K Nearest Neighbour Classifier

Contents

Eager Learners vs Lazy Learners

- Eager learners, when given a set of training tuples, will construct a generalization model before receiving new (e.g., test) tuples to classify.
- Lazy learners simply stores data (or does only a little minor processing) and waits until it is given a test tuple.
- Lazy learners store the training tuples or “instances,” they are also referred to as instance based learners, even though all learning is essentially based on instances.
- Lazy learner: less time in training but more in predicting.

-**k- Nearest Neighbor Classifier**

-Case Based Classifier

k- Nearest Neighbor Classifier

- **History**
- It was first described in the early 1950s.
- The method is labor intensive when given large training sets.
- Gained popularity, when increased computing power became available.
- Used widely in area of pattern recognition and statistical estimation.

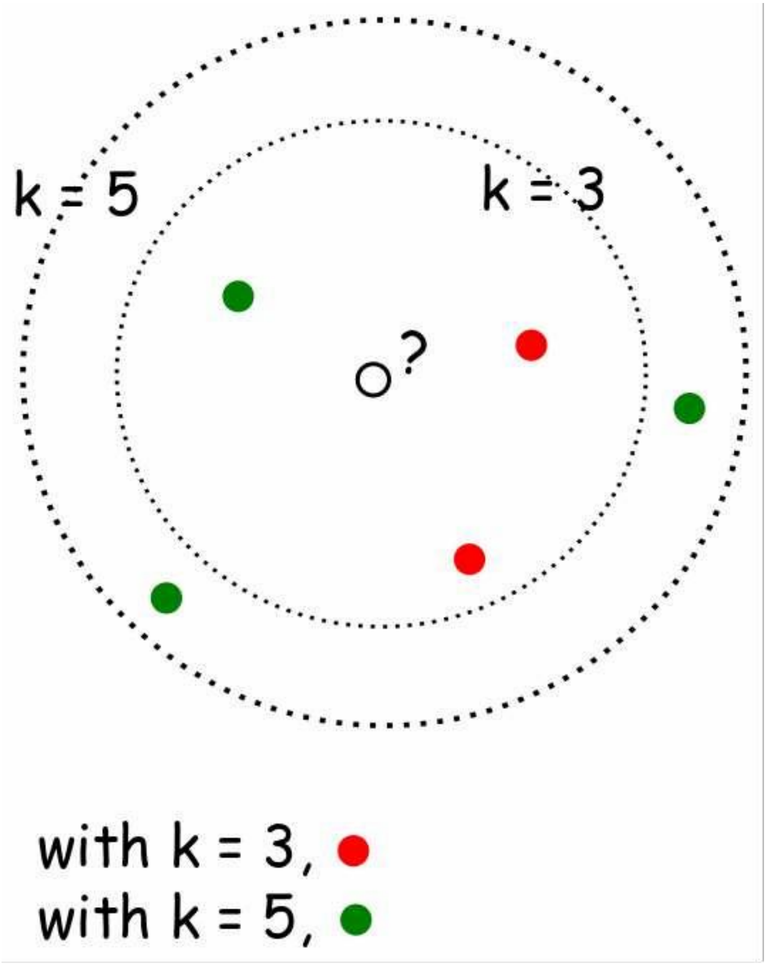
What is k- NN??

- Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.

- The training tuples are described by n attributes.

- When k = 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space.

When k=3 or k=5??



Remarks!!

- Similarity Function Based.
- Choose an odd value of k for 2 class problem.
- k must not be multiple of number of classes.

Closeness

- The Euclidean distance between two points or tuples, say,

X1 = (x11,x12,...,x1n) and X2 =(x21,x22,...,x2n), is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

- Min-max normalization can be used to transform a value v of a numeric attribute A to v0 in the range [0,1] by computing

$$v' = \frac{v - min_A}{max_A - min_A},$$

What if attributes are categorical??

- **How can distance be computed for attribute such as colour?**

-Simple Method: Compare corresponding value of attributes

-Other Method: Differential grading

What about missing values ??

- If the value of a given attribute A is missing in tuple X1 and/or in tuple X2, we assume the maximum possible difference.
- For categorical attributes, we take the difference value to be 1 if either one or both of the corresponding values of A are missing.
- If A is numeric and missing from both tuples X1 and X2, then the difference is also taken to be 1.

How to determine a good value for k?

