**Skill test Questions and Answers**

**1) [True or False] k-NN algorithm does more computation on test time rather than train time.**

A) TRUE  
B) FALSE

**Solution: A**

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the testing phase, a test point is classified by assigning the label which are most frequent among the *k* training samples nearest to that query point – hence higher computation.

**2) In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.**

A) 3  
B) 10  
C) 20  
D 50 

**Solution: B**

Validation error is the least when the value of k is 10. So it is best to use this value of k

**3) Which of the following distance metric can not be used in k-NN?**

A) Manhattan  
B) Minkowski  
C) Tanimoto  
D) Jaccard  
E) Mahalanobis  
F) All can be used

**Solution: F**

All of these distance metric can be used as a distance metric for k-NN.

**4) Which of the following option is true about k-NN algorithm?**

A) It can be used for classification  
B) It can be used for regression  
C) It can be used in both classification and regression

**Solution: C**

We can also use k-NN for regression problems. In this case the prediction can be based on the mean or the median of the k-most similar instances.

**5) Which of the following statement is true about k-NN algorithm?**

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

A) 1 and 2  
B) 1 and 3  
C) Only 1  
D) All of the above

**Solution: D**

The above mentioned statements are assumptions of kNN algorithm

**6) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?**

A) K-NN  
B) Linear Regression  
C) Logistic Regression

**Solution: A**

k-NN algorithm can be used for imputing missing value of both categorical and continuous variables.

**7) Which of the following is true about Manhattan distance?**

A) It can be used for continuous variables  
B) It can be used for categorical variables  
C) It can be used for categorical as well as continuous  
D) None of these

**Solution: A**

Manhattan Distance is designed for calculating the distance between real valued features.

**8) Which of the following distance measure do we use in case of categorical variables in k-NN?**

1. Hamming Distance
2. Euclidean Distance
3. Manhattan Distance

A) 1  
B) 2  
C) 3  
D) 1 and 2  
E) 2 and 3  
F) 1,2 and 3

**Solution: A**

Both Euclidean and Manhattan distances are used in case of continuous variables, whereas hamming distance is used in case of categorical variable.

**9) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?**

A) 1  
B) 2  
C) 4  
D) 8

**Solution: A**

sqrt( (1-2)^2 + (3-3)^2) = sqrt(1^2 + 0^2) = 1

**10) Which of the following will be Manhattan Distance between the two data point A(1,3) and B(2,3)?**

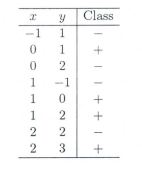
A) 1  
B) 2  
C) 4  
D) 8

**Solution: A**

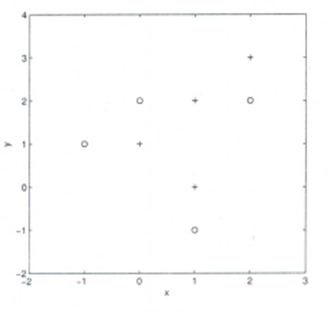
sqrt( mod((1-2)) + mod((3-3))) = sqrt(1 + 0) = 1

**Context: 11-12**

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/01184933/Pic_C_11_12.jpg)

Below is a scatter plot which shows the above data in 2D space.

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/01185059/Pic_C_11_12_b.JPG_.jpg)

**11) Suppose, you want to predict the class of new data point x=1 and y=1 using eucludian distance in 3-NN. In which class this data point belong to?**

A) + Class

B) – Class

C) Can’t say

D) None of these

**Solution: A**

All three nearest point are of +class so this point will be classified as +class.

**12) In the previous question, you are now want use 7-NN instead of 3-KNN which of the following x=1 and y=1 will belong to?**

A) + Class

B) – Class

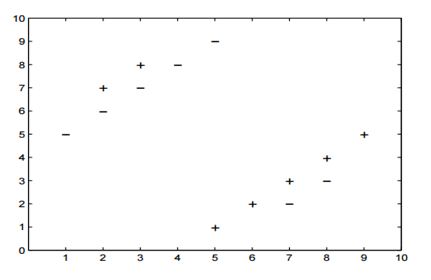
C) Can’t say

**Solution: B**

Now this point will be classified as – class because there are 4 – class and 3 +class point are in nearest circle.

