Mini-Projeto 1 - Regressão

1. Introdução ao Mini-projeto e descrição dos dados

A empresa **Home-Projects Brasil**, conhecida fabricante de imóveis climatizados tanto para o verão quanto para o inverno, decidiu realizar uma investigação quanto à **eficiência energética**: Existe impacto nas características em um **imóvel** (e.g., tamanho dos cômodos ou área das janelas) que podem impactar na eficiência energética tanto para a **refrigeração** quanto para o **aquecimento**? Para tanto, resolveu pedir a sua ajuda!

Após muitos anos no mercado coletando informações, a empresa forneceu a você tais dados referentes a estas experiências passadas. Os dados consistem em **oito** características do **imóvel**:

1 - compacidade relativa
2 - área da superfície
3 - área das paredes
4 - área do teto
5 - altura do imóvel
6 - orientação
7 - área com vidros
8 - distribuição da área com vidros

Para cada conjunto de características, foi medida a **carga** necessária para a refrigeração do ambiente assim como a carga necessária para o aquecimento do mesmo. Seu trabalho é identificar relações entre as características **fornecidas** e as cargas **necessárias** de modo a minimizar o custo elétrico com aquecimento e refrigeração. Utilize seus conhecimentos de **regressão** para alcançar este objetivo!

Este trabalho será dividido em **duas fases**: na primeira você receberá apenas um pequeno fragmento dos dados disponíveis, enquanto, na segunda fase, todos os dados estarão disponíveis.

2. Particionamento dos Dados

Em cada fase são disponibilizados dois *conjuntos de dados:* **treino** e **teste**. Na primeira fase, o conjunto de treino terá 10% dos dados totais da base **inclusive** os *atributos meta*, enquanto o conjunto de teste terá 5% dos dados totais mas **sem incluir** os *atributos meta*.

Por sua vez, na segunda fase, o conjunto de treino será composto por 60% dos dados totais e, o de teste, os 25% restantes.

Os dados de teste não serão acompanhados de seus respectivos *atributos meta* com o intuito de simular um teste no mundo real, em que conhecem-se as possíveis entradas mas desconhecem-se as devidas saídas. Desta forma, os bons modelos são aqueles que apresentam boas taxas de acertos no treino/validação mas que também generalizam seu poder

preditivo para outras instâncias inicialmente desconhecidas.

3. Seleção do Melhor Modelo

Os dados de treino disponibilizados em cada fase devem ser utilizados para *treino* e *validação* dos modelos propostos. A validação permitirá que os modelos sejam comparados e apenas um seja selecionado.

Para que o melhor modelo seja selecionado, pode-se utilizar qualquer técnica, ensinada em sala ou não, contanto que esta seleção possa ser explicada sob um ponto de vista estatístico.

Uma vez selecionado o melhor modelo, duas atividades serão realizadas:

- **Avaliação:** avaliar o modelo através da Soma do Quadrado dos Erros (SQE) sobre os dados de validação. Para tal, os dados de treino disponibilizados serão particionados em treino e validação, 75% e 25% respectivamente, de forma aleatória. Esse processo deve ser executado 30 vezes.
- **Predição:** predizer os *atributos meta* dos dados de teste.

4. Resumo das Entregas

Artefatos

Esses são os artefatos que serão entregues na Fase 1 e na Fase 2.

Fase 1

- Código da utilização dos 3 modelos de regressão estudados em sala (linear, múltipla e polinomial), mesmo se não utilizar todos;
- Código da Soma do Quadrado dos Erros (SQE) usado para comparar os modelos propostos;
- Duas Tabelas de Dados[1] com os resultados do SQE do melhor modelo (para cada atributo-meta) nas 30 execuções do experimento:
    - Os valores a compor o arquivo devem ser:
      número_da_execucao, sqe
    - O nome do arquivo deve identificar o atributo-meta.
- Uma Tabela de Dados com os *atributos meta* preditos pelo melhor modelo para os dados de *teste* disponibilizados:
    - Os valores a compor o arquivo devem ser:
      y_carga_aquecimento, y_carga_refrigeracao
- Texto[2] com a análise realizada durante a seleção do modelo e o por quê o melhor modelo é, de fato, o melhor. Essa análise deve ser embasada com argumentos estatisticamente válidos. O texto deve ser entregue no formato PDF.

[1] Todas as tabelas de dados geradas devem estar em arquivos CSV (comma-separated values) com o nome da coluna, mas sem o número das linhas.
[2] Um arquivo de texto (.txt) comum.

Fase 2
Na 2ª fase serão entregues os mesmos artefatos da 1ª com as modificações necessárias após receberem a segunda base de dados.

Lembretes

- Codificação em **R**;
- Lembrem-se que utilizar **funções** em R é a melhor forma de manter um código **legível**.
- Recomandamos que antes de codificar, comentem o passo-a-passo do script e ao término terão código e documentação.

Forma, Data e Hora das Entregas

Todos os artefatos devem ser compactados e enviados (no formato *.zip*) para o email analise.de.dados.2@gmail.com.
1ª Entrega: 14/11/2013 às 23h59
2ª Entrega: 15/11/2013 às 23h59

A pontualidade é necessária, pois os dados enviados serão utilizados no dia seguinte!
Mas, como o dia 15 é feriado e esse é o primeiro exercício, será dada uma pequena margem de atraso. Caso não entregue nas datas indicadas, você perderá 1 ponto pelo atraso mais 4 pontos por dia completo de atraso (e.g., caso entregue a 2a parte no dia 16/11/2013 a qualquer hora, você perderá 1 ponto, caso entregue no dia 17/11/2013 às 0h, você será penalizado em 5 pontos = 1 + 4).

Bom aprendizado!