# Aprendizaod de Máquina

### Lista 2 - Entrega 24/09

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.** Baixe o conjunto de dados `dadosReviewGoogle.RData` (utilize a função `load` para ler os dados). O banco de dados que contém textos com resenhas sobre aplicativos da App Store do Android. 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 é manipular o texto de diversas formas e determinar suas características:

1. Faça uma numvem de palavras com os textos e discuta.
2. Qual o tamanho do Corpus? e do Vocabulário?
3. Remova as stop-words, coloque todas as letras em minusculo e remova palavras com frequência no Corpus menor do que 3. Qual o tamanho do novo Vocabulário?
4. Utilize os métodos: binário, tf, idf e tf-idf para preencher a matrix de termos. Plote as matrizes obtidas (Dica: transforme para a classe matrix e use o comando `image()`). Qual métodos você considera mais adequado de maneira geral? Por quê?
5. Utilize as notas e crie um novo vetor na qual a classificação é dita positiva se a nota $geq3$ e negativa caso contrário. Com esse vetor em mãos faça:
   - Ache os 5 termos com maior tf-idf (para todo o banco)
   - Ache os 5 termos mais comuns para as revisões positivas
   - Ache os 5 termos mais comuns para as revisões negativas

Dica utilize o pacote `tm`.

**Exercício 2.** Baixe o conjunto de dados `titanic.txt`. Cada observação deste banco é relativa a um passageiro do Titanic. As covariáveis indicam características deste passageiros; a variável resposta indica se o passageiro sobreviveu ou não ao naufrágio.

Seu objetivo é criar classificadores para predizer a variável resposta com base nas covariáveis disponíveis. Para tanto, você deverá implementar os seguintes classificadores, assim como estimar seus riscos via conjunto de teste:

- **[1.5 pts]** Regressão Logística. Mostre os coeficientes estimados.

- **[1.0 pts]** Regressão Linear. Mostre os coeficientes estimados.
- **[1.5 pts]** Naive Bayes.
- **[1.5 pts]** Análise de Descrimiante Linear
- **[1.0 pts]** Análise de Descrimiante Quadrática
- **[1.0 pts]** K-NN. Para isso você precisará transformar as covariáveis categóricas em numéricas. Você pode usar variáveis dummies.

Responda ainda as seguintes perguntas:

- **[0.5 pts]** Qual o melhor classificador segundo o risco estimado? Discuta.
- **[2.0 pts]** Para os classificadores baseados em estimativas de probabilidade, faça também as curvas ROC com o conjunto de teste. Faça também a tabela de confusão quando o corte usado é 0.5 e também quando o corte é aquele que maximiza sensibilidade mais especificidade. Comente.