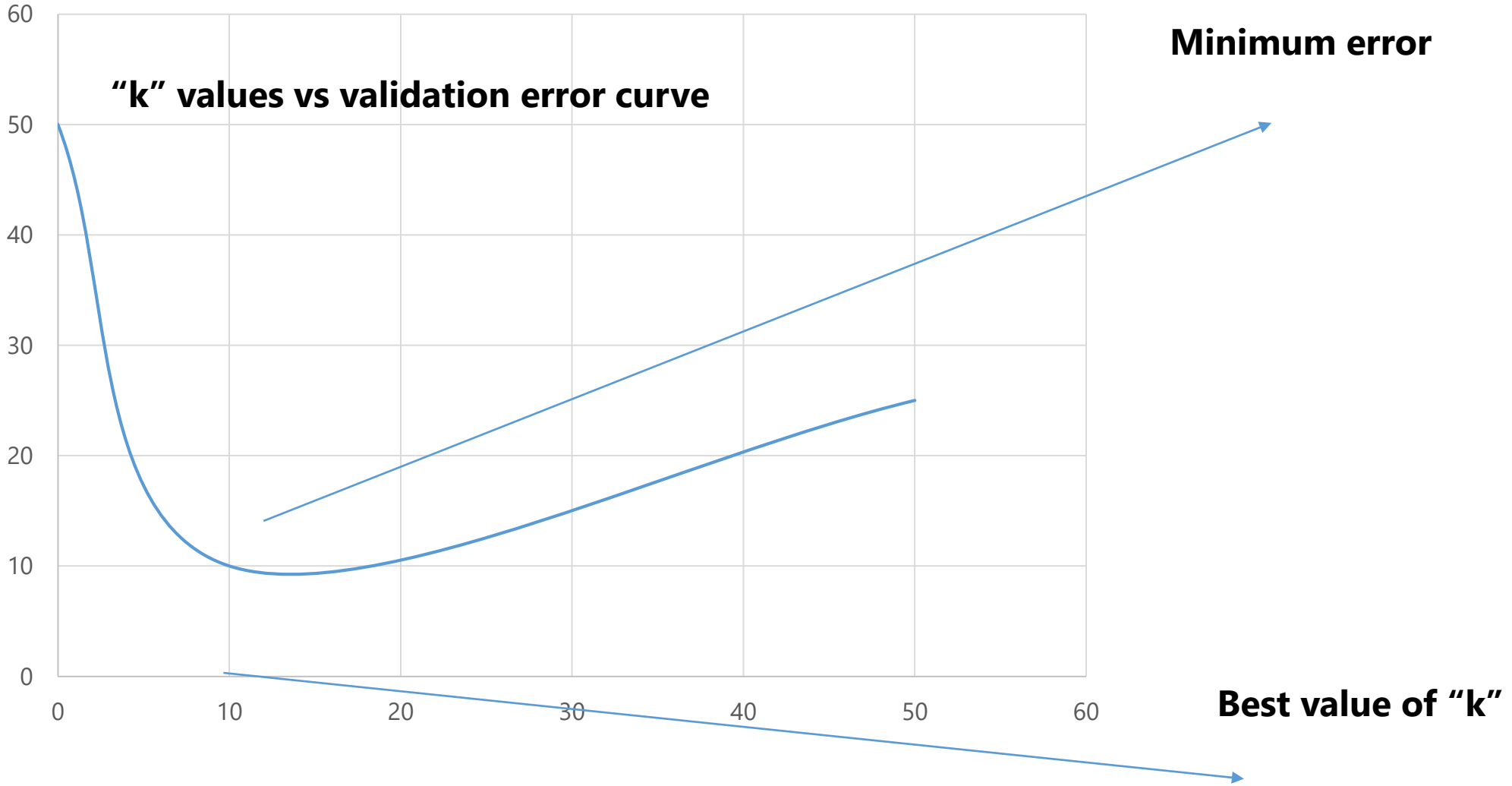
Amir Shokri

KNN



Amir Shokri

KNN



Amir Shokri

KNN

Pros.

- Non complex and very easy to understand and implement.
- Useful for non linear data as no assumptions about data.
- High accuracy (relatively), but no competitive compared to supervised learning algorithms.
- Can be used both for classification or regression.
- Best used where the probability distribution is unknown.

KNN

Cons.

- Computationally expensive.
- Lot of space is consumed as all the data points are stored.
- Sensitive to irrelevant features and the scale of the data.
- Output purely depends on k value chosen by user which can reduce accuracy for some values.

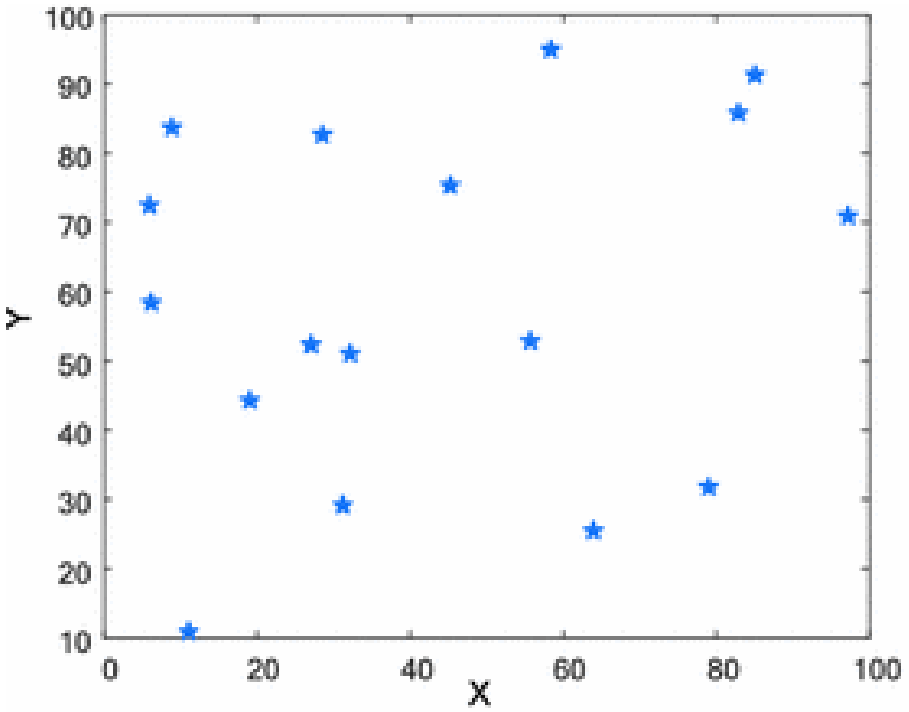
KNN

Applications.

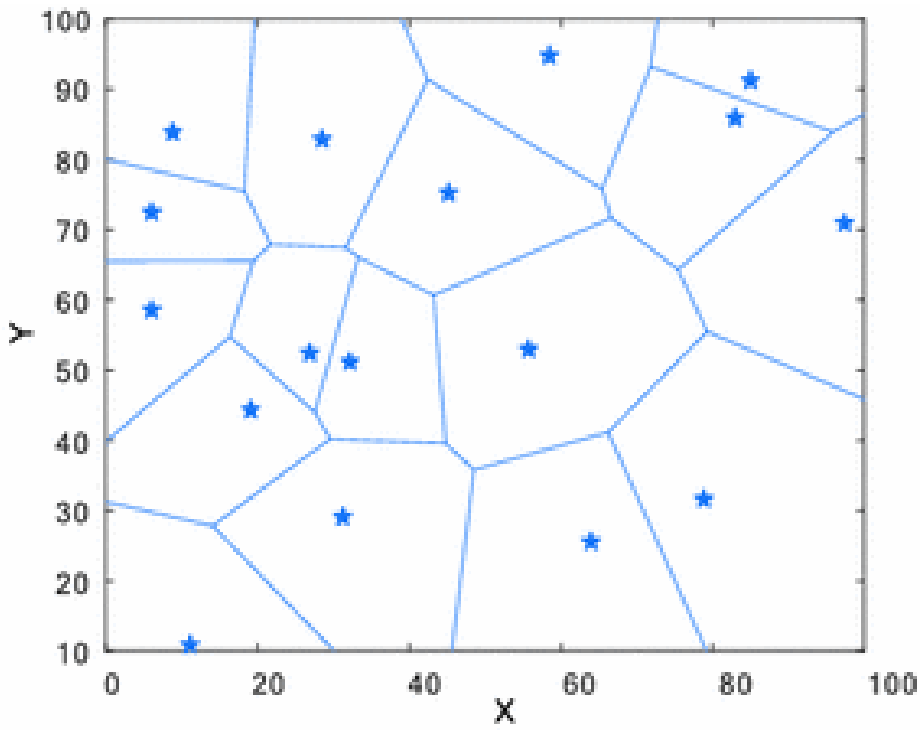
- Recommender systems
- Medicine
- Finance
- Text mining
- agriculture

Amir Shokri

KNN



(a) Initial points distribution



(b) Voronoi diagram by initial points

voronoi diagram

Amir Shokri

KNN

Bias : miss relevant relations within the data; everything gets skewed into specific classes -> underfitting.

Variance : too may hypotheses results in hallucination patterns -> overfitting

How To verify ? Split data in a test and training set

KNN

Bias : miss relevant relations within the data; everything gets skewed into specific classes -> underfitting.

Variance : too may hypotheses results in hallucination patterns -> overfitting

How To verify ? Split data in a test and training set

KNN

Will nearest-neighbours have high bias or high variance ?

-> high variance !

- Every instance is an hypothesis : the decision boundary is very/too detailed
- How can we solve this?
    - How to smooth the data obtain a more general model?
        - K-nearest neighbours (knn)
        - Weighted average
        - Prototypes
        - Locally weighted regresson

KNN

- For some value k take the k nearest neighbors of the new instance, and predict the class that is most common among these k neighbors
- Alleviates overfitting to a certain degree
    - Smoother decision boundaries
    - Influence of outliers is attenuated

KNN

- **What should k be ?**
  - Too low: overfitting
  - Too high : underfitting
- **Strategies:**
  - Cross-validation
  - Heuristics
  - algorithms