Lista 1 de Exercícios de Aprendizado por Máquina

Professores: Daniel S. Menasché, Edmundo de S. e Silva e Rosa M. M. Leão

1 Dependência do K-Means com Relação a Condições Iniciais

Como mostrado em sala, o algoritmo k-means converge para um ótimo local. Nessa questão, vamos mostrar que esse ótimo local pode estar muito distante do ótimo global.

Considere um conjunto de 5 grupos de pontos. Cada grupo de pontos possui N pontos, e a distância entre um grupo e cada um de seus vizinhos é B (vide figura abaixo). O raio de cada grupo de pontos é bem pequeno, e igual a δ, onde $\delta \ll B$.

1. Qual é a solução de clustering ótima global? Qual é, aproximadamente, a distorção associada a esse ótimo global?

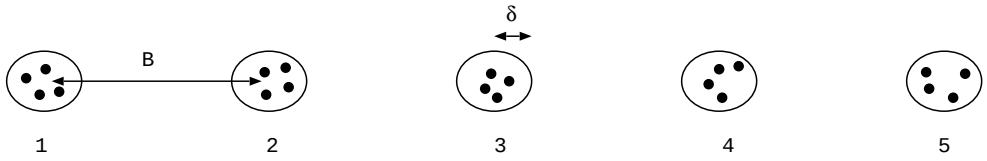

Figura 1: Exemplo de entrada do problema de clustering

2. Agora assuma que o algoritmo foi inicializado com um centróide no cluster 2, dois centróides no cluster 3, um centróide no cluster 4 e um centróide no cluster 5.

   (a) O que ocorrerá com as cores dos pontos após a primeira etapa E? E o que ocorrerá com a posição dos centróides após a primeira etapa M?

   (b) Quando é que o algoritmo convergirá?

   (c) Ao convergir, qual a distorção final?

(d) Na medida em B cresce, qual a diferença entre a distorção do ótimo local versus a distorção do ótimo global?

2 Misturas de Distribuições

O conjunto de pontos a seguir foi obtido considerando o seguinte experimento hipotético: um usuário de smartphone vai todo dia de casa para o trabalho e do trabalho para casa. Durante dois dias, a cada hora, o usuário envia um pacote de dados a um servidor que encontra-se em sua casa, e registra o atraso médio que leva para o pacote chegar ao servidor. O usuário armazena esses dados, e gera um trace, que indicamos na página a seguir. O i-ésimo número representa o atraso associado ao i-ésimo pacote enviado.

A partir do trace, queremos inferir quais pacotes foram emitidos de casa, e quais pacotes foram emitidos quando o usuário estava no trabalho.

1. Plote os dados das formas que achar mais convenientes.

2. Agrupe os pontos em dois clusters usando o algoritmo EM, assumindo que os pontos vêm de duas distribuições gaussianas. Indique claramente as condições iniciais utilizadas.

3. Assuma agora que os pontos vêm de duas distriuições exponencias.

   (a) Adapte o algoritmo EM para trabalhar com misturas de exponenciais. Em particular,

      i. resolva o problema de máxima verosimilhança para mistura de exponenciais

      ii. escreva o algoritmo completo claramente

      iii. agrupe os pontos em dois clusters usando seu algoritmo

4. Qual modelo melhor se adapta para clusterizar os pontos dados, mistura de exponenciais ou mistura de gaussianas? Justifique.

5. Teria sido vantajoso ter feito uso da correlação temporal entre os dados? Você pode verificar alguma anomalia em sua clusterização pelo fato de não ter usado correlação

temporal? Que ferramenta de aprendizado por máquina poderia ter sido usada para levar em conta a correlação temporal entre os dados? Como?

```
 7.8745
 9.0244
0.11303
 12.771
 12.041
 8.3501
 8.2138
 7.3548
 16.631
  6.984
 1.3099
 16.647
0.60308
0.19518
0.36838
0.34519
 1.6955
 1.1348
 1.8215
0.29858
 0.5679
0.29492
 1.2456
 2.3642
 50.471
 15.311
 6.1094
 6.3664
```

2.8233

4.3917

3.8202

3.3976

7.3414

12.589

21.353

6.2425

0.54136

2.4628

0.25492

0.34033

3.8651

1.5619

0.38951

0.79158

0.70681

1.0422

0.49119

0.78336