Übung 04

Algorithmik und Statistik 2 LAB, SS2019

Bis: Sonntag, 23. Juni 2019, 23:59 Uhr

Bitte um Beachtung der Übungs-Policy für genaue Anweisungen und einige Beurteilungsnotizen. Fehler bei der Einhaltung ergeben Punktabzug.

Aufgabe 1

[**10 points**] Für diese Frage verwenden wir die `OJ`Daten aus dem `ISLR`-Paket. Wir werden versuchen, die Variable "Purchase" vorherzusagen. Nachdem Sie `uin` zu Ihrem `UIN` geändert haben, verwenden Sie den folgenden Code, um die Daten aufzuteilen.

```
library(ISLR)
library(caret)
uin = 123456789
set.seed(uin)
oj_idx = createDataPartition(OJ$Purchase, p = 0.5, list = FALSE)
oj_trn = OJ[oj_idx,]
oj_tst = OJ[-oj_idx,]
```

(a) Stimmen Sie ein SVM mit linearem Kernel mit 5-facher Cross-Validierung auf die Trainingsdaten ab. Verwenden Sie das folgende Wertgitter für `C`. Berichten Sie die gewählten Werte aller Tuningparameter + Testgenauigkeit.

```
lin_grid = expand.grid(C = c(2 ^ (-5:5)))
```

(b) Abstimmung eines SVM mit Polynomkern auf die Trainingsdaten mittels 5-facher Cross-Validierung. Geben Sie kein Tuning-Grid an. (`caret` wird einen für Sie erstellen.) Berichten Sie die gewählten Werte aller Tuning-Parameter. Berichten Sie über die Genauigkeit der Testdaten.

(c) Stimmen Sie ein SVM mit Radialkernel mit 5-facher Cross-Validierung auf die Trainingsdaten ab. Verwenden Sie das folgende Wertgitter für `C` und `sigma`. Berichten Sie die gewählten Werte aller Tuningparameter. Berichten Sie über die Genauigkeit der Testdaten.

```
rad_grid = expand.grid(C = c(2 ^ (-2:3)), sigma  = c(2 ^ (-3:1)))
```

(d) Stimmen Sie einen Random Forest mit einer 5-fachen Kreuzvalidierung ab. Berichten Sie die gewählten Werte aller Tuningparameter. Berichten Sie über die Genauigkeit der Testdaten.

(e) Fassen Sie die obigen Genauigkeiten zusammen. Welche Methode hat am besten funktioniert? Warum?

Aufgabe 2

[**10 points**] Verwenden Sie für diese Frage die Daten in `clust_data.csv`. Wir werden versuchen, diese Daten mit k-means zu bündeln. Aber, welche k sollen wir verwenden?

(a) Wenden Sie k-means 15 mal auf diese Daten an, wobei Sie die Anzahl der Zentren von 1 bis 15 verwenden. Verwenden Sie jedes Mal `nstart = 10` und speichern Sie den Wert `tot.withinss` aus dem resultierenden

Objekt. (Hinweis: Schreiben Sie eine for-Schleife.) Die `tot.withinss` misst, wie variabel die Beobachtungen innerhalb eines Clusters sind, das wir gerne niedrig halten würden. Offensichtlich wird dieser Wert also mit mehr Zentren niedriger sein, egal wie viele Cluster es wirklich gibt. Zeichne diesen Wert gegen die Anzahl der Zentren auf. Suchen Sie nach einem "Ellenbogen", der Anzahl der Zentren, in denen die Verbesserung plötzlich wegfällt. Basierend auf dieser Darstellung, wie viele Cluster sollten Ihrer Meinung nach für diese Daten verwendet werden?

(b) Wenden Sie k-means für die von Ihnen gewählte Anzahl von Zentren erneut an. Wie viele Beobachtungen werden in jedem Cluster platziert? Was ist der Wert von `tot.withinss`?

(c) Visualisieren Sie diese Daten. Plotten Sie die Daten mit den ersten beiden Variablen und färben Sie die Punkte entsprechend des k-means clusterings. Basierend auf diesem Plot, denken Sie, dass Sie eine gute Wahl für die Anzahl der Zentren getroffen haben? (Kurze Erklärung.)

(d) Verwenden Sie PCA, um diese Daten zu visualisieren. Plotten Sie die Daten mit den ersten beiden Hauptkomponenten und färben Sie die Punkte entsprechend dem k-means Clustering. Basierend auf diesem Plot, denken Sie, dass Sie eine gute Wahl für die Anzahl der Zentren getroffen haben? (Kurze Erklärung.)

(e) Berechnen Sie den Anteil der Variation, der durch die Hauptkomponenten erklärt wird. Machen Sie eine Darstellung des kumulierten Anteils erklärt. Wie viele Hauptkomponenten sind notwendig, um 95% der Variation der Daten zu erklären?

Aufgabe 3

[10 points] Für diese Frage werden wir auf die `USArrests` Daten aus den Notizen zurückkommen. (Dies ist ein Standarddatensatz von `R`.)

(a) Führen Sie hierarchisches Clustering sechsmal durch. Berücksichtigen Sie alle möglichen Kombinationen von Verknüpfungen (Average, Single, Complete) und Datenskalierung. (Skaliert, Nicht skaliert.)

Linkage	Scaling
Single	No
Average	No
Complete	No
Single	Yes
Average	Yes
Complete	Yes

Schneiden Sie das Dendrogramm jedes Mal auf eine Höhe, die zu vier verschiedenen Clustern führt. Plotten Sie die Ergebnisse mit einer Farbe für jeden Cluster.

(b) Basierend auf den obigen Plots, erscheint eines der Ergebnisse nützlicher als die anderen? (Es gibt hier keine richtige Antwort.) Wählen Sie Ihren Favoriten. (Nochmals, keine richtige Antwort.)

(c) Verwenden Sie die Dokumentation zu `?hclust`, um weitere mögliche Verknüpfungen zu finden. Such dir einen aus und probiere ihn aus. Vergleichen Sie die Ergebnisse mit Ihren Favoriten von **(b)**. Ist es anders?

(d) Verwenden Sie die Dokumentation zu `?dist`, um andere mögliche Entfernungsmessungen zu finden. (Wir haben `euklidisch` verwendet.) Wählen Sie eine (nicht `binär`) und versuchen Sie es. Vergleichen Sie die Ergebnisse mit Ihren Favoriten von **(b)**. Ist es anders?