Klasyfikator oparty o drzewo decyzyjne

Łukasz Odwrot 218283

13.03.2018

Spis treści

1 Wstęp **2**

2 Badane zbiory **2**

3 Metody klasteryzacji **2**

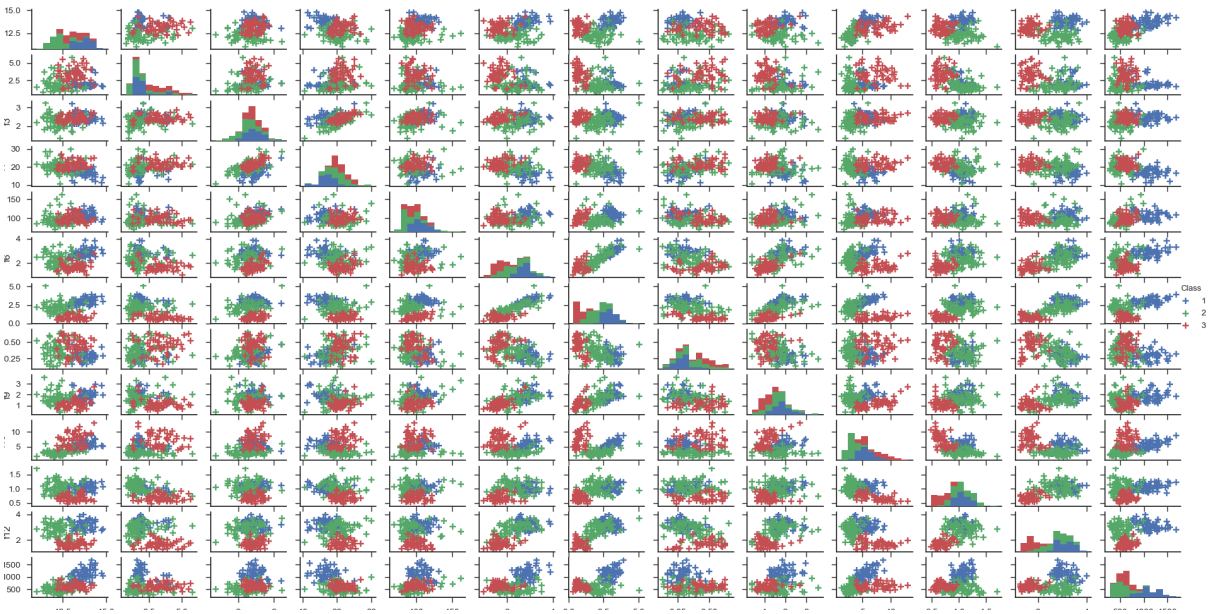
4 Ocena klasteryzacji **3**

1 Wstęp

Klasteryzacja to forma nienadzorowanego uczenia. Polega ona na przypisaniu obiektów ze zbioru na podstawie podobieństwa cech do klastrów, których ilość zwykle jest parametrem wejściowym algorytmu klasteryzacji.

2 Badane zbiory

Rozkłady cech dla poszczególnych klas przedstawiono na poniższych rysunkach.



Rysunek 1: Rozkład cech dla zbioru Wine

3 Metody klasteryzacji

Zbadane zostaną dwie metody klasteryzacji.

K-means

Metoda ta polega na przyporządkowaniu danych wejściowych, w których każda próbka należy do klastra z najbliższą wartością średnią. Każda próbka to wielowymiarowy wektor liczb rzeczywistych.

Na wejściu algorytmu podawane są parametry :

k - ilość klastrów

data - dane wejściowe

Algorytm działa następująco:

1. Rozmieszcza centroidy w losowych miejscach przestrzeni,
2. Dla każdego punktu znajduje najbliższą centroidę i przypisuje punkt do centroidy,
3. Na podstawie wszystkich próbek przypisanych do centroidy wyliczane są wartości średnie i powtarzany jest krok 2 dopóki warunki stopu nie zostaną spełnione lub nie nie zmieniło się przypisanie próbek.

Metoda przeznaczona jest jedynie dla danych numerycznych.

W funkcji klasteryzacji kmeans w R możemy zdefiniować maksymalną liczbę iteracji (domyślnie 10) oraz ilość losowań początkowych pozycji, z których wybrana zostanie najlepsza.

Partitioning Around Medoids
Używa ona zachłannego algorytmu, więc może nie znaleźć najlepszego rozwiązania, ale dzięki temu jest relatywnie szybka. Działa według następującego schematu.
1. Wybiera k reprezentatów, które będą centami klastrów. 2. Przypisuje wszystkie punkty do najbliższego klastra. 3. Dla każdej próbki będącą centrum medoidy i dla każdej próbki nie będącej centrum zamień je. Jeżeli konfiguracja pogorszyła się, cofnij zmianę.
Koszt wyliczany jest jako suma odległości próbek w klastrze od jego centrum.

4 Ocena klasteryzacji

Do oceny jakości klasteryzacji posłużą nam następujące miary:
Purity
Informuje w jakim stopniu klastry odpowiadają pojedynczym klasom. Warto zaznaczyć, że w przypadku takiej samej ilości klas co klastrów funkcja zawsze zwróci wartość 1.

$$\frac{1}{N} = \sum_{m \in M} \max_{d \in D} |m \cap d|$$

Rand measure Porównuje jak podobne są klastry względem wzorca. Miara może być interpretowana jako procentowa ilość podjętych prawidłowych decyzji.

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Dunn index Miara ta odzwierciedla gęstość i poprawność odseparowania klastrów. Wyliczana jest na podstawie stosunku minimalnej odległości wewnątrz klastra do maksymalnej odległości wewnątrz klastra.

Davies–Bouldin index Miarę tę można obliczyć na podstawie poniższego wzoru.

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Gdzie n jest liczbą klastrów, c - centroidą klastra, σ - średnia odległość od wszystkich elementów klastra. Algorytm dające niskie odległości wewnątrz klastra będą i duże odległości między klastrami będą dawały niskie wyniki.

$$\hat{y} = arg \max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$P(y \mid x_1, ..., x_n) = \frac{P(y) P(x_1, ...x_n \mid y)}{P(x_1, ..., x_n)}$$