**Context 13-14:**

Suppose you have given the following 2-class data where “+” represent a postive class and “” is represent negative class.

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/01185614/Pic_c_13.jpg)

**13) Which of the following value of k in k-NN would minimize the leave one out cross validation accuracy?**

A) 3  
B) 5  
C) Both have same  
D) None of these

**Solution: B**

5-NN will have least leave one out cross validation error.

**14) Which of the following would be the leave on out cross validation accuracy for k=5?**

A) 2/14  
B) 4/14  
C) 6/14  
D) 8/14  
E) None of the above

**Solution: E**

In 5-NN we will have  10/14 leave one out cross validation accuracy.

**15) Which of the following will be true about k in k-NN in terms of Bias?**

A) When you increase the k the bias will be increases  
B) When you decrease the k the bias will be increases  
C) Can’t say  
D) None of these

**Solution: A**

large K means simple model, simple model always condider as high bias

**16) Which of the following will be true about k in k-NN in terms of variance?**

A) When you increase the k the variance will increases  
B) When you decrease the k the variance will increases  
C) Can’t say  
D) None of these

**Solution: B**

Simple model will be consider as less variance model

**17) The following two distances(Eucludean Distance and Manhattan Distance) have given to you which generally we used in K-NN algorithm. These distance are between two points A(x1,y1) and B(x2,Y2).**

**Your task is to tag the both distance by seeing the following two graphs. Which of the following option is true about below graph ?**

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/01190245/Pic_17.jpg)

A) Left is Manhattan Distance and right is euclidean Distance  
B) Left is Euclidean Distance and right is Manhattan Distance  
C) Neither left or right are a Manhattan Distance  
D) Neither left or right are a Euclidian Distance

**Solution: B**

Left is the graphical depiction of how euclidean distance works, whereas right one is of Manhattan distance.

**18) When you find noise in data which of the following option would you consider in k-NN?**

A) I will increase the value of k  
B) I will decrease the value of k  
C) Noise can not be dependent on value of k  
D) None of these

**Solution: A**

To be more sure of which classifications you make, you can try increasing the value of k.

**19) In k-NN it is very likely to overfit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?**

1. Dimensionality Reduction
2. Feature selection

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

In such case you can use either dimensionality reduction algorithm or the feature selection algorithm

**20) Below are two statements given. Which of the following will be true both statements?**

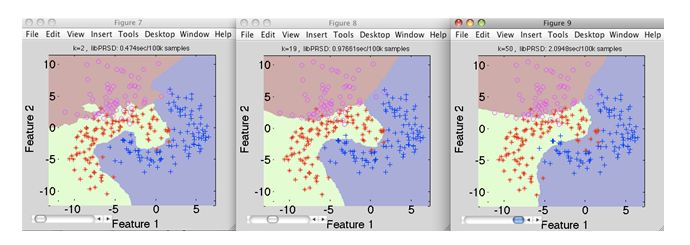
1. k-NN is a memory-based approach is that the classifier immediately adapts as we collect new training data.
2. The computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario.

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

Both are true and self explanatory

**21) Suppose you have given the following images(1 left, 2 middle and 3 right), Now your task is to find out the value of k in k-NN in each image where k1 is for 1st, k2 is for 2nd and k3 is for 3rd figure.**

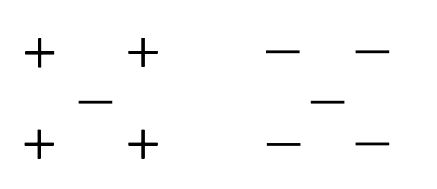
[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/02162508/Pic_341.jpg)

A) k1 > k2> k3  
B) k1<k2  
C) k1 = k2 = k3  
D) None of these

**Solution: D**

Value of k is highest in k3, whereas in k1 it is lowest

**22) Which of the following value of k in the following graph would you give least leave one out cross validation accuracy?**

[](https://cdn.analyticsvidhya.com/wp-content/uploads/2017/08/01193852/Pic_22.jpg)

A) 1  
B) 2  
C) 3  
D) 5

**Solution: B**

If you keep the value of k as 2, it gives the lowest cross validation accuracy. You can try this out yourself.