- Starting with k = 1, we use a test set to estimate the error rate of the classifier.
- The k value that gives the minimum error rate may be selected.

KNN Algorithm and Example

Distance Measures

$$\textit{Euclidean distance}: \ d(x, y) = \sqrt{(x_i - y_i)^2}$$

$$\textit{Squared Euclidean distance}: \ d(x, y) = (x_i - y_i)^2$$

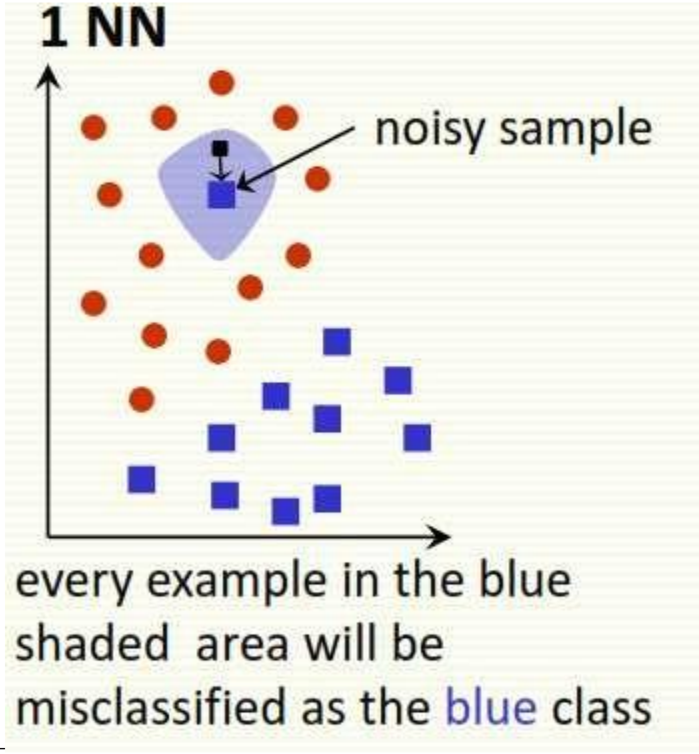
$$\textit{Manhattan distance}: \ d(x, y) = |(xi - yi)|$$

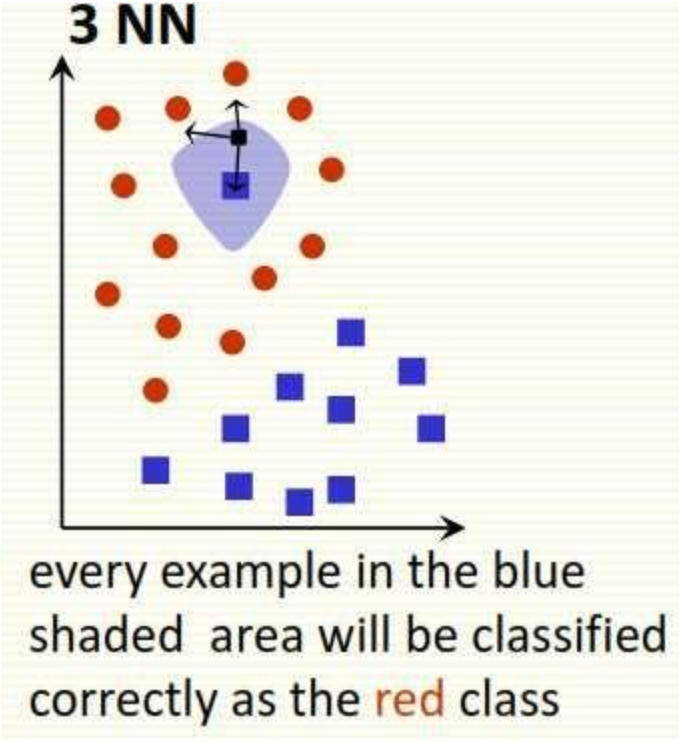
Which distance measure to use?

