Inspire…Educate…Transform.

KNN

Dr. Manish Gupta

Sr. Mentor – Academics, INSOFE

KNN-Algorithm

- Assign a class to a new data point based on its neighbors (mode)
- Identify a numeric value of a new data point based on its neighbors (mean/median)
- Weighted mean/mode of entire data
- It is also called as instance based learning (IBL), case based reasoning (CBR), lazy learning.

Process is simple

- Pick a number of neighbors you want to use for classification or regression (K)
- Choose a method to measure distances (same consideration as clustering)
- Keep a data set with records

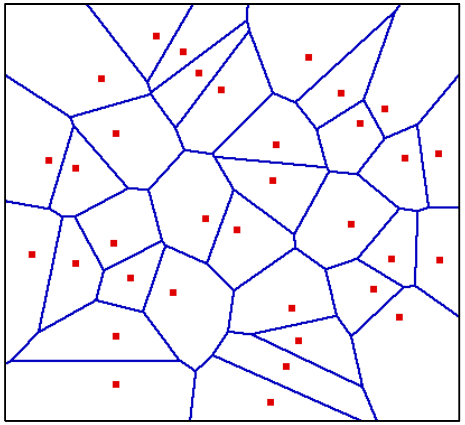
Process

- For every new point, identify the number of nearest neighbors you picked using the method you chose

- Let them vote if it is a classification issue or take a mean/median for regression!

Observations

- Decision surfaces created by KNN:
  - Voronoi Diagrams: Each point in a convex hull is closest to the sample inside the convex hull than to any other sample
  - Much more complex than decision trees!
    - http://www.raymondhill.net/voronoi/rhill-voronoi.html
    - http://www.pi6.fernuni-hagen.de/GeomLab/VoroGlide/
- Theoretical guarantee
  - Noisy training is an issue
  - Can overfit

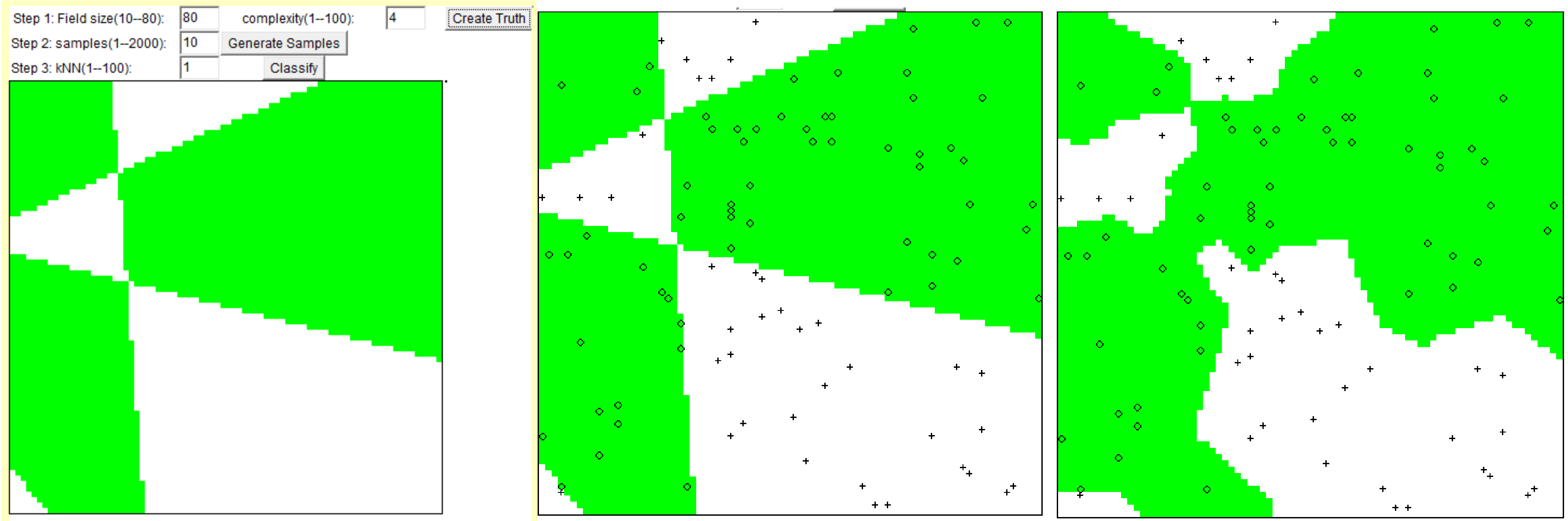


Let us play

- http://sleepyheads.jp/apps/knn/knn.html
  - (Two class example)
- http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
  - Changing K allows you to model true distribution

Behavior of KNN

http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html



K=1, error rate=9.55%

- Experiment with K to see how decision surface is obtained

Observations

- As the complexity of space grows, accuracy comes down and you need more data
- Increasing K can reduce the overfit, and accuracy would improve. But beyond a value, accuracy starts decreasing. There will be an optimum K.