**23) A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?**

**Note: Model has successfully deployed and no technical issues are found at client side except the model performance**  
A) It is probably a overfitted model  
B) It is probably a underfitted model  
C) Can’t say  
D) None of these

**Solution: A**

In an overfitted module, it seems to be performing well on training data, but it is not generalized enough to give the same results on a new data.

**24) You have given the following 2 statements, find which of these option is/are true in case of k-NN?**

1. In case of very large value of k, we may include points from other classes into the neighborhood.
2. In case of too small value of k the algorithm is very sensitive to noise

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

Both the options are true and are self explanatory.

**25) Which of the following statements is true for k-NN classifiers?**

A) The classification accuracy is better with larger values of k  
B) The decision boundary is smoother with smaller values of k  
C) The decision boundary is linear  
D) k-NN does not require an explicit training step

**Solution: D**

Option A: This is not always true. You have to ensure that the value of k is not too high or not too low.

Option B: This statement is not true. The decision boundary can be a bit jagged

Option C: Same as option B

Option D: This statement is true

**26) True-False: It is possible to construct a 2-NN classifier by using the 1-NN classifier?**

A) TRUE  
B) FALSE

**Solution: A**

You can implement a 2-NN classifier by ensembling 1-NN classifiers

**27) In k-NN what will happen when you increase/decrease the value of k?**

A) The boundary becomes smoother with increasing value of K  
B) The boundary becomes smoother with decreasing value of K  
C) Smoothness of boundary doesn’t dependent on value of K  
D) None of these

**Solution: A**

The decision boundary would become smoother by increasing the value of K

**28) Following are the two statements given for k-NN algorthm, which of the statement(s)**

**is/are true?**

1. We can choose optimal value of k with the help of cross validation
2. Euclidean distance treats each feature as equally important

A) 1  
B) 2  
C) 1 and 2  
D) None of these

**Solution: C**

Both the statements are true

**Context 29-30:**

Suppose, you have trained a k-NN model and now you want to get the prediction on test data. Before getting the prediction suppose you want to calculate the time taken by k-NN for predicting the class for test data.  
Note: Calculating the distance between 2 observation will take D time.

**29) What would be the time taken by 1-NN if there are N(Very large) observations in test data?**

A) N\*D  
B) N\*D\*2  
C) (N\*D)/2  
D) None of these

**Solution: A**

The value of N is very large, so option A is correct

**30) What would be the relation between the time taken by 1-NN,2-NN,3-NN.**

A) 1-NN >2-NN >3-NN  
B) 1-NN < 2-NN < 3-NN  
C) 1-NN ~ 2-NN ~ 3-NN  
D) None of these

**Solution: C**

The training time for any value of k in kNN algorithm is the same.

### 1. What is “K” in KNN algorithm?

K = Number of nearest neighbors you want to select to predict the class of a given item

### 2. How do we decide the value of “K” in KNN algorithm?

If K is small, then results might not be reliable because noise will have a higher influence on the result. If K     is large, then there will be a lot of processing which may adversely impact the performance of the algorithm. So, following is must be considered while choosing the value of K:

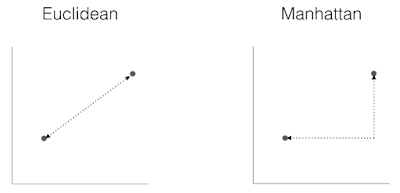
**a.** K should be the **square root** of n (number of data points in training dataset)  
**b.** K should be **odd**so that there are no ties. If square root is even, then add or subtract 1 to it.