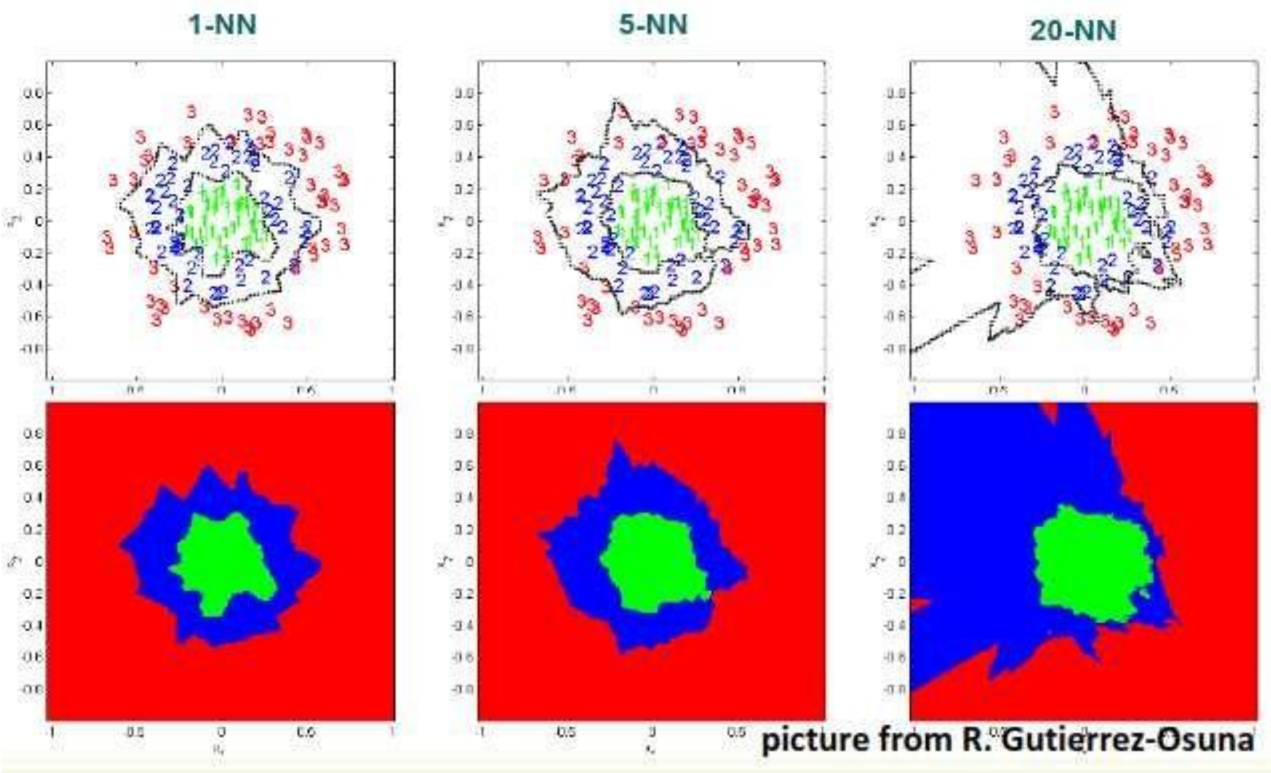
We use Euclidean Distance as it treats each feature as equally important.

How to choose K?

- If infinite number of samples available, the larger is k, the better is classification.
- k = 1 is often used for efficiency, but sensitive to “noise”

