pre_aula07.md

FA084 - Pre Aula 07

O objetivo desta atividade é solidificar os conceitos aprendidos até agora.

Desde a primeira aula, dividimos os conjuntos em treino e teste e testamos modelos diferentes. Mas isto, sem comparação e análise dos resultados pouco significa.

Nesta atividade vocês devem desenvolver um script que compare o desempenho de diferentes técnicas de modelagem em um conjunto de dados. Acredito que seja uma das últimas vezes que usaremos os dados do tinanic, não se preocupem.

O script

Forneci um script base com uma sugestão de rotina. Podem se ater a ordem dele mas se quiserem ser criativos, sem problemas.

O código deve conter:

1. Bibliotecas usadas
2. Comentários com descrição **geral** do que fazem em algum bloco de código ou etapa.
3. Organização. Ler código de outras pessoas é sempre complicado. Tem que haver um balanço entre não explicar nada e descrever demais.

O enunciado

Desenvolva um script para prever se um passageiro do Titanic sobreviveu usando o conjunto de dados fornecido. Os dados já estão divididos em treino e teste.

1. A rotina deve comparar o desempenho das ténicas KNN (`knn()`), regressão logística (`glm()`) e árvore de decisão (`rpart`).
2. Use cross validation com 5 folds.
3. Ajuste de hiperparâmetros:

- `knn` : escolha o intervalo de K que será testado e faça o ajuste para escolher o melhor.
- `rpart` : escolha o intervalo de variação adequado dos hiperparâmetros `cp` e `minbucket` e faça o ajuste que resulta na melhor combinação destes.
- `glm` : não há parâmetros para ajustar, use o threshold de probabilidade de `0.5` para fazer as predições. Ou seja se a probabilidade for `>= 0.5` o atributo meta é previsto como 1.

4. Use as métricas que julgar necessárias para comparar os modelos.

A entrega

Um arquivo `.zip` (`pre_aula_07_ra_primeiro_nome.zip`)com a estrutura de pastas que usamos desde o começo:

- Na pasta `data` , os dados.
- Na pasta `code` , um script com os detalhes da modelagem. Lembrem-se de comentar apenas o suficiente, É parte do aprendizado encontrar esse balanço. (`pre_aula_07_ra_primeiro_nome.R`)
- Na pasta `code` um pdf (até 2 páginas, não use print de códigos, apenas um relatório resumindo o que foi feito) com explicação dos resultados obtidos e conclusões da modelagem. (`pre_aula_07_ra_primeiro_nome.pdf`)

BÔNUS

Ajuste os parâmtros de alguma técnica de ensemble também (apenas uma das duas). Sugiro o pacote `randomForest` (para random forest) ou `xgboost` (para boosting).