Aprendizaod de Máquina

Lista 4 - Entrega 30/10

Marcos Prates

Lista em Duplas. Se você usou R ou outro programa orientado a programação para resolver os exercícios práticos, submeta junto com as respostas o código utilizado para resolvê-las **(os envios devem ser feito pelo Moodle)**. Se você usou algum programa visual, explique qual programa foi usado, e como o problema foi resolvido nele (que funcionalidades foram usadas etc).

Exercício 1. Seu objetivo é usar as técnicas de redução de dimensionalidade, de segmentação (clustering) e de regras de associação para entender melhor um banco de dados que contém textos com resenhas sobre aplicativos da App Store do Android. Para isso, use a função `load` para carregar o banco `dadosReviewGoogle.RData`. Este banco contém dois objetos, `textos`, que contém as diferentes resenhas sobre os aplicativos, e `notas`, que contém as respectivas notas atribuidas pelos usuários que escreveram essas resenhas. Seu objetivo **não** é o de predição de notas, mas apenas o de melhor entendimento dos reviews.

(a) **[0.2 pts]** Mostre 5 resenhas do banco juntamente com suas respectivas notas.

Para os itens que seguem, você pode trabalhar com um subconjunto dos dados originais.

(b) **[0.2 pts]** Use o código fornecido para converter os textos em uma matriz documento-termo binária (isto é, cada entrada da matriz indica se um termo está presente ou não no respectivo texto).

(c) **[3.2 pts]** Use duas técnicas de clustering para criar agrupamentos dos diferentes textos (não use as notas para isso). Para a técnica que foi fornecida, faça também duas variações. Interprete os grupos obtidos por cada um dos métodos. Eles concordam entre si?

(d) **[3.2 pts]** Mostre as 5 regras de associação encontradas (não use as notas para isso) usando o algoritmo *a priori* com maior suporte, as 5 com maior confiança e as 5 com maior lift. Interprete o valor do suporte, lift e confiança de um regra de sua escolha. Mostre ao menos 3 maneiras distintas essas regras visualmente.

(e) **[1.6 pts]** Implemente componentes principais para esses dados (não use as notas para isso). Mostre quais são as 5 variáveis que recebem os maiores coeficientes (cargas) no primeiro componente. Mostre também as 5 variáveis que recebem os menores coeficientes (cargas) no primeiro componente. É possível interpretar essas palavras? Faça o mesmo com o segundo componente. Faça um diagrama de dispersão dos dois primeiros componentes principais. Use uma cor para cada ponto de acordo com a nota atribuida. Há um relação entre os componentes encontrados e as notas atribuidas? Você consegue encontrar outliers com base nesses

gráficos? Mostre ao menos três textos outliers. Repita o procedimento usando os três primeiros componentes, isto é, usando um gráfico em 3d.

(f) **[1.6 pts]** Implemente kernel PCA para esses dados, e trabalhe com ao menos duas variações dela. Plote novamente o gráfico de dispersão para essas novas técnicas. Eles são muito diferentes entre si? E com relação a componentes principais? Repita o procedimento usando os três primeiros componentes, isto é, usando um gráfico em 3d.

Exercício 2. Baixe o arquivo lista4.R. Ele mostra um código para baixar o banco de dados IncomeESL, que será utilizado neste exercício. Este banco mede diversas covariáveis em indivíduos americanos, como salário, origem e nível de educação. O código fornecido coverte este banco para o formato `transactions`, que será usado para implementar as regras de associação vistas em aula. Em particular, o código discretiza as variáveis numéricas. Usando o algoritmo a priori:

- **[1.0 pts]** Mostre as 10 regras (juntamente com suporte, confiança e lift) com maior lift entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3
- **[1.0 pts]** Mostre as 10 regras (juntamente com suporte, confiança e lift) com maior confiança entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3
- **[1.0 pts]** Plote as 10 regras com maior lift entre regras com suporte de ao menos 0.001, confiança ao menos 0.8, e tamanho máximo 3
- **[1.0 pts]** Mostre todas as regras (juntamente com suporte, confiança e lift) com maior confiança entre regras com suporte de ao menos 0.001, confiança ao menos 0.7, tamanho máximo 3, tamanho mínimo 2 e que tenha "ethnic classification=hispanic"do lado esquerdo da regra.
- **[1.0 pts]** Explore os dados você mesmo. Mostre ao menos duas regras de associação que você achou interessante além das já apresentadas na lista. Justifique porque as achou interessantes.

Exercício 3. Neste exercício você irá explorar alguns sistemas de recomendação para o MovieLens Dataset. Para tanto, instale a biblioteca `recommenderlab`, e carregue os dados usando `data(MovieLense)`.

1. **[2.0 pts]** Usando 75% dos dados para treinamento e assumindo que são dadas 12 avaliações por usuário, compare a performance dos seguintes métodos com relação a quão boas a predições das notas são:
   - Filtro colaborativo com base nos produtos com $k = 2$
   - Filtro colaborativo com base nos produtos com $k = 5$
   - Filtro colaborativo com base nos produtos com $k = 8$
   - Filtro colaborativo com base nos usuários com $k = 2$
   - Filtro colaborativo com base nos usuários com $k = 5$
   - Filtro colaborativo com base nos usuários com $k = 8$

   Você deve estimar o EQM (MSE em inglês), o REQM (RMSE em inglês) e o MAE.

2. **[1.5 pts]** Compare os mesmo métodos que o descrito no item anterior, mas desta vez usando os métodos de avaliação com base nas N melhores recomendações. Você deve considerar uma avaliação como sendo boa quando sua nota é maior ou igual a 4. Você deve estimar a sensibilidade, 1-especificidade, precisão e lembrança (recall) para $N = 1, 5, 10, 20, 50$ e 100 recomendações.

3. **[1.0 pts]** Algum dos métodos foi uniformemente melhor que os outros? Justifique