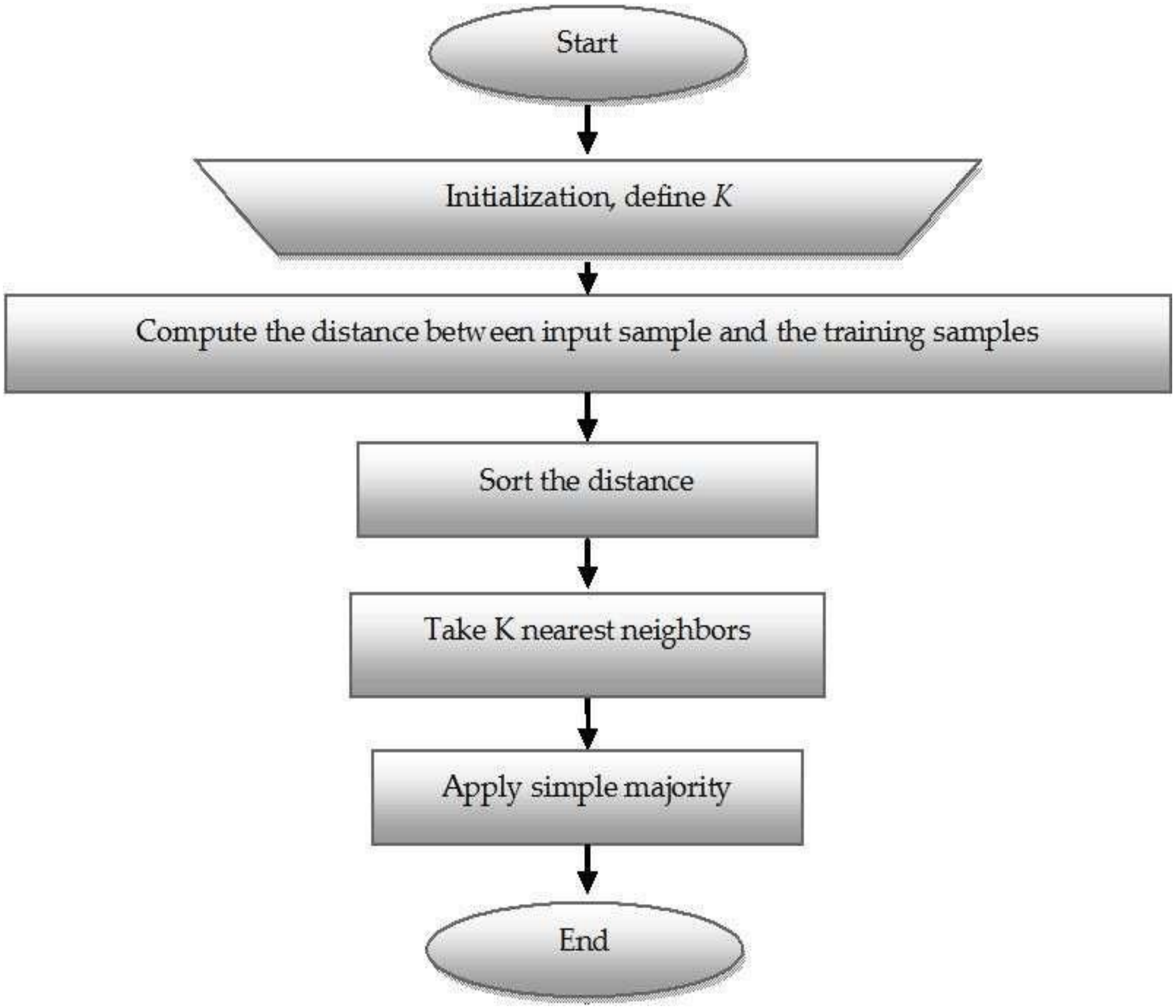






- Larger k gives smoother boundaries, better for generalization, but only if locality is preserved. Locality is not preserved if end up looking at samples too far away, not from the same class.
- Interesting relation to find k for large sample data : **k = sqrt(n)/2** where n is # of examples
- Can choose k through cross-validation

KNN Classifier Algorithm



Example

- We have data from the questionnaires survey and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here are four training samples :

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that passes the laboratory test with X1 = 3 and X2 = 7. Guess the classification of this new tissue.

- Step 1 : **Initialize and Define k.**

Lets say, k = 3

(Always choose k as an odd number if the number of attributes is even to avoid a tie in the class prediction)

- Step 2 : **Compute the distance between input sample and training sample**

- Co-ordinate of the input sample is (3,7).

- Instead of calculating the Euclidean distance, we calculate the Squared Euclidean distance.

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance
7	7	$(7-3)^2 + (7-7)^2 =$ **16**
7	4	$(7-3)^2 + (4-7)^2 =$ **25**
3	4	$(3-3)^2 + (4-7)^2 =$ **09**
1	4	$(1-3)^2 + (4-7)^2 =$ **13**

- Step 3 : **Sort the distance and determine the nearest neighbours based of the Kth minimum distance :**

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance	Rank minimum distance	Is it included in 3-Nearest Neighbour?
7	7	16	**3**	**Yes**
7	4	25	**4**	**No**
3	4	09	**1**	**Yes**
1	4	13	**2**	**Yes**

- Step 4 : **Take 3-Nearest Neighbours:**
- Gather the category Y of the nearest neighbours.

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance	Rank minimum distance	Is it included in 3-Nearest Neighbour?	Y = Category of the nearest neighbour
7	7	16	3	**Yes**	**Bad**
7	4	25	4	**No**	-
3	4	09	1	**Yes**	**Good**
1	4	13	2	**Yes**	**Good**

- Step 5 : **Apply simple majority**

- Use simple majority of the category of the nearest neighbours as the prediction value of the query instance.

