EBAC

Aluna: Cristina Freitas

Módulo 23 – Random Forest

Exercício 1: Monte um passo a passo de Random Forest, explique com suas palavras e qual a diferença entre Bagging e Random Forest.

Passo a passo:

1) É gerado a partir de um dataset alguns outros subdatasets com o mesmo número de linhas do dataset original com reposição e são escolhidas colunas aleatoriamente;
2) A partir de cada conjunto de dados gerados é aplicado um modelo de classificação;
3) Com o resultado dos modelos se faz a agregação;

O **Random Forest** é um algoritmo de Ciência de Dados que implementa parcialmente a técnica de **Bagging**. O **Random Forest** é um tipo de **Bagging**. O algoritmo serve para determinar a variável alvo que pode ser fruto de uma árvore binária – como 0 ou 1 ou fruto de uma regressão logística – valores contínuos como o preço de alguma coisa.

Esta fase é chamada de **Bootstrap + Feature Selection**. A técnica consiste em pegar um conjunto de dados - dataset - e gerar aleatoriamente outros subdatasets com o mesmo número de linhas com a possibilidade de reposição. Neste contexto, algumas linhas podem ser repetidas. O que o Random Forest tem de novo em relação ao Bagging é a escolha das colunas. As colunas são escolhidas aleatoriamente e o número delas foi determinado pelo inventor da técnica Leo Breiman (2001). Essa escolha de colunas é chamada **Feature Selection**. A regra é:

- Para classificação: $m = \sqrt{p}$
- Para regressão: $m = p/3$

Cada conjunto de dados é submetido a um **modelo de classificação** que é uma árvore de decisão. Cada um desses conjuntos submetidos aos modelos de classificação é chamado **base learner**.

Com o resultado de cada um dos modelos é feita uma agregação. No caso de valores binários a agregação tem como resultado a maior frequência. No caso de resultados contínuos a agregação tem como resultado a média. Esta fase da técnica **Random Forest** é conhecida como **Aggregating**.

Random Forest utiliza o conceito de “Sabedoria das Multidões” que propõe que o conhecimento coletivo é maior do que o conhecimento individual (Surowiecki, 2004). O algoritmo é robusto a *overfiting*, mas há casos em que ele acontece.