INF-0615 – Aprendizado de Máquina Supervisionado I

Exercício 02

Predição de Crise Financeira

1 Descrição do Dataset

Neste exercício, iremos predizer o nível de estabilidade de uma empresa, ou seja, o quão uma empresa está propensa a entrar ou não em crise econômica. O valor predito é contínuo e quanto maior este valor, menos suscetível esta a a impresa à crise. Segundo a descrição oficial da base[1], podemos considerar como limiar o valor -0.5, indicando que maior que este limiar a empresa está estável e, abaixo deste valor, está em crise. A base de dados apresenta os seguintes atributos:

- **Company**: Indentificador único para uma empresa;
- **Time**: período de tempo referente á séria temporal desta empresa;
- **x1 a x83**: variáveis anônimas relacionadas às características financeiras e não-financeiras da companhia. A variável x80 é categórica;
- **Financial Distress**: número que mensura a estabilidade financeira da empresa (valor alvo que vamos predizer);

2 Atividades:

1. Inspecione os dados. Quantos exemplos você tem? Como você irá lidar com as features discretas? Há exemplos com features sem anotações? Como você lidaria com isso?
2. Realize uma breve limpeza na base de dados e divida-a em treino, validação e teste. Verifique se os conjuntos estão disjuntos.
3. Normalize os dados de modo que eles fiquem mais bem preparados para o treinamento.
4. Como *baseline*, faça uma regressão linear para predizer o nível de estabilidade da empresa. Calcule o erro nos conjuntos de treino, validação e teste.
5. Implemente soluções alternativas baseadas em regressão linear através da combinação dos features existentes (multiplicação e/ou divisão) para melhorar os resultados obtidos no baseline. Compare suas soluções nos conjuntos de treino e validação.
6. Implemente soluções alternativas baseadas em regressão polinomial (elevando o grau de features) para melhorar os resultados obtidos no baseline. Plote o erro no conjunto de treino e de validação pelo grau do polinômio.
7. Tome a melhor solução baseado no erro no conjunto de validação, para cada item anterior, e reporte o erro no conjunto de teste.
8. Tome agora o melhor modelo de todos, ou seja, aquele com menor erro no conjunto de validação em toda a atividade. Ele mantém a mesma performance no teste ? Ou seja, ele apresenta o menor erro também no conjunto de teste ?

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *Financial_ Distress.csv*: conjunto de dados para ser dividido em treino, validação e teste;
- *Ex02.R*: Código de referência que implementa possíveis soluções para o exercício.

4 Referência

1. *Financial Distress Prediction.* `https://www.kaggle.com/shebrahimi/financial-distress`.