### 3. Why is the odd value of “K” preferable in KNN algorithm?

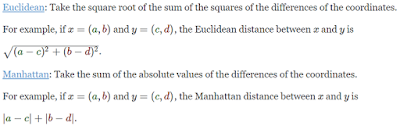
K should be odd so that there are no ties in the voting. If square root of number of data points is even, then add or subtract 1 to it to make it odd.

### 4. What is the difference between Euclidean Distance and Manhattan distance? What is the formula of Euclidean distance and Manhattan distance?

Both are used to find out the distance between two points.

[](https://4.bp.blogspot.com/-9iDGWtgrbh0/XErhbogDiBI/AAAAAAAABLE/tAhILG2rJ68Hs88XBSi5PP0Wkxi3F-U2ACLcBGAs/s1600/Euclidean_distance_and_Manhattan_distance.PNG)

**Euclidean Distance and Manhattan Distance Formula**

[](https://4.bp.blogspot.com/-Tr6BrJ4mZNw/XErlXGm8xrI/AAAAAAAABLc/ZWSmXXOrQBIyKmeO4DdPvckpUVjDFHW7wCLcBGAs/s1600/Euclidean_distance_and_Manhattan_distance_formula.PNG)

### 5. Why is KNN algorithm called Lazy Learner?

When it gets the training data, it does not learn and make a model, it just stores the data. It does not derive any discriminative function from the training data. It uses the training data when it actually needs to do some prediction. So, KNN does not immediately learn a model, but delays the learning, that is why it is called lazy learner.

### 6.Why KNN is non-parametric?

Non-parametric means not making any assumptions on the underlying data distribution. Non-parametric methods do not have fixed numbers of parameters in the model. Similarly in KNN, model parameters actually grows with the training data set – you can imagine each training case as a “parameter” in the model.

### 7.How to handle categorical variables in KNN?

Create dummy variables out of a categorical variable and include them instead of original categorical variable. Unlike regression, create k dummies instead of (k-1). For example, a categorical variable named “Department” has 5 unique levels / categories. So we will create 5 dummy variables. Each dummy variable has 1 against its department and else 0.

### 8.How to find best K value?

Cross-validation is a smart way to find out the optimal K value. It estimates the validation error rate by holding out a subset of the training set from the model building process.

Cross-validation (let’s say 10 fold validation) involves randomly dividing the training set into 10 groups, or folds, of approximately equal size. 90% data is used to train the model and remaining 10% to validate it. The misclassification rate is then computed on the 10% validation data. This procedure repeats 10 times. Different group of observations are treated as a validation set each of the 10 times. It results to 10 estimates of the validation error which are then averaged out.

## **Can KNN be used for regression?**

*Yes, K-nearest neighbor can be used for regression. In other words, K-nearest neighbor algorithm can be applied  when dependent variable is continuous. In this case, the predicted value is the average of the values of its k nearest neighbors.*

### 9. Why should we not use KNN algorithm for large datasets?

KNN works well with smaller dataset because it is a lazy learner. It needs to store all the data and then makes decision only at run time. It needs to calculate the distance of a given point with all other points. So if dataset is large, there will be a lot of processing which may adversely impact the performance of the algorithm.

KNN is also very sensitive to noise in the dataset. If the dataset is large, there are chances of noise in the dataset which adversely affect the performance of KNN algorithm.

### **10.How to choose optimal value of K in KNN Algorithm?**

**1. Square Root Method**: Take square root of the number of samples in the training dataset.

**2. Cross Validation Method:** We should also use cross validation to find out the optimal value of K in KNN. Start with K=1, run cross validation (5 to 10 fold),  measure the accuracy and keep repeating till the results become consistent.

K=1, 2, 3… As K increases, the error usually goes down, then stabilizes, and then raises again. Pick the optimum K at the beginning of the stable zone. This is also called **Elbow Method.**

**3. Domain Knowledge** also plays a vital role while choosing the optimum value of K.

**4.**K should be an **odd number**.

### **11.How is KNN different from k-means clustering?**

KNN and k-means clustering both are very different algorithms that solve different problems and have their own meanings of what the variable ‘k’ is.  KNN is a supervised classification algorithm that will label new data points based on the ‘k’ number of nearest data points and k-means clustering is an unsupervised clustering algorithm that groups the data into ‘k’ number of clusters.

