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

Exercício 1
Processamento e visualização de dados

Objetivo

Neste exercício, iremos trabalhar com o processamento de dados e visualização.

Ao resolver um problema utilizando Aprendizado de Máquina, a maneira como os dados são preparados pode causar até mais impacto do que o algoritmo escolhido. Dessa forma, é muito importante que os dados sejam compreendidos e devidamente tratados.

Técnicas de visualização são bastante úteis para exibir, de maneira sumarizada, características interessantes dos dados, e técnicas tradicionais de pré-processamento podem deixá-los em formatos melhores para serem consumidos por métodos de aprendizado.

Ao término deste exercício, espera-se que você entenda como é realizada uma etapa tradicional de pré-processamento na resolução de problemas de Aprendizado de Máquina. Todos os procedimentos implementados podem ser facilmente aplicados em qualquer outra base de dados.

O exercício

O exercício encontra-se distribuído em três seções.

A primeira corresponde a um processo completo de tratamento das amostras do conjunto de dados `iris.csv`.

Em seguida, na seção **EXERCÍCIOS**, tarefas de visualização são aplicadas sobre a base de dados `data2.csv`, assim como a implementação de certos conceitos vistos na primeira seção.

Por fim, na seção **AVALIAÇÃO**, as funções implementadas nas seções anteriores são exaustivamente testadas para garantir o correto funcionamento do código.

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

Note que apenas na primeira e segunda seções existem códigos para serem completados. A seção AVALIAÇÃO serve apenas para que testes sejam feitos no Judge Online, e NÃO deve ser alterada.

Os casos de teste

Este exercício possui **5 casos de teste**. Eles buscam avaliar cada uma das funções implementadas, por meio das células presentes na seção **AVALIAÇÃO**. São avaliadas 7 tarefas:

1. Tratamento de amostras faltantes;
2. Remoção de duplicatas;
3. Remoção de inconsistências;
4. Normalização de amostras;
5. Remoção de *outliers*;
6. Cálculo de covariância;
7. Cálculo de correlação.

Os casos de teste são incrementais, i.e., os casos iniciais corrigem apenas um subconjunto de tarefas, aumentando de acordo com o número do caso. A distribuição de tarefas é feita da seguinte forma:

- **Caso de teste 1:** corrige a tarefa 1;
- **Caso de teste 2:** corrige as tarefas 1 a 3;
- **Caso de teste 3:** corrige as tarefas 1 a 4;
- **Caso de teste 4:** corrige as tarefas 1 a 5;
- **Caso de teste 5:** corrige todas as tarefas.