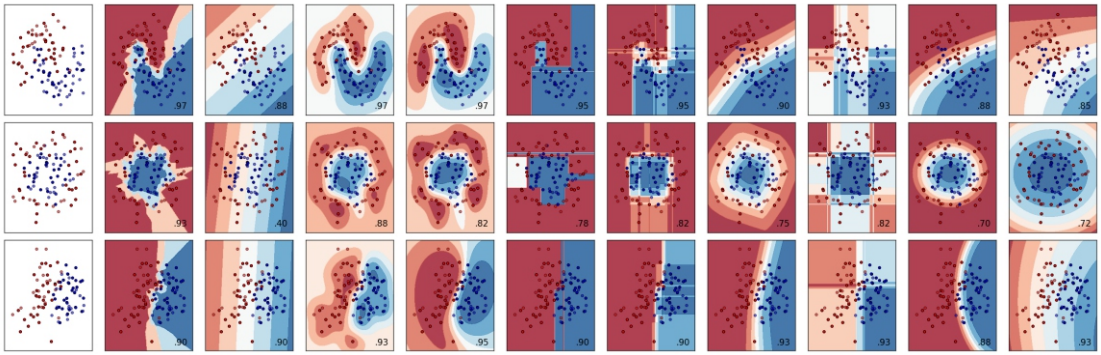
Engenharia e Software

Desafio Radix - Machine Learning



Informações e dicas importantes

- Para resolver o problema, é necessário ter acesso aos arquivos "p1_data_train.csv", "p1_data_test.csv" e "p1_predictions_example.csv". O link para download se encontra no enunciado do problema abaixo;
- Não é obrigatório apresentar solução completa para o problema, as ideias e criatividade dos candidatos serão valorizadas no processo de avaliação. Quem tiver interesse nas vagas de estágio, e tiver dificuldade em criar a solução, pode explicar em alto nível as suas ideias.
- O objetivo deste desafio é não somente avaliar os candidatos, mas também mostra aos interessados o tipo de problema que os selecionados irão resolver dentro da empresa. Dessa forma, os próprios alunos podem avaliar se tem interesse em trabalhar nessa área;
- No final desse documento, foram passadas referências que podem ajudar na busca de solução para o problema;
- Não há restrições quanto às ferramentas computacionais usadas para resolver o problema. Particularmente, as linguagems R e Python podem ajudar muito na solução. Na verdade, o próprio problema foi criado usando essas linguagens;
- Os candidatos interessados nas vagas devem enviar, além do currículo, um documento pdf com a solução (completa ou parcial) para o problema e também o arquivo "p1_predictions.csv", com as predições geradas para o problema;
- O arquivo enviado com predições deve estar no mesmo formato que o arquivo disponibilizado: "p1_predictions_example.csv";
- Cite no documento pdf com a descrição da solução as ferramentas computacionais utilizadas;
- Na solução pode-se também explicar o que não deu certo em hipotéticas tentativas sem sucesso.

Problema - Modelo Paramétrico

Link para download dos arquivos do problema (https://www.dropbox.com/s/y5zxefqamm5ww0u/desafio_radix_data.zip?dl=0)

Considere um processo industrial que opera em dois estados possíveis de qualidade: ótimo ou regular. A determinação do estado corrente do processo é feita por operadores humanos, que tem acesso a um painel de controle, exibindo uma série de variáveis medidas. Em intervalos regulares, os operadores classificam o estado de qualidade do processo, de acordo com os valores observados das variáveis medidas. Baseado na classificação produzida, ações adequadas são tomadas sobre o processo. Devido a limitações de informatização da planta, nem todas variáveis medidas são armazenadas em um banco de dados histórico.

Você foi contratado para investigar o processo de interpretação realizado pelos operadores. A sua investigação deve ser pautada em análises sobre uma base de dados amostrada do banco de dados histórico. A base ("p1_data_train.csv") a que você tem acesso possui as seguintes variáveis:

- Temp1, Temp2, Temp3 e Temp4, que são temperaturas medidas em diferentes pontos da planta;
- target , que é a variável de resposta, representando o estado de qualidade associado à amostra em questão. Esta variável é binária, com 0 representando o estado ótimo, e 1 o estado regular.

Para concluir sua análise, realize os seguintes passos:

- Faça uma análise exploratória dos dados, criando visualizações para as variáveis do problema;
- Ajuste um modelo de regressão logística para predição da variável de resposta;
- Com base na análise exploratória realizada, e nos coeficientes do modelo paramétrico treinado, enumere quais são as variáveis mais relevantes no processo de interpretação realizado pelos operadores;
- Para avaliar a qualidade da sua modelagem, o seu cliente enviou um conjunto de dados de teste ("p1_data_test.csv"). Utilize o seu modelo paramétrico treinado, para gerar predições para essas amostras de teste. As predições devem estar também no formato binário. Salve as suas predições em um arquivo chamado "p1_predictions.csv";
- Estime qual o erro de suas predições para o conjunto de teste;
- Enumere potenciais motivos de não ser possível ajustar um modelo com erro 0 para o conjunto de teste que lhe foi enviado.

Referências

Para os candidatos que tiverem dificuldades de expressar suas ideias através de alguma linguagem de programação, ou de encontrar soluções para os problemas anteriores, os seguintes links podem ajudar bastante na resolução dos problemas:

- http://www.numpy.org/ (http://www.numpy.org/)
- http://pandas.pydata.org/ (http://pandas.pydata.org/)
- http://scikit-learn.org/stable/ (http://scikit-learn.org/stable/)
- https://www.r-project.org/ (https://www.r-project.org/)
- http://topepo.github.io/caret/index.html (http://topepo.github.io/caret/index.html)