### **12.What are potential problems with implementing kNN on a very large dataset?**

                One must understand what operations happen during each iteration of the algorithm. For each new data point, the kNN classifier must:

1. Calculate the distances to all points in the training set and store them
2. Sort the calculated distances
3. Store the K nearest points
4. Calculate the proportions of each class
5. Assign the class with the highest proportion

Obviously this is a very taxing process, both in terms of time and space complexity. The first operation is a quadratic time process, and the sorting a O(nlogn) process. Together, one could say that the process is a O(n³logn) process; a monstrously long process indeed.

Another problem is memory, since all pairwise distances must be stored and sorted in memory on a machine. With very large datasets, local machines will usually crash.

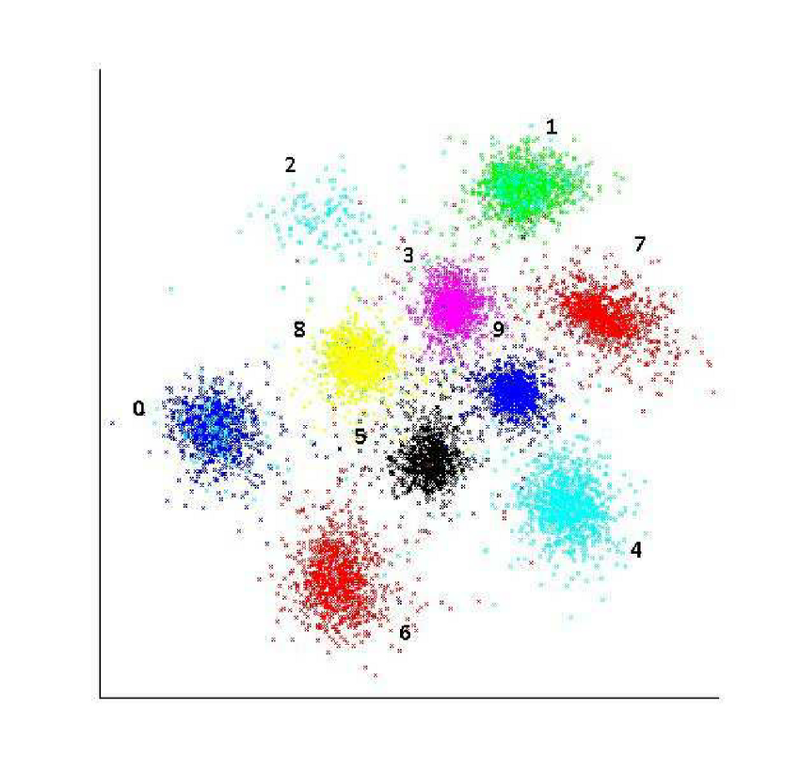
### **13.What are some ways of getting around the kNN-specific problems?**

Solution #1: Get more resources (computing power or larger memory).

This is obviously not the best answer to a scalability question, and not really applicable in real-life, industry problems.

Solution #2: Preprocessing the data.

Dimensionality reduction (via PCA (principal component analysis), or feature selection) to reduce the complexity of the distance calculation. You can also use clustering algorithms (like K-means or Rocchio) to reduce the number of points used to compute distances and sort, as illustrated below. In this case, the nontrivial task becomes assigning the test set point to the correct cluster.

Solution #3: Random sampling to reduce training set size.

