**Universidade Federal de São Carlos**
Pós-Graduação em Ciência da Computação
Aprendizado de Máquina
Prof. Tiago A. Almeida <talmeida@ufscar.br>

# Exercício 3
# Naive-Bayes

## Objetivo

Neste exercício, iremos implementar o Naive-Bayes: um método tradicional de aprendizado de máquina baseado no Terorema de Bayes.

O Naive-Bayes é um método de aprendizado supervisionado baseado em probabilidade. O algoritmo consiste em aplicar o Teorema de Bayes supondo independência total entre os atributos de uma amostra. Mesmo utilizando conceitos probabilísticos simples, o método é capaz de obter resultados tão bons quanto outros métodos mais sofisticados, sendo extremamente escalável. É um método que se destaca em tarefas de Processamento de Linguagem Natural (PLN), como classificação de texto e detecção de SPAM.

Ao término desse exercício, espera-se que você consiga implementar o método Naive Bayes, entendendo todas as etapas de cálculo de probabilidade até a previsão final. Também espera-se que você seja capaz de resolver um problema de detecção de SPAM, implementando funções que podem ser facilmente reutilizadas em outros problemas de PLN.

## O exercício

Ao longo do exercício, você deverá completar cinco funções:

- `calcularProbabilidades`: responsável por calcular a probabilidade de ocorrência de cada atributo por possíveis classes da base;
- `classificacao`: responsável por realizar a previsão final, utilizando as probabilidades previamente calculadas;
- `text2features`: responsável por transformar um texto em um vetor numérico, permitindo que o mesmo seja utilizado no método supervisionado;
- `calcularProbabilidades_Laplace`: responsável por calcular a probabilidade de ocorrência de cada atributo por possíveis classes da base usando correção de Laplace;

- `classificacao_texto`: responsável por realizar a previsão final, adotando uma estratégia recomendada para bases de dados extremamente esparsas, como normalmente ocorre em problemas de PLN;

Preencha o código apenas nos espaços delimitados por comentários, normalmente iniciados por um comentário "`COMPLETE O CÓDIGO AQUI`" e instruções para a implementação.

**As implementações devem ser o mais genéricas possíveis, funcionando para qualquer conjunto de dados. Durante a avaliação, as funções serão testadas em bases com quantidade de amostras e atributos diferentes das fornecidas com o exercício, tendo em comum apenas o nome da coluna que contém a classe das amostras. Não adicione comandos do tipo `print` ou `display` dentro das funções que serão completadas, apenas o código da função.**.

Aproveite para interagir com o *notebook*: teste valores diferentes para as amostras que serão classificadas e entenda como o método funciona.

# Os casos de teste

Este exercício possui **5 casos de teste** que buscam avaliar cada uma das funções implementadas. A distribuição de tarefas avaliadas por cada caso é feita da seguinte forma:

- **Caso de teste 1:** corrige a função `calcularProbabilidades`;
- **Caso de teste 2:** corrige a função `classificacao`;
- **Caso de teste 3:** corrige a função `text2features`;
- **Caso de teste 4:** corrige a função `calcularProbabilidades_Laplace`;
- **Caso de teste 5:** corrige a função `classificacao_texto`.