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

Exercício 9
Redução de Dimensionalidade

Objetivo

Na maioria dos problemas de aprendizado de máquina, a quantidade de atributos é bastante elevado. Embora grande volume de dados signifique mais capacidade para modelos inteligentes encontrarem relações, muitas vezes é comum desejar reduzir a quantidade de atributos. Essa redução pode ser feita para atender diferentes objetivos: (i) melhorar os dados na etapa de pré-processamento, (ii) reduzir o volume da base para torná-la adequada aos recursos computacionais disponíveis, (iii) descrever as amostras num espaço dimensional possível de ser graficamente representado, entre outros.

Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) é um dos método mais populares para redução de dimensionalidade. Sua origem data de 1901, no trabalho de Karl Pearson intitulado *"On lines and planes of closest fit to systems of points in space"* (*Philosophical Magazine, Series 6*, 2:11, 1901), primeiro a abordar conceitos da técnica. Décadas depois, Harold Hotteling consolidou e nomeou o método, em seu trabalho *"Analysis of a Complex of Statistical Variables into Principal Components"* (*Journal of Educational Psychology*, 24:6-7, 1933). O PCA foi introduzido na área da computação em trabalhos mais recentes, publicados no início do século XXI.

A redução de dimensionalidade através da Análise de Componentes Principais é feita através do cálculo de autovetores e autovalores sobre a matriz de covariância dos atributos da base de dados. Com estes dados, é possível projetar as amostras em espaços dimensionais menores, e reconstruí-los de forma aproximada para as dimensões originais.

Nesse exercício, você implementará um procedimento completo de redução de dimensionalidade. Através da implementação do PCA, e das etapas de projeção e reconstrução dos dados, espera-se que você compreenda claramente a teoria do método, assim como as possíveis aplicações de redução de dimensionalidade.

O exercício

Ao longo do exercício, você deverá completar 3 funções:

- `pca`: responsável por calcular a matriz de covariância e encontrar os autovalores e autovetores.

- `projetarDados`: responsável por utilizar os autovetores encontrados para projetar os dados em um espaço de dimensão reduzida; e
- `reconstruirDados`: responsável por utilizar os autovetores encontrados para reconstruir os dados para a dimensão original.

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 genéricas e funcionar para qualquer conjunto de dados. Na 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..

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:

- **Casos de teste 1:** testa a função `pca`;
- **Caso de teste 2 e 3:** testa a função `projetarDados`;
- **Caso de teste 4 e 5:** testa a função `reconstruirDados`.