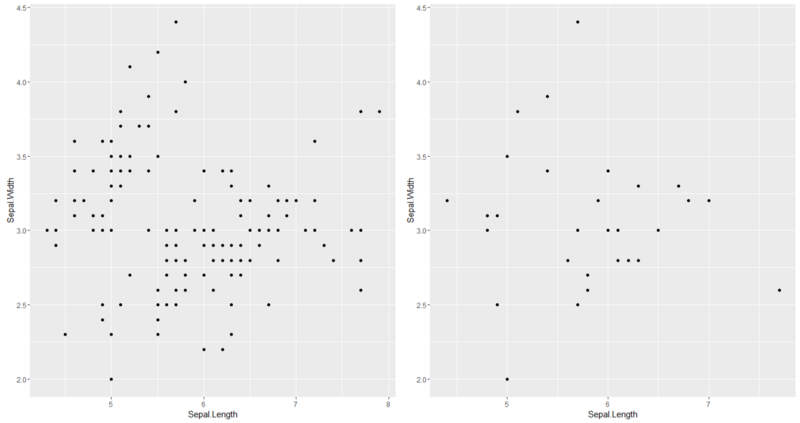
If you use a good random number generator and a decently large sample size, the sample should be a fairly good representation of the original.

### **14.If using random sampling only once and supposing we know a good k value to use for the original data, how should k be adjusted in accordance to a change in the input size?**

Two important points must be clarified to tackle this problem:

* What effect does sampling have on the kNN model?
* What effect does changing k have on the kNN model?

## Point #1: Effects of sampling:



ggplot2 comparison of a sample 20% the size of the original dataset(iris)

As illustrated above, sampling does several things in the perspective of a single data point, since kNN works on a point-by-point basis.

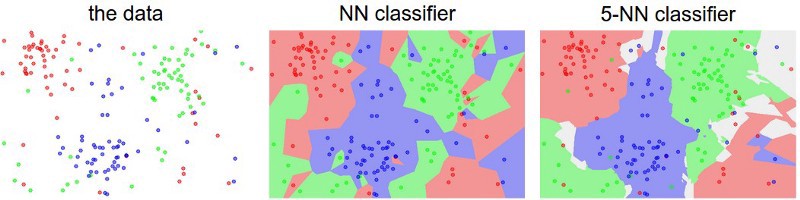
* The average distance to the k nearest neighbors increases due to increased sparsity in the dataset.
* Consequently, the area covered by k-nearest neighbors increases in size and covers a larger area of the feature space.
* The sample varianceincreases.

A consequence to this change in input is an **increase in variance**. When we talk of variance, we refer to the variability in the predictions given different samples from the population. Why would the immediate effects of sampling lead to increased variance of the model?

Notice that now a **larger area of the feature space** is represented by the same k data points. While our sample size has not grown, the population space that it represents has increased in size. This will result in higher variance in the proportion of classes in the k nearest data points, and consequently a higher variance in the classification of each data point.

## Point #2: Effects of Changing the k Parameter in kNN

Let us first examine the visual changes of changing k from k=1 to k=5 on a particular dataset.



Notice from the comparison that:

* The number of distinct regions (in terms of color) goes down when the k parameter increases.
* The class boundaries of the predictions become more smooth as k increases.

