Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Guilherme Thiago Gomes dos Santos

05 de maio de 2018

Proposta

Histórico do assunto

Desde o meu conhecimento de estudos de análises sobre big data, uma coisa que sempre me veio a mente foi o estudo de casos para a manutenção e melhoria dos nossos sistemas de saúde atuais.

Ambientes hospitalares necessitam de melhores formas de controle do ambiente, desde equipamentos, até monitoramento dos pacientes, com o objetivo de diminuir as fatalidades e sequelas, devido à problemas derivados dos quadros clínicos de cada um.

Descrição do problema

Minha proposta de projeto baseia-se em analisar informações de pacientes cardíacos, cujo sofreram um ataque cardíaco. A Partir dos dados recolhidos dos pacientes e utilizando modelos preditivos de aprendizagem supervisionada, deveremos analisar qual é o risco de que um novo paciente possa vir à óbito, e com isso, aumentar a observação dele e tomar medidas preventivas para que tal incidente não ocorra.

Conjuntos de dados e entradas

Os dados que estou utilizando foram obtidos através do site de Machine Learning da Universidade da California em Irvine (UCI).

O Dataset apresenta metadados sobre a idade do paciente ao ter um ataque cardíaco, se ele sobreviveu ao ataque cardíaco, por quantos meses ele ficou vivo após o ataque cardíaco, e se sofre de efusão pericárdica, entre outros.

Descrição da solução

Com o dataset acima citado, irei primeiramente validar quais registros estão aptos a serem analisados; Caso o registro não esteja, será descartado.

Após isso, verificarei via gráfico, quais são as idades em que as pessoas que sofreram um ataque cardíaco vieram a óbito e as que sobreviveram, para podermos ter uma noção da faixa etária mais vulnerável.

Também treinarei alguns modelos preditivos de aprendizado supervisionado, para ver qual é o mais eficiente para a nossa situação. Após esta etapa, pretendo realizar uma checagem para ver se todos os atributos que nós estamos verificando são imprescindiveis para realizar uma predição com um score elevado.

Modelo de referência (benchmark)

Utilizarei alguns modelos para testar a eficiência deles através do score e do tempo gasto para realizar o treino, o teste e a predição.

Dentre os modelos, utilizarei os algoritmos de:

- Support Vector Machines (SVM)
- K-N Neighbors
- Naive bayes

Também utilizarei o algoritmo de GridSeach para tuning do modelo escolhido.

Métricas de avaliação

Avaliarei qual é o melhor método preditivo através da comparação dos scores alcançados por cada um. O que possuir o score mais elevado e com um tempo de treino e testes adequado, será o escolhido.

Uma vez que nosso modelo é de classificação, utilizarei o método de f1-score para validar a precisão de cada modelo.

Design do projeto

Primeiramente, irei fazer o download do dataset através do site do repositório de Machine Learning da UCI :

- https://archive.ics.uci.edu/ml/machine-learning-databases/echocardiogram/echocardiogram.data (apesar do formato, ele é um arquivo .csv)

Após esta etapa, irei adequar o dataset, removendo as linhas que possuem caracteres inválidos, que possam interferir na nossa análise.

Também irei remover as colunas 'mult', 'name' e 'group', pois de acordo com a descrição do dataset, elas são descartáveis.

Em seguida, realizarei uma análise sobre os registros, como porcentagem de óbitos, sobreviventes, média de idade entre os pacientes, paciente mais novo a sofrer um ataque e também o mais velho.

Após esta etapa, irei fazer as análises preditivas, que vão desde a parte de cross validation, treinamento e teste, até análise do score e escolha do modelo preditivo melhor adequado para o dataset selecionado.