GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula de exercícios

Exercício

Exercício: Este conjunto de dados contém informações genéticas usadas para a classificação de ancestralidade humana com base em análise de componentes principais (PCA). O conjunto é composto por dois arquivos: `genetic_data_train.csv` e `genetic_data_test.csv`.

Fonte: *Kaggle.*

As bases de dados estão disponíveis aqui.

Exercício

Dados de treino:

O arquivo `genetic_data_train.csv` contém dados de treinamento utilizados para ajustar um modelo que prediz a ancestralidade de um indivíduo. Ele inclui informações genéticas de $n = 183$ indivíduos, amostrados de diferentes populações ao redor do mundo. Os dados genéticos foram projetados nas $p = 10$ principais componentes (PC1 a PC10), que capturam uma proporção significativa da variabilidade (0,2416).

Cada linha representa um indivíduo, e as colunas contêm os valores dos componentes principais e os respectivos rótulos de ancestralidade. As categorias de ancestralidade disponíveis são: Africana, Europeia, Leste Asiática, Oceânica e Nativo-Americana.

Exercício

Dados de teste:

O arquivo `genetic_data_test.csv` contém os dados de teste usados para avaliar o desempenho preditivo do modelo treinado. Ele inclui informações genéticas de $n = 111$ indivíduos, cada linha representando um indivíduo, com colunas contendo os valores dos componentes principais e os respectivos rótulos de ancestralidade. O conjunto de teste também inclui indivíduos com ancestralidade "Desconhecida", bem como indivíduos Mexicanos e Afro-Americanos. Os cinco indivíduos com ancestralidade "Desconhecida" pertencem a uma das cinco categorias presentes no conjunto de treinamento, enquanto os indivíduos Mexicanos e Afro-Americanos apresentam diferentes graus de mistura histórica entre ancestrais.

Objetivo:

O objetivo deste exercício é desenvolver um modelo preditivo utilizando regressão logística (multinomial). Usando as componentes principais como variáveis explicativas, o modelo busca classificar com precisão a ancestralidade de um indivíduo com base em seus dados genéticos. O modelo treinado será então aplicado ao conjunto de teste para realizar as predições de ancestralidade. Avalie utilizar algum método de regularização.