Outras familias de clusterização

Jacques Wainer

IC – Unicamp

Novembro 2015

Clusterização hierárquica

- aglutinativa: variações na definição de *linkage*
  - single
  - average
  - complete
  - Wald
- divisiva

Clusterização hierárquica

- ir agrupando as 2 "coisas" mais próximas entre si.
- "coisas" podem ser dados e e clusters
- mas é preciso definir qual a distancia de um cluster a um dado e de dois clusters - essa é a função de *linkage*

Linkage

- single: a distancia de 2 clusters é a menor distancia entre pontos de um para os pontos do outro
- complete: a distancia entre 2 clusters é a maior distancia entre os pontos de um para os pontos do outro
- average: a distancia é a média da distancia entre os pontos de um e do outro
- Ward

Dendograma

- Uma representação grafica dos pontos (e clusters) sendo agregados em clusteres maiores, até a ultima etapa onde cria-se o cluster final com todos os dados
- representado como uma arvore com a raiz no topo
- a barra horizontal indica que 2 clusters (ou pontos) formaram um cluster, e a altura da barra indica a distancia entre os dois subclusters originais
- ▸ link
- outros formatos: horizontal ▸ link polar ▸ link

Clusterização hierárquica - segunda etapa

- Normalmente o dendograma por si só não nos da informação útil.
- usa-se o dendograma para definir onde "vai ser cortado", que define o numero de clusters ▸ link

Clusterização hierárquica em R

- Função `hclust` link faz o dendograma. Note que a função recebe uma matrix de distancia e não os dados. `method=` define o linkage
- `plot` plota o dendograma (2 versões, `hang=-1`)
- `cutree` corta o dendograma em k clusters e retorna o clustering dos dados
- `rect.hclust` link desenha no dendograma retângulos nos clusteres finais (dado o k)

Clusterização hierárquica em R

```
iris2d=prcomp(iris[,-5])$x[,1:2]
ii=iris2d[1:20,]
dd=dist(ii)
h.c=hclust(dd,"complete")
plot(h.c,main="Complete")
plot(h.c,main="Complete" ,hang=-1)
```

Clusterização hierárquica em R

```
h.s=hclust(dd,"single")
plot(h.s,main="Single")
h.a=hclust(dd,"ave")
plot(h.a,main="Average")
h.w=hclust(dd,"ward")
plot(h.w,main="Ward")
```

Clusterização hierárquica em R

```
plot(iris2d)
dd=dist(iris2d)
c.s=cutree(hclust(dd,"single"),k=5)
plot(xx,col=c.s,mail="Single")
c.c=cutree(hclust(dd,"complete"),k=5)
plot(xx,col=c.c,main="Complete")
c.a=cutree(hclust(dd,"ave"),k=5)
plot(xx,col=c.a,main="Average")
c.w=cutree(hclust(dd,"ward.D"),k=5)
plot(xx,col=c.w,main="ward")
```

Clusterização hierárquica divisiva

- Use um 2-means para dividir o conjunto todo em 2 grupos.
- pego o cluster com maior diametro (= maior distancia entre 2 membros do cluster) e divida-o usando o 2-means
- repita o processo ate que so reste um dado em cada cluster.
- parece ser usado em processamento de texto.
- implementado pela função `diana` link em R. (A função pode usar o k-medoides em vez do k-means)

Clusterização baseada em densidade: DBScan

- Cluster são regiões de alta densidade de dados separados por regiões de baixa densidade
- um ponto é **denso** se num raio `Eps` existem pelo menos `MinPts` dados. Os pontos dentro do raio `Eps` são chamados de vizinhança
- um ponto é **fronteiriço** se ele pertence a vizinhança de um ponto denso mas ele proprio não é denso.
- um ponto é um **outlier** caso contrario.
- ▸ link

DBScan

- encontre um ponto denso sem cor: ele iniciar um cluster novo
- pontos densos propagam seu cluster para os vizinhos,
- pontos fronteiriços recebem u,a indicação de cluster mas nao propagam
- outliers nao recebem indicação de cluster
- ▸ link
- DBscan define o numero de clusters automaticamente, nao é completo (nem todos os pontos pertencem a um cluster),
- mas tem 2 parametros que precisam de ajuste (em vez de um para o k-means) `Eps` e `MinPts`

Como escolher Eps e MinPts?

- determinar para cada ponto a distancia do k-ezimo vizinho mais próximo.
- varios lugares recomendam k=4 ou k=5 - esse será o `MinPts`.
- plote a distancia ordenada do k-esimo vizinho mais proximo de cada dado. Deve haver um cotovelo a partir do qual a distancia do k-esimo vizinho aumenta - essa distancia sera o `Eps`
- exemplo ▸ link
- a intuição é que a maioria dos pontos são densos e portanto o MinPts-ezimo vizinho deve ser menor que o Eps (já que ele é denso)

DBScan em R

- Função `dbscan` do pacote `dbscan` ▸ link
- ha também oa função `dbscan` do pacote `fpc` (usando na aula anterior).
- o pacote `dbscan` tem a função `kNNdist(x, k, ...)` e `kNNdistplot(x, k = 4, ...)` para plotar as distancias do k-ezimo vizinho.

Ensemble de clusterizadores

- Ensembles são muito usados em modelos preditivos, geralmente com muito sucesso
- Sao um conjunto de classificadores da mesma familia (mas treinados com dados levemente diferentes) ou de familias diversas (usualmente treinados nos mesmos dados)
- cada membro do ensemble faz uma previsão, e essas previsões são agregadas (maioria, maioria ponderada, um algoritmo de aprendizado para combinar as previsões, etc)
- a mesma ideia para clusterizadores.

Ensemble de clusterizadores

- Varios clusterizadores (ou da mesma familia ou de familias diferentes).
- clusteres mais robustos e estáveis (menos sensíveis a outliers, e a parametros dos clusterizadores)
- clusteres mais sutis, novidade (?)

Ensemble de clusterizadores

Por exemplo vários k-means

- com k diferentes
- com subconjuntos diferentes dos dados
- com inicializações diferentes
- com todos os dados mas subconjuntos das dimensões originais

Ensemble de clusterizadores

- Ate agora nos temos variado a inicialização (`nstart`) e o k para *escolher* o melhor k
- ensembles não escolhem o k, eles usam os varios resultados para construir uma nova representação dos dados e clusteriza-los nessa nova representação.
- por exemplo, para cada par de dados podemos medir a proporção de vezes que eles estão num mesmo cluster - uma nova representação relacional entre os dados
- ou criar um grafo bipartido de dados e clusters (dos varios clusterizadores) e particionar este grafo
- capitulo 13 do livro texto é sobre ensembles de clusterizadores

Tarefa

- Use as 4 primeiras dimensões do iris padronizadas
- faça uma clusterização hierárquica com linkage single, complete e ward.
- Use a dimensão 5 do iris e meça o corrected Rand index para determinar o numero de clusters para cada linkage
- qual o linkage da o melhor resultado segundo o corrected Rand
- use o DBScan e as técnicas discutidas para determinar o Eps. Compute o corrected Rand para os dados *nao outliers*.
- compare os resultados da clusterização hierárquica, do DBScan e do k-means da lista passada.