Notatki do wykładów: algorytm grupowania k-średnich (k-Means)

(c) Marcin Sydow

Zadanie grupowania (ang. clustering)

Grupowanie jest przykładem zadania uczenia maszynowego bez nadzoru.

Dany jest zbiór danych składających się z N obserwacji.

Każda obserwacja opisana jest wektorem atrybutów. Ponieważ jest to uczenie bez nadzoru, nie ma atrybutu decyzyjnego.

Celem grupowania jest podział zbioru obserwacji na kilka grup w taki sposób, żeby obserwacje w każdej grupie były jak najbardziej "podobne" parami do siebie, natomiast obserwacje przydzielone do różnych grup były jak najbardziej "niepodobne" do siebie.

W wielu algorytmach liczba grup k, na jakie należy podzielić dane jest dana jako argument.

Sformalizowanie pojęcia "podobieństwa"

Aby sformalizować pojęcie "podobieństwa" (i "niepodobieństwa") występujące w opisie problemu grupowania, używa się pojęcia odległości pomiędzy parami obiektów.

Jeżeli wszytkie atrybuty są numeryczne, do mierzenia stopnia "podobieństwa" używa się np. odległości euklidesowej (podobnie jak w przypadku algorytmu klasyfikacji k-NN). 2 obiekty są tym bardziej podobne im mniejsza jest odległość między nimi.

Sformalizowanie problemu grupowania dla k grup

Podział zbioru danych X na k-grup tak, żeby wewnątrz grup obiekty były jak najbardziej do siebie parami podobne a pomiędzy grupami jak najmniej podobne można sformalizować następująco.

Rozważmy sumę wszystkich odległości parami obiektów w zbiorze danych.

Suma ta, dla ustalonego zbioru oczywiście jest stała, nazwijmy tę sumę S:

$$S = \sum_{x,y \in X} d(x,y)$$

Sformalizowanie problemu grupowania dla k grup

Dla ustalonego podziału na k grup można podzielić zbiór wszystkich par na dwa rozłączne podzbiory par:

- I: zbiór takich par, że oba elementy należą do tej samej grupy podziału
- O: zbiór takich par, że elementy należą do dwóch różnych grup podziału

Wtedy wielkość S zdefiniowana poprzednio rozbija się na dwa składniki:

$$S = \sum_{x,y \in X} d(x,y) = \sum_{x,y:(x,y) \in I} d(x,y) + \sum_{x,y:(x,y) \in O} d(x,y)$$

Wobec stałości wielkości S i ustalonej liczby grup k, minimalizacja pierwszego składnika sumy (podobieństwo wewnątrz grup) jest więc równoznaczna z maksymalizacją drugiego (niepodobieństwo pomiędzy grupami).

Problem grupowania na k grup jest więc dobrze zdefiniowanym problemem optymalizacyjnym

Algorytm grupowania k-średnich (ang. k-means)

Wejście: n elementowy zbiór danych do pogrupowania (dane opisane są za pomocą n-wymiarowych wektorów atrybutów numerycznych), liczba naturalna k (liczba grup)
Wyjście: przyporządkowanie każdego z n elementów danych do dokładnie jednej z k grup, tak, aby każda z k grup była niepusta

Inicjalizacja: każdy z n elementów zostaje początkowo przyporządkowany do losowo wybranej z k grup

Następnie, aż do ustania jakichkolwiek zmian wykonywane są na przemian dwie fazy:

- oblicz centroid każdej grupy (średnia arytmetyczna wszystkich wektorów w danej grupie)
- przyporządkuj każdy element do tej grupy, której centroid jest najbliższy danemu elementowi

Własności algorytmu k-średnich

- po skończonej liczbie kroków algorytm k-średnich zatrzyma się (nie będzie dalej żadnych zmian w obliczanych centroidach ani przyporządkowaniach elementów do grup)
- algorytm k-średnich minimalizuje sumę kwadratów odległości elementów od centroidów swoich grup.
- algorytm k-średnich nie daje gwarancji znalezienia podziału na grupy dającego minimalną sumę kwadratów dla danego zbioru danych i liczby grup k.

Dziękuję za uwagę