Prática: Lidando com Dados do Mundo Real (II)

Eduardo Prasniewski

1 Seção 6

KNN: Usado para classificar um ponto de acordo com os K's vizinhos mais próximos. É passado um exemplo prático usando os dados do Movielens, no notebook os eixos utilizados para detecção de vizinhos são a popularidade e a similiridade de cossenos entre os gêneros.

PCA: Principal Component Analisys, visa diminuir dimensões a partir de transformações lineares que preservam a informação. Muito usado para compressão de imagens ou extração de features. Assim como KNN, o professor passa um conteúdo prático usando dados de flores íris que já vem instalado na biblioteca scikit-sklearn, no qual o dataframe possuía 4 dimensões e foi possível reduzir a 2 com PCA, podendo assim até mesmo plotar um gráfico.

ETL ou ELT: Extract, Transform e Load. Extrair, transformar e carregar, essa sigla se refere as etapas de processamento que um dado é submetido até chegar no datawarehouse. Para aplicações mais simples é comumente usado ETL, porém para aplicações de grande porte, ELT, pois a transformação requer muito processamento computacional e isso pode ser abstraído na camada de Load, dependendo do software que é encarregado do datawarehouse (como o Hadoop).

Reinforced Learning: Explora o estado em que o objeto está e com base em estatística e algebra linear (vindas do método de Markov, eu procurei um pouco sobre assunto pois não gosto de apenas usar funções sem saber a origem matemática e lógica das mesmas) define qual é a melhor decisão a ser tomada, com em recompensas e punições. Uma das aplicações do aprendizado reforçado é Q-learning, que foi passado como prática no Jupyter, usando a biblioteca gym (percebi que algumas pessoas tiveram dificuldades com a biblioteca mas basta instalar a versão que o professor usa na vídeo aula), utilizando conceitos como look-up table (programação dinâmica) para otimizar o aprendizado.

Confusion Matrix: Basicamente uma matriz que relaciona os dados previstos com os dados reais, se aplicado de forma correta gera uma matriz com a diagonal de valores mais intensos em relação ao resto.

2 Seção 7

Bias: quão longe da média real está a média prevista

Variance: quão espalhados estão os dados previstos.

K-fold-cross-validation: trata-se de treinar com vários "buckets"ao invés de separar uma única vez entre dados de treino e validação.

Limpeza e padronização de dados é um dos pilares do Big Data, com dados confiáveis de qualidade é possível criar com modelos simples ótimos resultados. Na video aula o professor deseja pegar os dados dos logs do seu servidor web, de uma página de notícias de Orlando. Para isso ele teve que fazer inúmeros filtros principalmente para limpar dados que vinham de bots e para sua surpresa através dos logs conseguiu identificar que seu site estava sob um ataque malícioso.

Para lidar com pontos fora da curva, como por exemplo a média da maioria da população (ou seja removendo os bilionários) é necessário trabalhar com o desvio padrão assim, se estiver além dos limites do desvio, é retirado do conjunto de dados.

Após a limpeza é necessário saber as características (dimensões) do dataset, pois nem todas serão de utilidade e se tornam um "lixo"custoso, uma vez que estão sendo armazenadas e computadas. Para resolver isso pode ser usado algoritmos de redução de dimensão, como PCA e K-means.

Para substituir dados nulos de uma coluna, é possível usar a média ou a mediana (caso existam pontos fora da curva) porém este método não é muito recomendado, assim como remover a linha. Na verdade a melhor maneira é usando machine learning em conjunto com os outros dados ou adquirir mais dados.

Para lidar com dados desbalanceados (os quais podem gerar um modelo desbalanceado) pode-se replicar os dados minoritários, apagar dados majoritários, usar KNN para adicionar novas amostras minoritárias (SMOTE) ou ajustar o limite da probabilidade.

Binning trata-se de transformar dados numéricos em categórico, por exemplo dividir em buckets. Transformar os dados (pode ser criando uma nova dimensão) pode ser de grande ajuda, por exemplo aplicar uma função a um determinado valor. Uma outra ferramenta útil é encodig, assim como scaling e shuffling.