K-means parte 2

Jacques Wainer

IC – Unicamp

Novembro 2015

Como escolher o k

- medidas internas - usando apenas os dados originais
- medidas externas - usando informação extra sobre os dados, em particular a classe que eles pertencem.
- medidas externas fazem sentido quando se tem a classe de *alguns* dados e usa-se isso para ajustar os grupos. Se voce tem a classe de todos os dados, então use um classificador.

Medidas internas

- **coesão** dentro de um grupo (intra-grupo): os pontos dentro de um mesmo grupo devem estar próximos entre si.
- **separação** dos diferentes grupos (inter-grupo) os grupos devem estar distantes entre si.
- varias definições e intuições sobre o que são distancia intra cluster e distantcia inter-clusters, e como agrega-las

(Familia de) indice de Dunn

$$DI = \frac{\min \text{inter}_{ij}}{\max \text{intra}_k} \tag{1}$$

onde inter_{ij} é alguma medida de distancia entre grupos i e j e intra_k é alguma medida de distancia dentro de um grupo k
Coesão implica em um intra pequeno e separação um inter gande, portanto valores maiores de indice de Dunn são preferiveis

(Familia de) indice de Dunn

Medidas inter_{ij}

- distancia entre os centros dos grupos
- menor distancia entre os pontos de i e de j
- a media das distancias entre pontos de i e j

Medidas intra_k

- soma das distancias do centro aos pontos de k
- maior distancia ente dois pontos de k
- media das distancias entre pontos de k

Silhueta

Para cada dado i

- $a(i)$ é a distancia de i até o centro do seu cluster
- $b(i)$ é a distancia média de i até os dados do cluster mais próxmo
- $s(i) = \frac{b(i)-a(i)}{\max[a(i),b(i)]}$ é a siljueta do dado
- a silhueta é a média dos $s(i)$ para todos os dados
- $-1 \leq s(i) \leq 1$ e $s(i) \approx 1$ significa $a(i) << b(i)$ - portanto silhuetas altas são preferiveis

Medidas internas

- Há mutas outras medidas internas de qualidade da clusterização,
- Os criterios que se baseaim em uma visão que a distancia intra deve ser pequena e a inter grande são criterios pensados para o k-means onde os grupos são convexos.
- algumas métricas: Gap, figure of merit, intracluster variability, connectivity
- O livro texto tem um capitulo (14) bem detalhado sobre medidas de qualidade de cluster que deve ser consultado. Nas transparencias eu não fiz distinção entre medidas internas e medida relativas (que o livro faz).

Medidas externas

- Se os (ou alguns dos) dados clusterizados pertencem a classes já conhecidas, mas não usadas na clusterização, então essa informação pode ser usada para avaliar a qualidade da clusterização
- um criterio é a **concordancia**, ou seja os grupos devem corresponder as classes externas
- um criterio é **pureza** ou seja cada grupo deve conter dados de apenas uma classe. Note que pode haver mais grupos/cluster do que classes, mas cada grupo deve ser puro em relação a uma classe

Medidas externas

- A maioria dos indices externos se baseaim em 4 conjuntos:
- SS os **pares** de dados que pertencem ao mesmo cluster (clusterização) e a mesma classe (informação externa)
- SD os pares que pertencem a um mesmo cluster mas a classes diferentes
- DS cluster diferentes mas mesma classe
- DD cluster e classes diferentes
- para concordancia, SS e DD são bons. Para pureza SS, DS e DD são bons.

Algumas medidas externas

- Indice Rand $= \frac{|SS|+|DD|}{|SS|+|SD|+|DS|+|DD|}$
- Indice de Jaccard $= \frac{|SS|}{|SS|+|SD|+|DS|}$
- Indice de Fowlkes e Mallows (entre 0 e 1)
- Hubert normalizado (entre -1 e 1)
- Rand corrigido (entre -1 e 1 e 0 signfica que as concordancias são devido ao acaso).
- variação da informação

Medidas de clusterização em R

pacote `fpc` função `cluster.stats` ▸ link

- matriz de distancia em vez dos dados originais `dist(dados)`
- resultado do clustering `kmeans()$cluster`
- alt.clustering as classes externas (se existem)
- computa dunn, silhuette `avg.silwidth`, corrected rand e variation of information `vi` se classificação externa fornecida.

Como escolher o k do k-means

- a ideia é usar uma metrica (interna ou externa) e variar o k até achar um maximo (ou mínimo).
- infelizmente isso é mais teorico que pratico
- algumas metricas decrescem (ou crescem) sistematicamente como o k e ai voce deve procurar uma descontinuidade na curva geral
- link (ruidoso varias metricas) e link (várias medidas do Dunn e não apenas uma)

Como escolher o k do k-means em R

```
> library(fpc)
> dat=iris[,-5]
> ddat=dist(dat)
> clus=lapply(2:20,function(x) kmeans(dat,x,nstart=10)$cluster
> clus[1:2]
..
> stats=lapply(clus,function(x) cluster.stats(ddat,x))
> stats[1]
..
> out=sapply(stats,function(x) c(n=x$cluster.number,dunn=x$dun
> out
..
> par(mfrow=c(1,2))
> plot(out[1,],out[2,],xlab="k",ylab="dunn",type="l")
> plot(out[1,],out[3,],xlab="k",ylab="sil",type="l")
```

