**Clustering**

Learning non supervisionato. Non vi è più il dataset di test.

**Distance based cluster**

Un algoritmo di cluster funziona valutando la distanza tra le istanze e mettendo quelle simili nello stesso e quelle diverse in altri.

Il problema diventa trovare una partizione che minimizzi la dispersione all’interno del cluster. Questo porterebbe alla soluzione banale: ogni cluster, una istanza.

* Usare un approccio euristico, per una soluzione buona, più che la migliore possibile: K means.
* Si potrebbe avere un problema di Soft Clustering, assegnano alle istanze un grado di appartenenza a più di un cluster: Expectation-Maximization.

**Tipologie di Clustering**

* Esclusivo: Ogni istanza a un cluster unico: K means
* Overlapping: Ogni punto a due o più cluster con un certo grado: Fuzzy C means
* Gerarchico o Probabilistico

Clustering è usato per trovare oggetti insoliti, gli outlier.

**K Means**

Apprende un clustering predittivo. Apprende un modello che può essere usato per assegnare nuovi dati ai cluster.

Vuole trovare una partizione che minimizzi la dispersione totale all’interno del cluster. L’algoritmo raggiungerà un punto stazionario dove non sono più possibili miglioramenti.

Itera la partizione dei dati usando la regola del centroide più vicino. Le istanze vengono assegnate al cluster con il centro più vicino.

La formula della funzione obiettivo: , con m media dei punti.

Scelta iniziale dei centri:

* Prime k istanze
* Etichettare le istanze con numeri progressivi e scegliere valori precisi: m/k, 2m/k
* Scelta random
* Generare k punti, scegliendo le coordinate.
* Generare una partizione del dataset in k sottoinsiemi esclusivi e considerare i centroidi.

Per sapere se i cluster sono buoni:

* SSE: Somma delle differenze2 tra ogni osservazione x e la media. Valori bassi, punti vicini al centroide, buon clustering.

**Fuzzy C means**

Permette a un punto di appartenere a più cluster. Usato nel Pattern Recognition. Tenta di suddividere una collezione finita di n esempi in c cluster fuzzy rispetto a un criterio. Return: un elenco di cluster, una matrice di partizione U, con ogni elemento che rappresenta il grado di appartenenza al cluster. Uij è il grado di xi al cluster j, compreso tra [0,1]

Funzione obiettivo:

Procedimento: Si inizializza la matrice. A ogni k step calcolo il vettore dei centri. Aggiorno Uk, Uk+1. Se la differenza in modulo è minore di una soglia, stop, altrimenti ritorno al passo 2.

Procedura che converge a un minimo locale. Diverse inizializzazioni causano diverse evoluzioni dell’algoritmo.

La matrice di partizione: contiene tutti i fattori presi dalle funzioni di appartenenza. La somma di ogni riga = 1. Colonne = #Cluster, Righe = #Dati.

**Clustering gerarchico**

Per identificare i gruppi nel dataset e produce una rappresentazione ad albero. Può essere:

* **Bottom up**: Ogni oggetto è come un cluster di un elemento (foglie), alla fine tutti i punti sono membri dello stesso (root).
* **Top down**: Inverso. Parte dalla root, con tutti in un unico cluster. Iterando finché ognuno non sono nel loro cluster.

Bottom up: Si assegna ogni elemento a un cluster. N elementi, N cluster. Si trova la coppia più vicina e si unisce in uno unico. Si ricalcolano le distanze. E si reitera fino ad arrivare a un unico cluster. Per ottenerne k, si tagliano i k-1 collegamenti più lunghi. La funzione per determinare la distanza può essere:

* Single linkage: Distanza più breve tra qls membro di uno e qls membro dell’altro.
* Complete linkage: Distanza più grande tra qls membro di uno e dell’altro.
* Average linkage: Distanza media tra qls membro di uno e dell’altro.

Il grafico risultante è il dendrogramma. Altezza di ogni nodo è proporzionale alla dissimilarità tra i suoi figli.

**Clustering probabilistico**

Trovare l’insieme di cluster più probabile dati i dati.

Mixture model: modello probabilistico dove i dati sono generati da un mixture finito di distribuzioni con parametri sconosciuti.

Ogni cluster è rappresentato da una distribuzione parametrica. Ogni distribuzione dà la probabilità che un’istanza abbia valori per i suoi attributi, se si sapesse a quale cluster appartiene.

Quello più usato è il Gaussian mixture model. Un modello di GMM è una funzione composta da diverse gaussiane, ognuna identificata dal numero di cluster del dataset. Distribuzioni centrate nei baricentri 𝜇, e varianza 𝜎.