Apprendimento basato sulle istanze

L’apprendimento è la registrazione degli esempi di training nel database. Non costruisco un modello, memorizzo i dati. La classificazione di una nuova istanza, è il reperimento degli esempi simili e classificazione di quella sulla base di questi esempi. Assegno alla nuova istanza, la classe di maggioranza degli esempi più vicini.

* Metodo lazy: Ritarda l’elaborazione finchè non è necessario classificare una nuova istanza.
* Costo di classificazione può essere elevato, devono essere considerati tutti gli esempi.

KNN - K Nearest Neighbour

Classificatore di apprendimento supervisionato non parametrico, che utilizza la prossimità per effettuare classificazioni o previsioni sul raggruppamento di un singolo punto dati.

Usa la distanza euclidea.

Trova i #k elementi più vicini secondo la distanza euclidea e restituisce il valore di classe più frequente.

* Tutti gli attributi delle istanze sono considerati per reperire gli esempi simili. Se il concetto dipende da pochi attributi, gli esempi possono essere a grande distanza.

Altre tipologie di distanze:

* Minkowski, con ordine p>0: distanza = || xi -xj ||, (norma)
* Distanza euclidea, con p = 2,
* Distanza di Manhattan, con p = 1
* Distanza di Chebyshev, con p = ∞.

Gli attributi possono avere scale diverse, la distanza può essere diversa. Prima della distanza si cambia la scala, in modo che ognuno varii tra [0,1].

Se gli attributi sono nominali: Si possono usare distanze definite ad hoc.

KNN pesato sulla distanza

Il contributo di ciascun k esempio vicino è pesato sulla distanza tra lui e l’istanza da classificare. Gli esempi più vicini sono pesati di più rispetto quelli lontani. Peso w = 1/d2.

Se si considerano tutti gli esempi invece dei k più vicini. Lo svantaggio è computazionale. Si parlerebbe di metodo globale invece che locale.

Curse of dimensionality

k-NN considera tutti gli attributi degli esempi, in contrasto con gli algoritmi per l’apprendimento di alberi di decisione che selezionano solo gli attributi più rilevanti. Esempio: 20 attributi di cui solo 2 rilevanti per la classificazione delle istanze. Istanze che hanno valori identici per questi 2 attributi potrebbero essere distanti nello spazio delle istanze: 20-NN.

**Pesare ogni attributo diversamente** quando si calcola la distanza tra 2 istanze corrisponde al variare la lunghezza degli assi nello spazio euclideo di un valore costante pari a wj: accorciando gli assi che corrispondono ad attributi meno rilevanti e allungando gli assi che

corrispondono ad attributi più rilevanti

Definizione di K

Valori bassi di k: overfitting (controllo pochi esempi)

Valori alti di k: underfitting (mi allargo e sono più generico)

Si consiglia sempre un valore dispari (3-NN, 5-NN) per non avere lo stesso #esempi positivi e negativi