Issues with KNN and instance based techniques

- Curse of dimensionality
  - For the k nearest neighbor rule to perform well, we want the neighbours to be representative of the population densities at the query point (the given value x to be classified). Which is to say that the k nearest neighbours should typically fall near that point x.
  - We can expect that to happen in low dimensions: a unit interval, for example, with 5000 points the 5 nearest neighbours would be in a neighborhood of length 0.001, in average, which seems right; the 5000 points will cover decently our space, even when taken in groups of 5: we can expect that the 5 neighbours will be quite near x.
  - But, say, in 6 dimensions, we cannot be so optimistic: 5000 points in this cube means that we have in average 5 points for each 6-dim-cube of size length=0.31, so the 5 nearest neighbours for a given query point will not be, in average, very near to it.
- Requires more memory and more time

Attributes

Records

Search process

ENGINEERING K-NN

Attributes

- Scaling the attributes is important
  - Attributes with larger range can dominate
  - Categorical variables and Ordinal variables need to be handled nicely

Curse of dimensionality

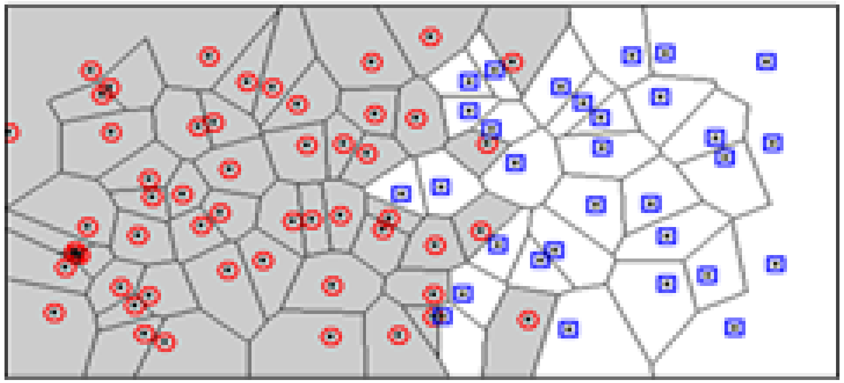
- K-NN is heavily impacted by high dimensionality.
- Reduce the dimensions
  - Correlation
  - Info gain (Can lose some that are important; it assumes independent attributes)
    - Wrapper methods
      - Forward selection, Backward elimination
  - Weighting attributes
    - Think PCA

R-KNN

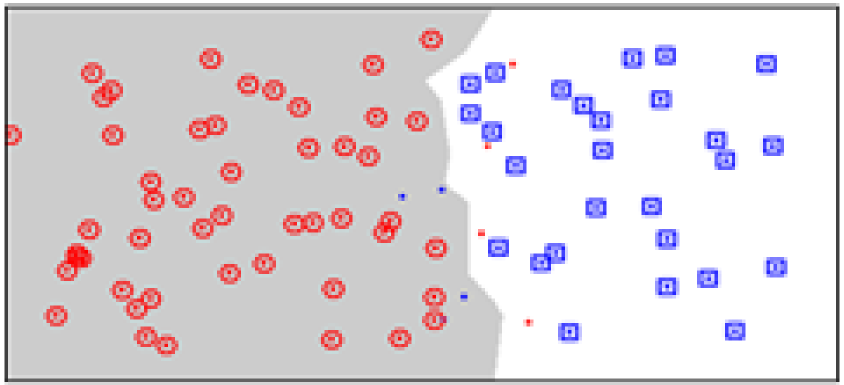
- The same concept of randomforest but for k-nn
  - http://www.biomedcentral.com/1471-2105/12/450
  - Specifically, a collection of r different KNN classifiers will be generated. Each one takes a random subset of the input variables.
  - Each KNN classifier classifies a test point by its majority, or weighted majority class, of its k nearest neighbors. The final classification in each case is determined by majority voting of r KNN classifications.
- Can be used for feature selection
  - In order to select a subset of variables that have classification capability, the key is to define some criteria to rank the variables. We define a measure, called support.
  - Each feature f will appear in some KNN classifiers, say, set C(f) of size M, where M is the multiplicity of f. In turn, each classifier c in C(f) is an evaluator of its m features, say, set F(c).
  - We can take its accuracy as a performance measure for those features. The mean accuracy of these KNN classifiers (support) is a measure of the feature relevance with the outcome. Thus we have a ranking of the features.
  - We call this scheme bidirectional voting. Each feature randomly participates in a series of KNNs to cast a vote for classification. In turn, each classification result casts a vote for each participating feature.

