|  |  |  |
| --- | --- | --- |
|  | KMeans | SVM |
| What does Unsupervised/ Supervised mean? | Unsupervised: Data is not pre labelled | Pre-labelled data (by user) |
| Provide Main Idea + Sketch  & Pseudocode | Unlabelled Data  Preset number of clusters k  Centroid based: data points are assigned by distance measure  Euclidian distance  Ein Bild, das Farbigkeit enthält.  Automatisch generierte Beschreibung  **Pseudocode:**   1. Choose a cluster number $k$ and select randomly $k$ cluster centres (centroids) for initialisation. 2. assign each data point to the cluster whose cluster centre (centroid) is closest to this point. ‘Closeness’ is determined by the distance measure, which can be Euclid's distance 3. recalculate all $k$ cluster centres: The mean value (in all dimensions) of all data points in a cluster results in the new cluster centre (centroid).   cancellation criterion fulfilled? (e.g. ‘maximum number of iterations reached’ or ‘cluster centres change only slightly’) Otherwise go to 1. | * separate different groups of data in a multidimensional space. * Find the maximum margin between the closest data points of opposite groups, creating the optimal hyperplane * The number of features in the input data determines if the hyperplane is a line in a 2D space or a plane in an N-dimensional space. * can handle both linear and nonlinear classification tasks * **Support Vectors = the** data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier. * **Hyperplane =** decision plane which separates between a set of objects having different class memberships. * **Margin =** gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.   Ein Bild, das Screenshot, Text, Reihe, Farbigkeit enthält.  Automatisch generierte BeschreibungFig. by Wikipedia |
| error/accuracy | Test-test Splitting with prior labelling of the data! | Classification Error + Margin Error = Total Error  (Missclassific. + Points within the margin = Total error)  Train-test splitting |
| Feature Scaling | YES (Distance based) | YES (Distance based) |
| Advantages | Easy to implement, use for any number of dimensions, scales to large datasets, it will always converge, fits clusters in varying shapes and sizes, | * SVM works relatively well when there is a clear margin of separation between classes. * SVM is more effective in high dimensional spaces. * SVM is effective in cases where the number of dimensions is greater than the number of samples. * SVM is relatively memory efficient |
| Disadvantages | Number of clusters are chosen manually  Clusters are dependent on initial values, sensitive to outliers, k-means is limited to linear cluster boundaries | * Choosing a “good” kernel function is not easy. * Long training time for large datasets. * Memory-intensive: SVMs can be memory-intensive, as the algorithm requires storing the kernel matrix, which can be large for large datasets. * Difficult to understand and interpret the final model, variable weights and individual impact. * Since the final model is not so easy to see, we can not do small calibrations to the model |
| Application example (what should the data set look like?) | -species (e.g. plants) segmentation based on diff. parameters  - general finding patterns  e.g. email spam detection | What **Croptype** to plant for different climate/soil/hydrogeologic parameters  **Fire (Yes/No)** prediction based on diff. rem.sen. paramters  Email spam detection |