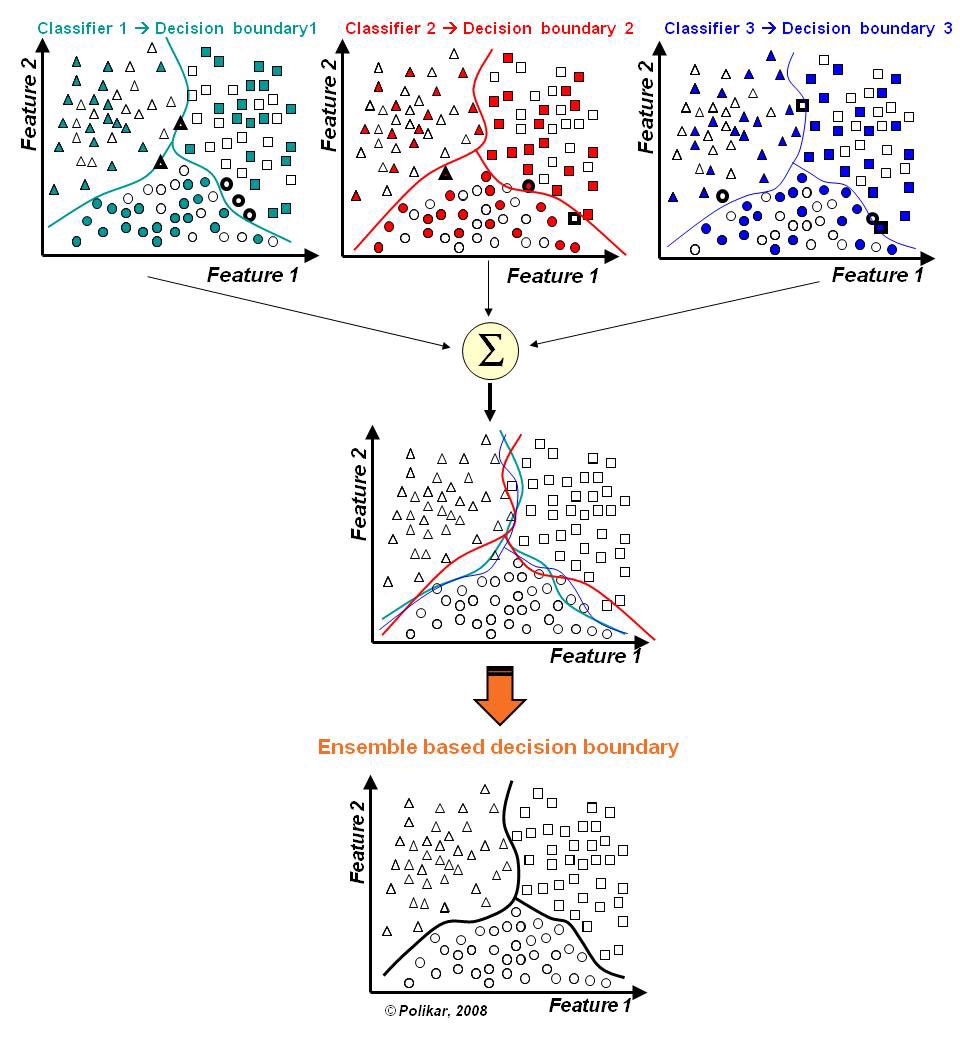
What really is the significance of these effects? First, it gives hints that a lower k value makes the kNN model more “sensitive.” That is, it is more sensitive to the local changes in the dataset. The “sensitivity” of the model directly translates to its variance.

All of these examples point to an **inverse relationship between variance and k.** Additionally, consider how kNN operates when k reaches its maximum value, k=n, where n is the number of points in the training set) In this case, the majority class in the training set will always dominate the predictions. It will simply pick the most abundant class in the data, and never deviate, effectively resulting in **zero variance.**Therefore it seems to reduce variance, k must be increased**.**

Final Verdict: In order to offset the increased variance due to sampling, k can be increased to decrease model variance.

### **15. If not restricted to a single sample, what could be a fairly simple method to reduce increased variance of the model other than changing k?**

                       If not restricted in the number of times, one can draw samples from the original dataset, a simple variance reduction method would be to sample,many times, and then simply take a majority vote of the kNN models fit to each of these samples to classify each test data point. This variance reduction method is called **bagging.** You might have heard of bagging, since it is the core concept in randomforest, a very popular tree ensemble method. We will explore this technique in greater detail in future posts.



### 17. What are the advantages and disadvantages of KNN algorithm?

**Advantages of KNN**

**1. No Training Period:** KNN is called **Lazy Learner (Instance based learning)**. It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc.

**2.** Since the KNN algorithm requires no training before making predictions, **new data can be added seamlessly** which will not impact the accuracy of the algorithm.

**3.** KNN is very **easy to implement**. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

**Disadvantages of KNN**

**1. Does not work well with large dataset:**In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.

**2. Does not work well with high dimensions:**The KNN algorithm doesn’t work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

**3. Need feature scaling:** We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we don’t do so, KNN may generate wrong predictions.

**4. Sensitive to noisy data, missing values and outliers**: KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers.