Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Aluno: Thiago Meireles Grabe

Github: https://github.com/EstopaGrabe

Origem: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/

Proposta

Este documento apresenta uma proposta de projeto para a finalização do curso de Engenheiro de Machine Learning pela Udacity.

Histórico do assunto

Em engenharia mecânica o termo confiabilidade é relacionado com a disponibilidade de um equipamento ou sistema em operar em seu estado ótimo, ou seja, em pleno trabalho sem gerar perdas na produção ou operação em que ele foi projetado. Num contexto mais específico, o estudo de confiabilidade pode associar-se à análise de manutenção, determinando instantes ótimos de tempo em que a mesma deve ser realizada. Tal análise tem como objetivo evitar que a falha ocorra, ampliando o tempo de funcionamento da máquina e buscando reduzir os custos.

O tempo de vida útil remanescente (*Remaining Useful Life – RUL*) de determinado sistema, máquina ou componente é definido como a extensão do tempo atual até o final de sua vida útil. Assim, desenvolver um plano de prognóstico de falha que possa monitorar e administrar adequadamente o sistema, estimando o RUL de maneira precisa, é de grande valia. Este plano visa permitir que as ações de manutenção sejam executadas antes que a falha ocorra, impactando positivamente a disponibilidade do sistema, evitando custos envolvidos com manutenção corretiva e relacionados com a indisponibilidade (BEZERRA SOUTO MAIOR *et al.*, 2016).

Tendo em vista a importância de termos uma manutenção preventiva e preditiva nos diversos setores industriais, alguns componentes mecânicos tem sido alvo de estudos importantes utilizando métodos computacionais ou *Data-driven* para reconhecimento de padrões e predição de quando um componente ou sistema irá falhar. Nesse aspecto, itens como rolamentos (ELFORJANI, 2016), filtros analógicos (HU *et al.*, 2015) e Turbinas aeroespaciais (FORNLÖF *et al.*, 2016).

Nesse aspecto, o meu interesse pessoal no assunto se dá pela minha formação acadêmica em engenharia mecânica. As análises convencionais de dados para ativos importantes como os citados acima ainda são precárias e baseadas em planilhas extensas de Excel. Essas planilhas são eficientes até certo ponto, pois quando se pensa em escalabilidade, confiabilidade de modelos e mesmo cenários flexíveis não podemos nos prender à uma ferramenta essencialmente de manipulação de dados manual. Com isso, ao estabelecer um modelo para determinação de RUL em um componente ou sistema mecânico terei uma

experiência em manipulação de dados e estabelecer um modelo de predição ou classificação para um *dataset* específico.

Descrição do problema

O problema a ser tratado neste projeto foi apresentado como uma competição internacional chamada *PHM08 Prognostics Data Challenge Dataset "https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository".* Neste evento o desafio era um *set* de Turbinas aeroespaciais do mesmo tipo. Cada Turbina tem um início de degradação e desgaste diferente, ou seja, os pontos de operação inicial e final não são conhecidos para os Turbinas. Sabe-se apenas que os Turbinas iniciam a operação em condições de operar normalmente e a degradação ocorre ao longo do tempo.

Os dados são estruturados e não apresentam textos ou imagens, apenas valores distribuídos no conjunto dos números reais.

A saída para o problema deve ser um valor real indicando o número de ciclos restantes para a turbina em teste.

Através desse cenário, o objetivo é encontrar o RUL de cada Turbina aeroespacial e predizer quando a Turbina irá falhar após os testes feitos. Essa métrica deve ser feita através dos números de ciclos restantes para cada Turbina após o teste realizado. Para isso, o problema será abordado como uma regressão dos dados para obtenção de um resultado final.

Conjuntos de dados e entradas

Os dados para o projeto estão separados por dados de treino e de teste. Os arquivos são de formato *.txt* com os respectivos nomes:

- Dados para treino: *Train.txt*
- Dados para teste: *Test.txt*

Em ambos arquivos, vemos 26 colunas e 45918 linhas de dados que descrevem a condição da turbina em um determinado ponto ou uma referência para o teste executado no equipamento.

Há duas colunas iniciais de referência para os testes. Nelas há informações da turbina em teste ("*Unit Number"*) e outra que detalha o número de ciclos da determinada turbina (*"Cycle Number*). Para o trabalho proposto a solução será desenvolvida para encontrar o "Cycle Number" em que a falhar ocorrerá. Logo, a nossa coluna *target* será Cycle Number".

As três próximas colunas trazem informações do teste em si, ou seja, os parâmetros de operação das turbinas durante o teste.

Há ainda 21 colunas que são dados de sensores que mostram as condições das turbinas durante a operação do teste. A maior dos sensores tem seus valores entre [-1 , 1] e seus valores ao longo do tempo mostram uma ligeira queda e, ao final do teste, extrapolam seus valores médios indicando alguma variação brusca nas turbinas.

Uma vantagem do *dataset* proposto é que ainda é possível verificar o resultado encontrado atráves do site da competição na qual pode-se submeter os resultados encontrados e obter o RUL para a solução encontrada.

Descrição da solução

Para solucionar o problema proposto pela competição, deve-se atentar ao cálculo do parâmetro (*Remaining Useful Life – RUL*). Com esse parâmetro é possível estimar sobre determinada condições de operação (Colunas de dados "*Setting"*) e medições de diversos pontos da turbina (Colunas de dados *"Sensor"*) o tempo de sobrevida das turbinas.

Essa estimativa é de suma importância para grandes empresas aéreas ao programarem manutenções preventivas. Através deste método, manutenção preditiva, se faz possível uma gestão mais refinada da frota de turbinas e saber em tempo hábil se um equipamento ou sistema irá falhar ou não durante a operação em que foi projetado para operar.

Por se tratar de dados numéricos e uma métrica de avaliação (*Remaining Useful Life – RUL*) também mensurável, a solução do problema será objetiva. O trabalho deve ser feito em função dos dados dos sensores e dos parâmetros de operação e fornecer uma métrica para *RUL*.

O desenho da solução do problema pode ser modificado ao longo do desenvolvimento das atividades, mas em suma o meu pensamento ao tratar os dados será utilizar as bibliotecas *Pandas* e *Numpy* para uma análise preliminar dos dados e encontrar correlação entre as *features*. O algoritmo *Boruta* também será utilizada para análise e dar uma direção de quais *features* são relevantes ao problema. Para esse problema de regressão dos dados, utilizarei algumas opções como *KNN, Decision Tree Regressor* e *Polynomial Regression*.

KNN é um algoritmo que utiliza os K vizinhos próximos para predizer seus valores. Essa abordagem é interessante e pode gerar bons resultados, pois os testes com as turbinas se dá de forma contínua em relação ao número de ciclos.

Decision Tree Regressor é um algoritmo que prediz o valor da variável alvo através do aprendizado de simples decisões inferidas pelas *features* dos dados. Essa abordagem para o problema pode representar um resultado satisfatório através de métricas bem definidas para a regressão aplicada. Além disso, esse algoritmo não necessita uma preparação dos dados mas os dados serão preparados através de algumas técnicas o que pode beneficiar essa abordagem.

Polynomial Regression é o algoritmo que expande os modelos de regressão linear para funções de graus maiores. A abordagem por regressão polinomial tende a ser através do método "tentativa e erro" para encontrar um grau da função que tenha um resultado melhor. Contudo, o algoritmo será utilizado para verificação da sua aplicabilidade no problema em questão.

Ao aplicar cada algoritmo verificarei a eficácia de cada um através das méticas R^2, *Explained Variance* e *Mean squared log error.* Pretendo utilizar essas três métricas para avaliar não somente o quanto as predições estão corretas mas também a variância e os erros.

Ao se estabelecer o melhor algoritmo de regressão para o problema das turbinas, utilizarei uma etapa de refinamento dos dados através do *GridSearch* para então ter os melhores parâmetros para o algoritmo escolhido.

A partir dessa análise, podemos escalar os estudos para outros modelos de turbinas ou mesmo transformar em um método de análise *online* das turbinas em operação.

Modelo de referência (benchmark)

Tanto na literatura como na internet não foram achados *Notebooks* ou textos explicitamente apontando para uma solução ao problema proposto, apenas referências teóricas para a solução que são apresentadas abaixo.

Para referência ao projeto a ser desenvolvido, o trabalho descrito por RAMASSO; SAXENA, (2014) traz alguns pontos para análise dos dados e então entendermos melhor o que cada coluna representa em um modelo de predição de falhas em turbinas.

Na descrição da competição (SCENARIO, 2008) são apresentados os vinte melhores resultados durante a competição para a métrica RUL.

No.	Score
1	436.841
2	512.426
3	737.769
4	809.757
5	908.588
6	975.586
7	1,049.57
8	1,051.88
9	1,075.16
10	1,083.91
11	1,127.95
12	1,139.83
13	1,219.61
14	1,263.02
15	1,557.61
16	1,808.75
17	1,966.38
18	2,065.47
19	2,399.88
20	2,430.42

Esses resultados podem ser utilizados como referência inicial para a solução implementada. Por se tratar de uma competição internacional os resultados acima apresentam um alto nível de complexidade de modelos de diversos níveis de maturidade em *Machine Learning.* Tendo em vista esse ponto, além dos 20 melhores resultados será utilizado o modelo criado através de regressão linear simples (chamarei de modelo *benchmark* ao longo do trabalho) para validar melhorias dos outros algoritmos implementados.

Métricas de avaliação

A métrica de avaliação é simples e objetiva. Para o cálculo da RUL, utilizamos a equação abaixo:

$$s = \begin{cases} \sum_{i=1}^{n} e^{-\left(\frac{d}{a_1}\right)} - 1 \text{ for } d < 0 \\ \sum_{i=1}^{n} e^{\left(\frac{d}{a_2}\right)} - 1 \text{ for } d \geq 0, \end{cases}$$

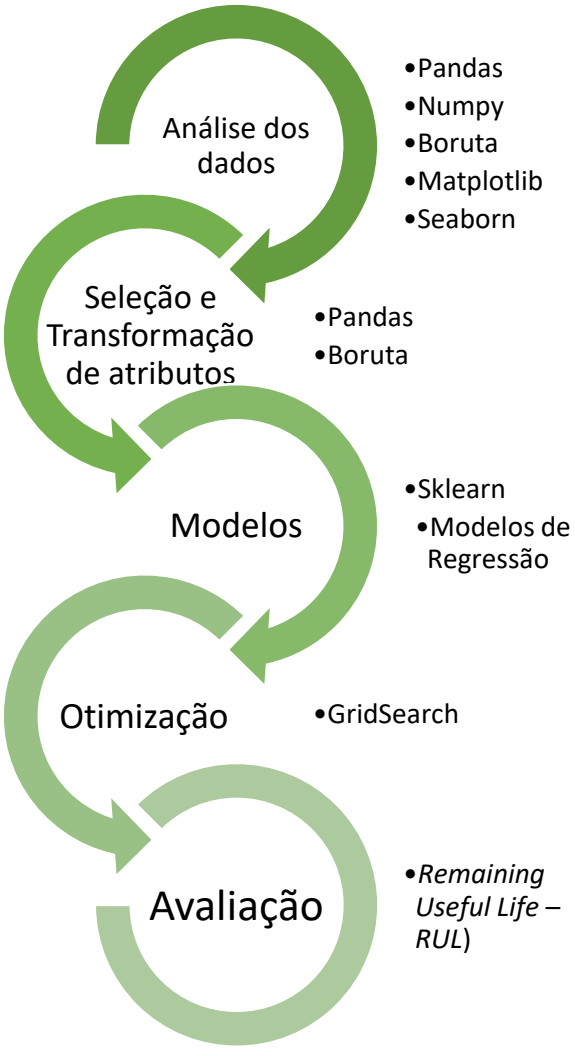
where,
s is the computed score,
n is the number of UUTs,
d = (Estimated RUL – True RUL),
$a_1 = 13$, and $a_2 = 10$.

Para essa avaliação deve-se obter o tempo remanescente para cada linha do *dataset* e comparar com os dados reais para a obteção do parâmetro "d". A partir disso, temos uma algebra relativamente simples para se obter o resultado.

Quanto mais próximo de zero o valor de "s", melhor foi a estimativa do modelo criado.

Design do projeto

Pretende-se utilizar o seguinte fluxo de trabalho:



Durante a etapa de análise de dados, pretende-se encontrar correlações entre as *features* e entender o comportamento de cada coluna de dados, bem como otimizar o entendimento de como a coluna alvo se comporta em relação à cada atributo. Um exemplo é a análise demonstrada na figura 1. Pode-se observar uma alta correlação entre os *features* de parâmetros do teste.

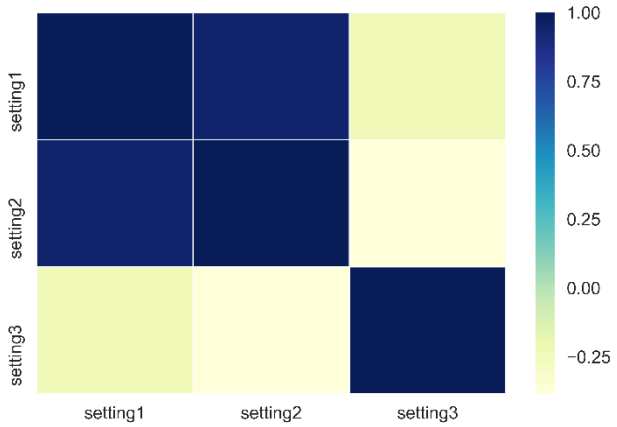


Figura1 – Correlação entre os parâmetros de teste

Na etapa de seleção e transformação de atributos, prentende-se entender quais atributos são de fato relevantes ao projeto além de agrupar aquelas *features* com maiores correlações com o intúito de se obter um *dataset* menor com mais significância dos dados. Com essa ação, podemos diminuir a fonte de erros.

Um exemplo simples é a transformação dos atributos *"setting1"* e *"setting2"* devido á alta correlçaõ. Pode-se então criar uma nova *feature* chamada, por exemplo, *"setting1/2":*

$$setting1/2 = (\text{setting1} + \text{setting2})/2$$

Após essas etapas de análise e transformação dos atributos, modelos de *Machine Learning* para regressão serão treinados e, através de métricas como R^2, *Explained Variance* e *Mean squared log error*, avaliados para posterior otimização através do *GridSearch.*

Por fim, avaliar os resultados de previsão do modelo elaborado e gerar o valor *RUL.* De posse desse resultado, pode-se iterar o ciclo procurando por pontos de melhorias como outros agrupamentos de *features* ou novos modelos de regressão e otimização.

Bibliografia

BEZERRA SOUTO MAIOR, C. *et al.* Remaining Useful Life Estimation by Empirical Mode Decomposition and Support Vector Machine. *IEEE Latin America Transactions*, v. 14, n. 11, p. 4603–4610, 2016. Disponível em: <http://ieeexplore.ieee.org/document/7795836/>.

ELFORJANI, M. Estimation of Remaining Useful Life of Slow Speed Bearings Using Acoustic Emission Signals. *Journal of Nondestructive Evaluation*, v. 35, n. 4, p. 1–16, 2016.

FORNLÖF, V. *et al.* RUL estimation and maintenance optimization for aircraft engines: a system of system approach. *International Journal of Systems Assurance Engineering and Management*, v. 7, n. 4, p. 450–461, 2016.

HU, Z. *et al.* Incipient Fault Diagnostics and Remaining Useful Life Prediction of Analog Filters. *Journal of Electronic Testing: Theory and Applications (JETTA)*, v. 31, n. 5–6, p. 461–477, 2015. Disponível em: <http://dx.doi.org/10.1007/s10836-015-5543-3>.

RAMASSO, E.; SAXENA, A. Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets. *International Journal of Prognostics and Health Management*, n. ISSN2153-2648, p. 1–15, 2014.

SCENARIO, E. PHM08 Prognostics Data Challenge Dataset. *Management*, p. 8–10, 2008.