Mini-Projeto 2 - Classificação

1. Introdução e Descrição dos Dados

A *Hangover S.A.*, empresa de produção de bebidas alcóolicas, buscava formas de aumentar suas vendas de vinhos. Para tanto, teve a ideia de tentar identificar quais as características que mais influenciavam na qualidade final do vinho. Caso seja possível identificar esta característica (ou conjunto de características) seria possível investir mais nos itens que mais tinham impacto na qualidade do vinho.

Para tanto, realizaram algumas análises sobre os vinhos. Os dados disponíveis são relativos à acidez (fixa, volátil e cítrica), o açúcar residual, o clorido, o dióxido (tanto o livre quanto o que possui enxofre), a densidade, o pH, sulfatos e o teor alcoólico. Todos numéricos, mas com faixas variadas de valores. Para cada combinação de características, produziu-se um vinho e este foi apresentado a um conjunto de enólogos campinenses que, sem conhecer as características, precisavam dar duas avaliações: na primeira, deviam indicar se o vinho era *bom* ou *ruim* e, na segunda, deviam graduar o vinho em *muito bom*, *bom*, *mais ou menos, ruim* ou muito *ruim*.

Entretanto, o trabalho desses enólogos é lento, caro e geralmente resulta em problemas de saúde. De posse de tantas informações, resolveram economizar neste processo e para isso, decidiram procurar os melhores analistas de campina grande para analisar os dados: vocês. Assim, espera-se que vocês encontrem uma forma automática de, partindo de outras combinações de vinho, identificar o quão bom um novo vinho será considerado pela crítica especializada!

2. Fases

Juntamente com esse documento foi disponiblilizado um arquivo de nome *mp2_fase1_vinhos_rotulados.csv* que contém 70% da base de dados rotulada, ou seja, todas as linhas possuem suas classes preenchidas nas variáveis-meta. Esses dados serão utilizados durante todo o projeto.

2.1. Fase 1: Análise e Seleção de Atributos

Objetivo

Analisar e selecionar os atributos mais importantes para as tarefas de classificação.

Descrição

Nessa fase você deverá entregar um documento com a análise realizada para seleção dos atributos mais relevantes para as duas classificações (binária e múltipla). A análise deve ser embasada em gráficos e práticas “estatisticamente corretas” (lembrem-se que esse será o

documento que atestará o seu aprendizado). Além disso, deverá ser enviado também o código fonte utilizado.

Entrega

- **26/11** (terça) até 11h59

Artefatos

- Análise em arquivo **PDF** (*mp2_analise.pdf*);
- **Código** Fonte em R.

2.2. Fase 2: Seleção de Modelos de Classificação

Objetivo

Comparar modelos de classificação, analisar seus resultados, selecionar o melhor para cada tarefa de classificação e utilizá-los para a predição.

Descrição

Nessa segunda fase serão disponibilizados os 30% restantes da base de dados (arquivo *mp2_fase2_vinhos_nao_rotulados.csv*), no entanto, esses dados não estarão rotulados. As rotulagens deverão ser previstas pelos modelos selecionados. Lembrem que um modelo é formado por: um algoritmo, seus hiper-parâmetros e seus dados de entrada, (i.e. os atributos escolhidos).

Cada grupo deverá então definir alguns modelos, treinar e selecionar os melhores para cada variável meta (utilizar os 70% dos dados para essa tarefa). Depois que tiver **ao menos 2 modelos para cada variável meta** (usando ao menos **2 algoritmos** diferentes dentre os vistos em sala[1]), repetir o seguinte processo 30 vezes:

1. Dividir os dados aleatóriamente em partições com 75% para treino e 25% para teste;
2. Treinar e testar cada modelo; e
3. Gerar um arquivo CSV (*mp2_validacao_modelos.csv*) com as seguintes colunas (contendo o cabeçalho e com aspas duplas nos nomes das colunas):

tipo_classificacao, nome_modelo, execucao, classe_prevista, classe_real

tipo_classificacao: {binaria, multipla}
nome_modelo: {um nome que diferencie os modelos}
execucao: [1, …, 30]
classe_prevista: {bom, ruim} OU {muito_bom, bom, mais_ou_menos, ruim, muito_ruim}
classe_real: {bom, ruim} OU {muito_bom, bom, mais_ou_menos, ruim, muito_ruim}

Após serem gerados os modelos, analisar os seus resultados, selecionar o melhor tanto para a variável *qualidade_binaria* quanto para *qualidade_multipla* (lembrar de explicar bem porque o modelo escolhido é o melhor) e gerar as predições para a **base com 30% dos dados**.

[1] Lembrar que, muito provavelmente, o R possui o algoritmo implementado e pronto pra usar.

Ambas as predições serão armazenadas em um arquivo CSV (*mp2_predicao.csv*) da seguinte forma (contendo o cabeçalho e com aspas duplas nos nomes das colunas):

classe_prevista_binaria, classe_prevista_multipla

classe_prevista_binaria: {bom, ruim}
classe_prevista_multipla: {muito_bom, bom, mais_ou_menos, ruim, muito_ruim}

Ao término **o mesmo documento (formato PDF) da Fase 1** deve ser atualizado com as **novas análises** realizadas para a seleção dos modelos. O mesmo deve ser feito para o **código-fonte**. Lembre-se que por serem duas variáveis-meta e 2 melhores modelos o processo de seleção deve ser analisado em separado.

Entrega

28/11 (quinta) até 23h59

Artefatos

- Análise em arquivo PDF (mesmo da Fase 1, *mp2_analise.pdf*, adicionadas as novas análises)
- **Código Fonte** em R
- Arquivo CSV com as **predições** para os dados de validação gerados em 30 execuções para cada modelo com o particionamento aleatório 75/25 (usar a base **com** os 70% dos dados com as variáveis-meta)
- Arquivo CSV com as **predições** para a *qualidade_binaria* e para a *qualidade_multipla* (usar a base com os 30% dos dados **sem** as variáveis-meta como entrada para a predição)

3. Aulas

3.1 Aula 26/11: Terça-Feira no Reenge-08

Essa aula terá duas partes:

1. Resolução do MP 1 ao vivo;
2. Tira dúvidas de R e MP 2.

Haverá lista de presença.

3.2 Aula 29/11: Sexta-Feira no Reenge-08

Apresentação dos resultados para todos. Inclusive para os professores. Haverá lista de presença. Atenção! Você será argüido(a) neste momento pelos professores, o que constará em sua avaliação! Esteja preparado(a)!

4. Lembretes

- São permitidos **grupos de 1 ou 2 alunos**; Em caso de duplas, ambos devem ser capazes de responder sobre TODO o projeto. Respostas como “essa parte quem fez foi X” não serão aceitas.
- Todos os artefatos devem ser compactados e enviados (no formato *.zip*, com o nome **mp2_<nome_grupo>_fase<1 ou 2>.zip**) para o email analise.de.dados.2@gmail.com;
- Codificação em **R**;
- Lembrem-se que utilizar **funções** em R é a melhor forma de manter um código **legível**. Mas não esqueçam de enviar também o “main”.
- Recomendamos que antes de codificar, comentem o passo-a-passo do script e ao término terão código e a documentação.
- **A pontualidade da 2ª fase é necessária, pois os dados enviados serão utilizados no dia seguinte**! Caso, na aula do dia 29/11 a sua resposta não esteja disponível, você receberá 0 pelo exercício.
- Não há provas na disciplina, as notas são dadas através do desempenho nos mini-projetos e pela presença nas aulas.

Bom aprendizado! =)