Datorlaboration 5

Josef Wilzén

15 september 2020

Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- Kodmanual: länk
- Dataset till vissa uppgifter finns här.
- **ISL**: An Introduction to Statistical Learning,
  - Allmän info: länk
  - Boken: länk
  - R-kod till labbar: länk
  - Dataset: länk

Notera att ni inte behöver göra alla delar på alla uppgifter. Det viktiga är att ni får en förståelse för de olika principerna och modellerna som avhandlats. Dessa uppgifter ska inte lämnas in, utan är till för er övning.

Datauppdelning

För att motverka överanpassning bör ni dela upp data till träning-, validering-, (och testmängd). Detta kan göras med `createDataPartition()` från `caret`-paketet. Argument till den funktionen som är av vikt här är p som hur stor andel av observationerna som ska användas till träningsmängden. Ni kan också använda `subset()` för att göra detta också, men det blir svårare att tydligt ange de observationer som ska tilldelas till valideringsmängden. Denna uppdelning ska ske slumpmässigt. Notera att om en testmängd ska skapas måste uppdelningen ske en gång till från valideringsmängden.

Del 1: K-means klustering

1. Gör lab 10.5.1 i **ISL**.
2. **ISL**: 10.7 Exercises Conceptual: 3

Del 2: Startcentroider i K-means klustering

Använd ”data1_alt.csv” för att beräkna en K-means klusteranalys.

1. Visualisera datamaterialet för att se med blotta ögat kluster i materialet. Hur ser de naturliga klustrena ut och vars är deras centroider?
2. Genomför en k-means klustring med de angivna startcentroiderna i koden samt ange antalet maximala iterationer till 1. Visualisera de resulterande klustren och tolka resultatet.

```
> cluster_result <- kmeans(x = data1,
+                  centers = rbind(c(0,0), c(1, 0), c(-1, 0), c(1, 1)),
+                  iter.max = 1, nstart = 1)
```

3. Öka antalet iterationer till något rimligt värde som tillåter algoritmen att konvergera och kör algoritmen igen och repetera steg 2. Blev resultatet något bättre? Varför inte?
4. Ange nu k slumpmässiga centroider istället. Tillåt algoritmen konvergera och jämför resultatet med steg 2. och 3.
5. Återkoppla till vilken funktion som K-means avser att optimera och bedöm vilken av diagrammen som visar de “korrekta” klustren.

Del 3: Densitetsbaserad klustring

Använd ”data3.csv” för att beräkna densitetsbaserad klusteranalys.

1. Visualisera materialet i två dimensioner och definiera antalet naturliga kluster.
2. Genomför klustring med DBSCAN och följande värden: `eps = 0.18`, `minPts = 3` och `borderPoints = TRUE`. Vad innebär dessa inställningar? Visualisera de resulterande klustren och bedöm ifall de faller in med de naturliga klustren som identifierats tidigare.
3. Testa ändra på `eps` argumentet till mindre och större värden. Hitta det minsta värdet då alla observationer tilldelas till ett kluster. Beskriv hur förändringen av detta argument påverkar klusterresultatet, d.v.s. vad händer i algoritmen i och med förändringen.
4. Testa att ändra på `minPts`- och `borderPoints`-argumenten. Hur förändras de resulterande klustren?
5. Genomför k-means klustring med några olika värden på k. Visualisera de resulterande klustren och jämför med steg 2. Vilken av metoderna anses producera “bättre” resultat? Om resultaten skiljer sig åt, försök förklara varför.

Använd nu det inbyggda datasetet iris. Notera att här finns det tre klasser (Species). Använd inte den variabeln i klusteringen, utan endast de fyra numeriska variablerna.

1. Genomför klustring med DBSCAN. Välj valfria värden på hyperparametrarna. Hur många kluster erhåller ni? Undersök hur väl klustringen matchar de befintliga grupperna (klasserna).
   (a) Gör parvisa punktdiagram för variablerna (blir sex stycken), låt klustren ha olika färger. Låt klasserna representeras av olika symboler. Hur blir uppdelningen?
   (b) Om matchningen blir dålig: försök att ändra på hyperparametrarna för att få en bättre gruppering.
2. Genomför klustring med k-means. Testa k=2,3,4,5. Undersök hur väl klustringen matchar de befintliga grupperna (klasserna).
   (a) Gör parvisa punktdiagram för variablerna (blir sex stycken), låt klustren ha olika färger. Låt klasserna representeras av olika symboler. Hur blir uppdelningen?
3. Jämför resultatet i 1) och 2). Vilken modell föredrar ni här?

Del 4: Teorifrågor

1. Ge ett exempel på ett datamaterial som innehåller kluster av olika tätheter där DBSCAN inte kommer kunna hitta de “korrekta” naturliga klustren?
2. Ge ett exempel på ett datamaterial där k-means kommer att fungera dåligt.
3. Förklara varför man i K-means algoritmen beräknar centroidernas position genom ett medelvärde av alla klustrets punkter om vi använder det Euklidiska avståndsmåttet och vill minimera SSE.
4. Beskriv två alternativa lösningar till initialiseringsproblemen för val av startcentroider i K-means algoritmen utöver att slumpmässigt välja dessa.
5. Vilka olika hyperparameterar finns det i k-means?
6. Vilka olika hyperparameterar finns det

Del 5

Gå in på denna sida Clustering basic benchmark. Ni ska nu testa att klustra några dataset härifrån.

1. Ladda ner datasetet ”Unbalance” och läs in i R.
   (a) Plotta och undersök data

(b) Testa att göra klustra data med
   i. k-means: testa olika värden på k
   ii. DBSCAN: Välj valfria värden på hyperparametrarna.

(c) Hur fungerar klustringen i de olika fallen?

2. Ladda ner dataseteten ”S-sets” och läs in i R.

(a) Börja med att undersöka data: Vad skiljer S1-S4 åt?

(b) Gå igenom dataseten S1-S4 och klustra med
   i. k-means: testa olika värden på k
   ii. DBSCAN: Välj valfria värden på hyperparametrarna.

(c) Hur fungerar klustringen i de olika fallen?

Del 6: Frivillig fördjupning

Nedan följer några uppgifter för frivillig fördjupning på områden som tas upp i kursen multivariata metoder. Notera att det är tillåtet att använda dessa metoder i era projekt i denna kurs.

Hierarkisk klustring

1. Gör lab 10.5.1 i **ISL**.
2. **ISL**: 10.7 Exercises Applied: 9, 11

Länkningsmetoder i hierarkisk klustring

Använd data2.csv för att beräkna hierarkisk klusteranalys.

1. Visualisera materialet i två dimensioner och definiera antalet naturliga kluster.
2. Beräkna en hierarkisk klustring med följande länkningsmetoder och visa lämpliga dendrogram:
   (a) Enkel länkning
   (b) Fullständig länkning
   (c) Ward’s metod
3. Definiera från varje dendrogram hur många kluster som hittats och visualisera resultatet i ett diagram. Verkar metoden ha hittat de naturliga klustren? Vilken länkningsmetod anser ni vara lämpligast för denna sorts data?
4. Genomför K-means klustring på samma data och jämför med resultaten med enkel länkning. Vilken egenskap utav de två metoderna visas här?

Principalkomponent analys (PCA)

1. Kör lab 10.4 i **ISL**.
2. **ISL**: 10.7 Exercises Applied: 10