Como escolher o k em R

Função `kmeansrun` ▸ link faz a busca acima, mas para apenas 2 metricas

- asw - silluette
- ch - Calinski-Harabasz (?)

```
> bestkm=kmeansruns(dat,2:20,runs=10,criterion="asw")
> bestkm
..
> b2=kmeansruns(dat,2:20,runs=10,criterion="ch")
> b2
...
```

Variações do k-means

Gerar os centros iniciais, e

1. atribuir cada dado ao grupo cujo centro esta mais próximo
2. recomputar o centro do grupo como sendo a média dos dados que pertencem ao grupo

Outras definições de centro: k -medianas

- Mudanças no passo 2 acima:
- k-medianas: computa-se o cento como sendo a mediana dos dado do grupo.
- k-medianas é menos sensivel a outliers: pontos anomalos não puxam o centro para eles.
- implementado pela função `kcca` do pacote `flexclust` ▸ link

Outras definições de centro: k -medoids

- o centro é o dado do grupo cuja soma das distancias aos elementos do grupo é a menor
- k-medoids é o unico algoritmo da familia do k-means que funciona para dados relacionais (grafos) - os centros não são "pontos novos" do espaço mas sim um dos dados.
- implementado pela função `pam` do pacote `cluster` ▸ link

Outras definições de centro: k-medoids

```
> library(cluster)
> med=pam(iris[,1:4],k=3)
> med
Medoids:
      ID Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]   8          5.0         3.4          1.5         0.2
[2,]  79          6.0         2.9          4.5         1.5
[3,] 113          6.8         3.0          5.5         2.1
Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [79] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3 3 3 2 2 3 3
Objective function:
    build      swap
0.6709391 0.6542077
...
```

Variações algoritmicas

2 truques para big data:

- rode o k-means numa amostra (pequena) dos dados. Isto gera os centros, classifique os dados restantes pela proximidade dos centros. Se a amostra é grande o suficiente, os centros das amostras não é muito diferente do centro de todos os dados
- batch vs online

```
para todos os dados:
    compute a atribuiçao
para todos os dados:
    compute o centro
```

Online k-means

Versão online:

```
para todos os dados
    compute a atribuicao
    ande com o centro um pouquinho em direcao ao dado
```

- big data: escolha os dados aleatoriamente - *stocastic gradient descent*
- stream learning: use o dado assim que ele chega e nunca mais - *online learning*

Variações sobre a atribuição: fuzzy C-means

- cada ponto pertence a um grupo com "intensidade" proporcional ao inverso de uma potencia da distancia (normalizado, para que as intensidades somem 1 no final)
- fuzzy c-means tem um parametro extra (*m fuzzyfication*) relacionado com a potencia da distancia. *m* grande torna os intensidade de cada ponto mais ou menos parecida, e portanto os grupos tem grande interseção.
- $m = 1$ faz a intensidade ser 1 para o grupo mais perto e 0 para os outros, portanto o k-means tradicional.
- $m = 2$ é normalmente usado.
- o centro é a media ponderada (pela intensidade) dos pontos.

Fuzzy C-means em R

Implementado pela função cmean do pacote e1071 ▸ link

```
> library(e1071)
> cl=cmeans(iris[,1:4],3,m=2)
> cl
...
> head(cl$membership)
...
> cl5=cmeans(iris[,1:4],3,m=5
> head(cl5$membership)
...
```

GMM - distancia que depende dos grupos

- Nas variações anteriores, a noção de distancia se mantem. E se tivermos uma definição de distancia que depende do grupo (ou melhor dos dados do grupo)?
- distancia de Mahalanobis - distancia dividido pelo desvio padrao dos dados naquela direção (aproximadamente) link
- isto resolveria o problema de clusters "mais compactos" link

GMM - graus de pertencer a um grupo

- GMM (Gaussian MIxture Models) além de incluir a noção de distancia de Mahalanobis, inclui também a noção de graus de pertencimento.
- cada grupo é modelado como uma gaussiana (mutlidimensional). O centro é a media da gaussiana
- a *matriz de covariança* é o equivalente para varias dimensões do desvio padrao da gaussiana.
- Gaussianas são elispoides onde a matriz de covariancia indica quao não esferico o elispoide e a direção do raio maior.
- cada ponto tem um grau de pertencer a cada grupo dado pela probabilidade (normalizada) do ponto na gaussiana correspondente.
- o novo centro e a nova matriz de covariancia são calculados pela media ponderada (pelo grau) de todos os pontos.
- ▸ link r ▸ link

GMM em R

Varios pacotes implementam variações do GMM. O mais simples é o `Mclust` do pacote `mclust` ▸ link

```
> library(mclust)
> cl=Mclust(iris[,1:4],G=3)
> summary(cl,parameters=T)
...
> plot(cl,what="cla",dimens = c(2,3))
```

GMM em R

- GMM permite restrições no formato das gaussianas.
- esfericas ou elipsoidal
- mesmo tamalho, ou não
- mesmo formato se elispoidal, ou nao

```
> cl2=Mclust(iris[,-5],G=3,modelNames="VII")
> plot(cl2,what="cla",dimens = c(2,3))
```

Pacotes do R mencionado nesta aula

- fpc - métricas
- mclust - GMM
- e1071 - fuzzy c-means
- flexclust - k-medianas

Tarefa

- Use as 4 primeiras dimensões do iris mas agora padronize cada uma das 4 dimensões
- repita a analise da variação do indice de Dunn e da silhuette para valores de k entre 2 e 20.
- faça a mesma analise masa as medidas externas de corrected Rand e variation of information. Use a dimensão 5 do iris que é a classe de cada dado.
- discuta se há um ou mais valores que parecem apropriados para o k.