|  |  |
| --- | --- |
| K Nearest Neighbors | K Means |
| * is a classification (or regression) algorithm that in order to determine the classification of a point, combines the classification of the K nearest points. * It is supervised because you are trying to classify a point based on the known classification of other points.   KNN is a supervised learning algorithm which means training data is labeled. Consider the task of classifying a green circle between class 1 and class 2.  https://qphs.fs.quoracdn.net/main-qimg-32a7eff79169c81b19f548248f0962ae  If we choose k=1, then green circle will go into class 1 as it is closer to the class 1. If K=3, then there are ‘two’ class 2 objects and ‘one’ class one object. So KNN will classify the green circle in class 2 as it forms the majority.   * large K = siimple model = underfit = low variance & high bias * small K = complex model =overfit = high variance& low bias * **When K increases to inf, the model is simplest.** All test data point will belong to the same class: the majority class. This is **under-fit, that is, high bias and low variance.** * When K decreases, **lets say K=1, the granularity or resolution is too fine, which is overfit.**   **Overfit = high variance** | * is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other. * It is unsupervised because the points have no external classification. * How does K-means algorithm work?   In unsupervised learning, the data is not labeled so consider the unlabelled data. Our task is to group the data into two clusters.  https://qphs.fs.quoracdn.net/main-qimg-304e2b5fd617a9e0ce3937fcb9451781  This is our data, the first thing we can do is to randomly initialize two points, called the cluster centroids.  https://qphs.fs.quoracdn.net/main-qimg-0bccbb44c188253d7acbb0bca4d343a4  In k-means we do two things. First is a cluster assignment step and second is a move centroid step.  In the first step, algorithm goes to each of the data points and divide the points into respective classes, depending on whether it is closer to the red cluster centroid or green cluster centroid.  https://qphs.fs.quoracdn.net/main-qimg-0850a8a05ad1fb9334e05c769398d556  In the second step, we move the centroid step. We compute the mean of all the red points and move the red cluster centroid there. We do the same thing for green cluster.  This is an iterative step so we do the above step till the cluster centroid will not move any further and the colors of the point will not change any further.  https://qphs.fs.quoracdn.net/main-qimg-9a58150d109450e1aade2ca3633e70e3  This is very layman explanation of how k-means works. |