- We have 2 “good” and 1 “bad”. Thus we conclude that the new paper tissue that passes the laboratory test with X1 = 3 and X2 = 7 is included in the “good” category.

Iris Dataset Example using Weka

- Iris dataset contains 150 sample instances belonging to 3 classes. 50 samples belong to each of these 3 classes.
- **Statistical observations** :
- Let's denote the true value of interest as θ ($expected$) and the value estimated using some algorithm as θ. ($observed$)
- Kappa Statistics : The kappa statistic measures the agreement of prediction with the true class -- 1.0 signifies complete agreement. It measures the significance of the classification with respect to the observed value and expected value.
- Mean absolute error:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\theta}_i - \theta_i|$$

- Root Mean Square Error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2}$$

- Relative Absolute Error:

$$\mathrm{RAE} = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta} - \theta_i|}$$

- Root Relative Squared o Err

$$\mathrm{RRSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2}{\sum_{i=1}^{N}\left(\bar{\theta} - \theta_i\right)^2}}$$

Complexity

- Basic kNN algorithm stores all examples
- Suppose we have n examples each of dimension d
- **O(d)** to compute distance to one examples
- **O(nd)** to computed distances to all examples
- Plus **O(nk)** time to find k closest examples
- Total time: **O(nk+nd)**
- Very expensive for a large number of samples
- But we need a large number of samples for kNN to to work well!!

Advantages of KNN classifier :

- Can be applied to the data from any distribution for example, data does not have to be separable with a linear boundary
- Very simple and intuitive
- Good classification if the number of samples is large enough

Disadvantages of KNN classifier :

- Choosing k may be tricky
- Test stage is computationally expensive
- No training stage, all the work is done during the test stage
- This is actually the opposite of what we want. Usually we can afford training step to take a long time, but we want fast test step

Applications of KNN Classifier

- Used in classification
- Used to get missing values
- Used in pattern recognition
- Used in gene expression
- Used in protein-protein prediction
- Used to get 3D structure of protein
- Used to measure document similarity

Comparison of various classifiers

Algorithm	Features	Limitations
C4.5 Algorithm	- Models built can be easily interpreted - Easy to implement - Can use both discrete and continuous values - Deals with noise	- Small variation in data can lead to different decision trees - Does not work very well on small training dataset - Over-fitting
ID3 Algorithm	- It produces more accuracy than C4.5 - Detection rate is increased and space consumption is reduced	- Requires large searching time - Sometimes it may generate very long rules which are difficult to prune - Requires large amount of memory to store tree
K-Nearest Neighbour Algorithm	- Classes need not be linearly separable - Zero cost of the learning process - Sometimes it is robust with regard to noisy training data - Well suited for multimodal	- Time to find the nearest neighbours in a large training dataset can be excessive - It is sensitive to noisy or irrelevant attributes - Performance of the algorithm depends on the number of

Naïve Bayes Algorithm	- Simple to implement - Great computational efficiency and classification rate - It predicts accurate results for most of the classification and prediction problems	- The precision of the algorithm decreases if the amount of data is less - For obtaining good results, it requires a very large number of records
Support vector machine Algorithm	- High accuracy - Work well even if the data is not linearly separable in the base feature space	- Speed and size requirement both in training and testing is more - High complexity and extensive memory requirements for classification in many cases
Artificial Neural Networks Algorithm	- It is easy to use with few parameters to adjust - A neural network learns and reprogramming is not needed. - Easy to implement - Applicable to a wide range of problems in real life.	- Requires high processing time if neural network is large - Difficult to know how many neurons and layers are necessary - Learning can be slow

Conclusion

- KNN is what we call *lazy learning (vs. eager learning)*
- Conceptually simple, easy to understand and explain
- Very flexible decision boundaries
- Not much learning at all!
- It can be hard to find a good distance measure
- Irrelevant features and noise can be very detrimental
- Typically can not handle more than a few dozen attributes
- Computational cost: requires a lot computation

References

- “Data Mining : Concepts and Techniques”, J. Han, J. Pei, 2001
- “A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining”, Sakshi, S. Khare, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 8, ISSN: 2321-8169
- “Comparison of various classification algorithms on iris datasets using WEKA”, Kanu Patel et al, IJAERD, Volume 1 Issue 1, February 2014, ISSN: 2348 - 4470

Thank you ☺