Wilson editing

Overlapping classes



Original data



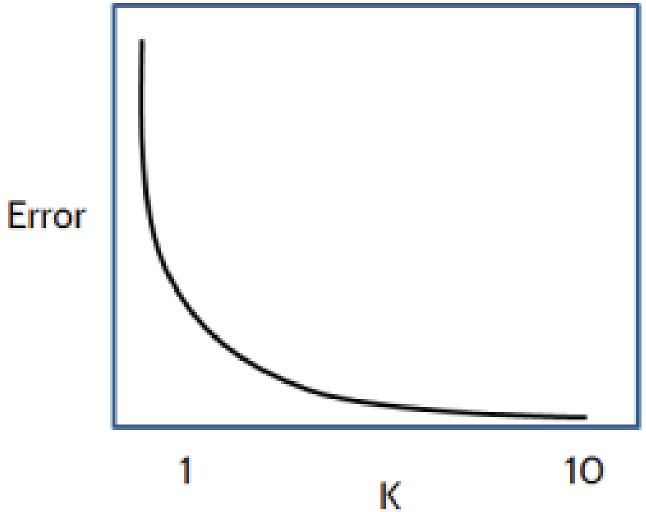
Wilson editing with k=7

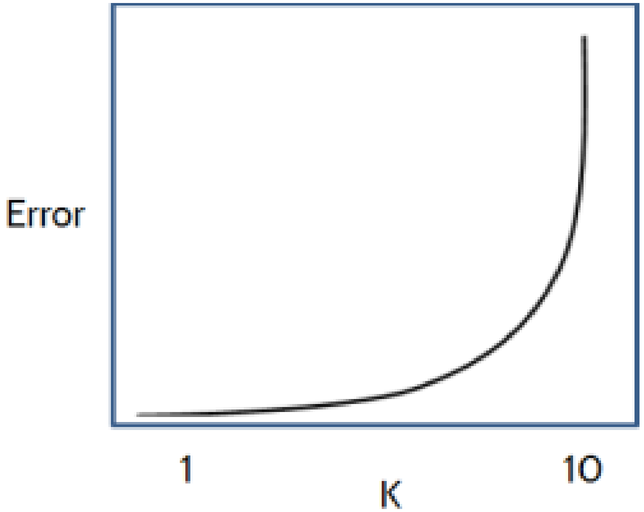
One method for outlier removal

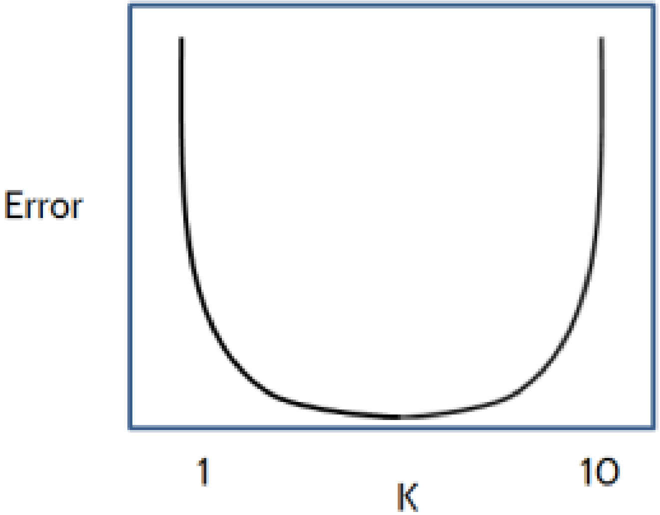
Remove points that do not agree with the majority of their k nearest neighbours

CSE 7306c

Use a K that gives least error on test data







Records: Handling missing values

- K-NN is impacted heavily by missing values
- Imputation is one option and can be done using multiple methods like
  - Case deletion: discarding all instances (cases) with missing values for at least one feature
  - Mean imputation
  - Median imputation
  - KNN imputation

Datasets	KNN			
	CD	MI	MDI	KNNI
Hepatitis	28.95	38.32	37.67	39.23
Heartc	19.42	18.79	18.62	18.70
Crx	25.09	25.20	24.71	24.58
Breastw	3.41	3.84	3.88	3.61

Cross-validation errors for KNN classifier using the four methods to deal with missing data

Acuna, Edgar, and Caroline Rodriguez. "The treatment of missing values and its effect on classifier accuracy." *Classification, clustering, and data mining applications*. Springer Berlin Heidelberg, 2004. 639-647.

Speeding up Using K-Means

- Come up with rough approximations to eliminate most points (distance between centroids of K-Means) and then apply elaborate measurements (K-NN) on closest points
  - Find the closest cluster and then compute exact K neighbors from the cluster members.

Summary

- Scaling the data
- Address dimensionality:
    - Correlation
    - Coarse approximations
    - R-KNN
- Remove outliers
    - Condensation of data
- Addressing speed
    - Condensation
    - KMeans
    - LSH

Inspire...Educate...Transform.

International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering