Relatório 10 - Lidando com Dados do Mundo Real

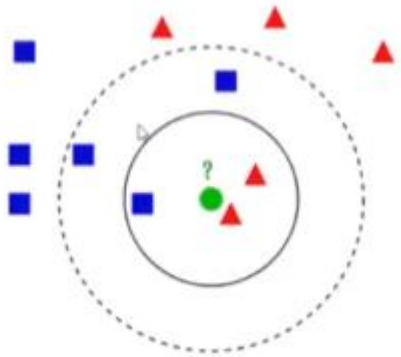
Guilherme Loan Schneider

Descrição da atividade

K-Nearest-Neighbors

É uma forma de classificar dados baseado na sua distância desde um dado conhecido. Esse tipo de classificação é aplicado em vários aspectos, mas é comumente encontrado ao aproximar as recomendações de usuários em algum tipo de serviço, como Spotify, Netflix, compras online, etc.

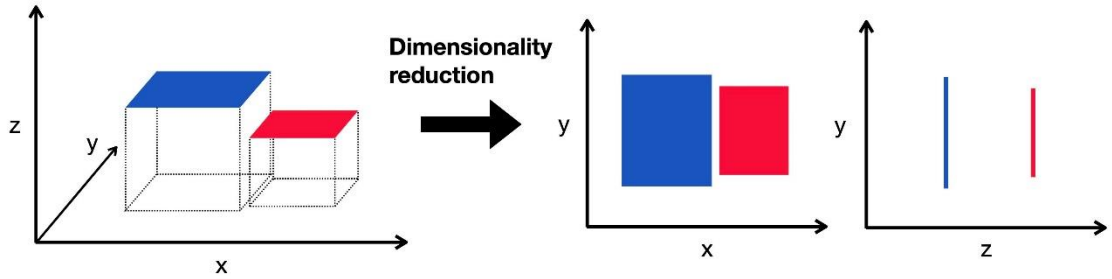
O "K" utilizado no nome do algoritmo refere-se ao tanto de pontos que serão encontrados por ele. Na figura abaixo, desde a origem até a linha contínua, o K é igual a 3, já da origem até a linha pontilhada, o K é igual a 5.



Dimensionality Reduction

A redução dimensional consiste em simplificar problemas que utilizam várias dimensões (chamadas de features), como o plano 4D e adiante. Essa técnica reduz as features para um plano observável, como o 2D, comumente chamado de plano cartesiano, enquanto mantém ao máximo a variância.

Na figura abaixo, é ilustrado esse tipo de técnica, que encontra planos facilmente analisáveis e aplica no cenário que faz sentido para o usuário. Esses planos encontrados são chamados de hiperplanos.

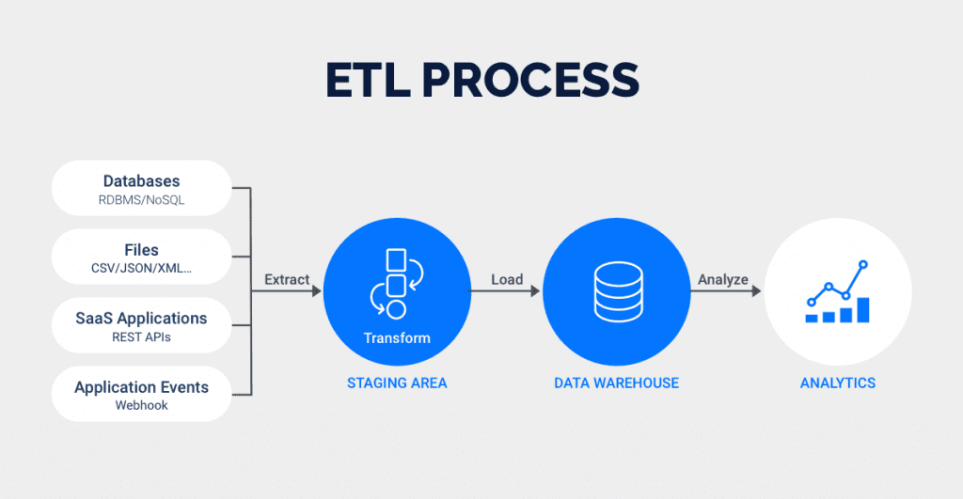


Data Warehousing

Data Warehousing é uma forma de armazenar dados de forma não estruturada ou estruturada. Esses dados podem ser de inúmeras fontes e de diferentes tipos, logo, é de extrema importância que seja feito o tratamento e limpeza desses dados.

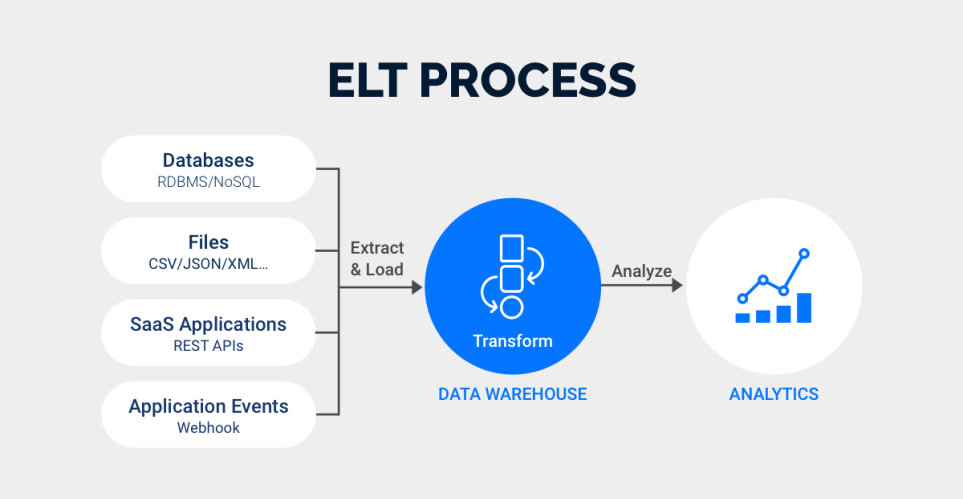
O processo ETL (Extract, Transform, Load) é frequentemente abordado para o tratamento e limpeza de dados. Consiste na extração dos dados, que serão obtidos de fontes

diferentes e formatos variados. Em seguida, passa para a fase de transformação, onde será aplicadas as técnicas de tratamento de dados, normalizando essas informações para o padrão adequado. Por fim, os dados são carregados no DW para posteriormente serem utilizados em análises, banco de dados, dentre outros.



Em sua grande maioria, o processo ETL lidará muito bem com dados de tamanho mediano, no entanto, ao nos depararmos com Big Data, o processo de Transformação acaba se tornando um problema.

Para solucionar esse problema, cria-se o processo ELT (Extract, Load, Transform), que basicamente permitirá que você utilize de recursos computacionais em nuvem para executar o processamento de tratamento de dados, tornando-o "menos custoso".



Reinforcement Learning

Esse tipo de aprendizado de máquina é utilizado quando um agente precisa explorar um determinado espaço, onde cada movimento escolhido possui um custo atrelado a ele, como se fosse um valor que representa o quão ótimo é aquele movimento.

Q-Learning

O Q-Learning é um tipo de aprendizagem por reforço, onde o Q representa o custo de determinado movimento.

Esse algoritmo possui:

- Um conjunto de estados representados por *s*

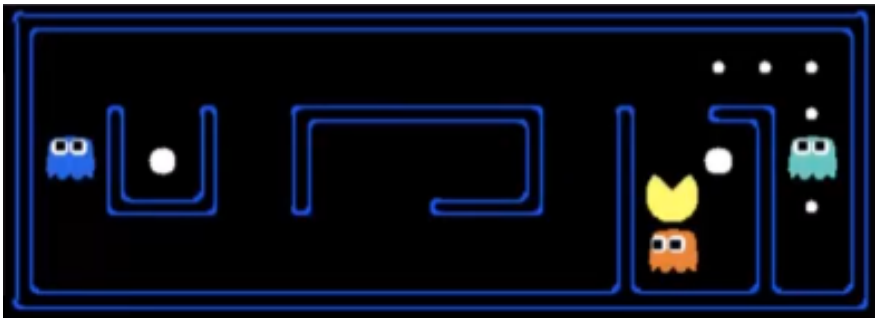
- Um conjunto de possíveis ações nesses estados: *a*
- Um valor para cada estado/ação: Q
- Inicia os valores de Q igual a 0

Ao iniciar o algoritmo, o agente explora o ambiente e quantificando (Q), onde caso algo ruim aconteça, o Q é reduzido, e se algo bom aconteça, o Q é incrementado. Preferencialmente, a decisão será sempre o melhor caminho (Q mais alto), no entanto, ao utilizar essa abordagem, é muito provável de haver perdas de caminhos. Para que isso não aconteça, é interessante adicionar uma variável μ, que funcionará da seguinte forma: caso um número aleatório seja menor que μ, o algoritmo não escolherá o Q mais alto, mas sim um caminho aleatório.

Na figura abaixo, é representado o jogo Pac-Man. O estado representado por *s* é o atual do agente (dado que ele pode ter feito outros movimentos anteriores a esse), as ações disponíveis são:

- Andar para cima;
- Andar para o lado direito;
- Andar para baixo.

O Q calculado para cada ação poderia ser, respectivamente, 5, 5 e -10.



Confusion Matrix

A matriz de confusão é uma tabela que mostra o desempenho de um modelo de classificação, comparando as previsões com os valores reais. É composta por quatro quadrantes que representam as combinações entre os valores previstos e os valores reais.

Os quadrantes da matriz de confusão são: Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN).

Measuring Classifiers

Ao analisar os resultados de uma matriz de confusão, é importante analisar os seguintes aspectos:

$$Recall = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoNegativo}$$

O recall pode ser utilizado quando os falsos negativos são importantes, como por exemplo, quantificar fraudes de um determinado serviço.

$$Precisão = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoPositivo}$$

Pode ser utilizado quando é importante a precisão de um determinado dispositivo, como por exemplo teste de drogas, reconhecimento de itens suspeitos.

$$Specificity = \frac{VerdadeiroNegativo}{VerdadeiroNegativo + FalsoPositivo}$$

$$F1Score = \frac{2VerdadeiroPositivo}{2VerdadeiroPositivo + FalsoPositivo + FalsoNegativo}$$

ou também

$$F1Score = 2\frac{Precisão * Recall}{Precisão + Recall}$$

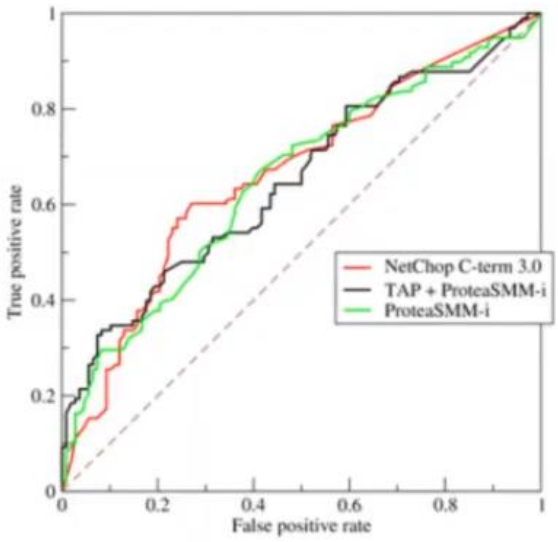
Utilizado quando a precisão e o recall são importantes.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

RMSE mede a acurácia, levando em conta apenas respostas certas e erradas.

ROC Curve

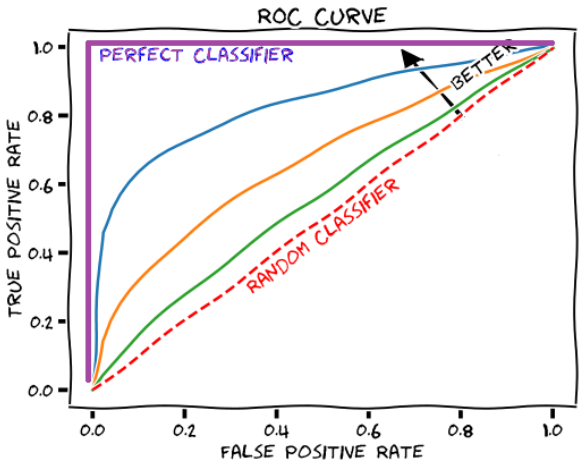
Chamada de Receiver Operating Characteristic Curve, consiste na representação da taxa de Verdadeiros Positivos (recall) pela taxa de Falsos Positivos.



As linhas em vermelho, preto e verde representam diferentes dispositivos classificadores de locais de clivagem do proteassoma humano, onde a melhor classificação é a linha que mais se aproxima do canto superior esquerdo do gráfico.

AUC

Chamada de Area Under the Curve, representa a probabilidade de que o modelo, se receber um exemplo positivo e negativo aleatoriamente, classificará positivo maior que negativo, ou seja, é utilizado para comparar diferentes classificadores. Uma classificação de 0.5 significa que é ruim, e uma classificação de 1 é perfeita.



Bias and Variance

Esses termos são utilizados como tipo de medição de algum tipo de dado. O Bias ou desvio, consiste em quantificar quão deslocados os dados estão do real valor esperado, ou seja, se eles estão deslocados, mas majoritariamente perto uns dos outros. A variância, por sua vez, indica o quão dispersos estão esses dados do real valor, ou seja, haverá alta variância se os dados estiverem muito "longe" uns dos outros.

No entanto, ao fazer uma determinada análise, é importante que você leve em consideração o erro, não apenas a minimização do desvio ou a variância.

$$Err = Bias^2 + Variance$$

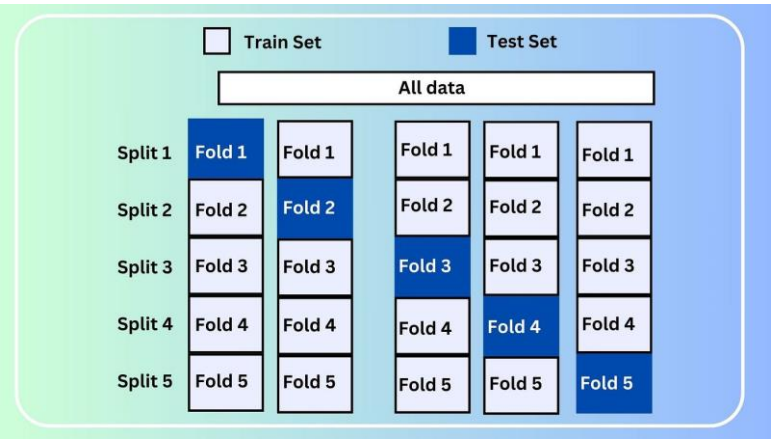
No caso do algoritmo KNN, quando se aumenta o valor de K, a variância é diminuída, por conta de existirem mais dados, no entanto, o desvio aumenta, visto que pode ser incluído muitos valores de um tipo, tornando-o tendencioso.

K-Fold-Cross-Validation to avoid overfitting

Esse processo consiste em dividir a base de dados de forma aleatória em K subconjuntos (em que K é definido previamente) com aproximadamente a mesma quantidade de amostras em cada um deles.

A cada iteração, treino e teste, um conjunto formado por K-1 subconjuntos são utilizados para treinamento e o subconjunto restante será utilizado para teste gerando um resultado de métrica para avaliação. Esse processo garante que cada subconjunto será utilizado para teste em algum momento da avaliação do modelo. Após obter os resultados para cada iteração, utiliza-se a média desses resultados.

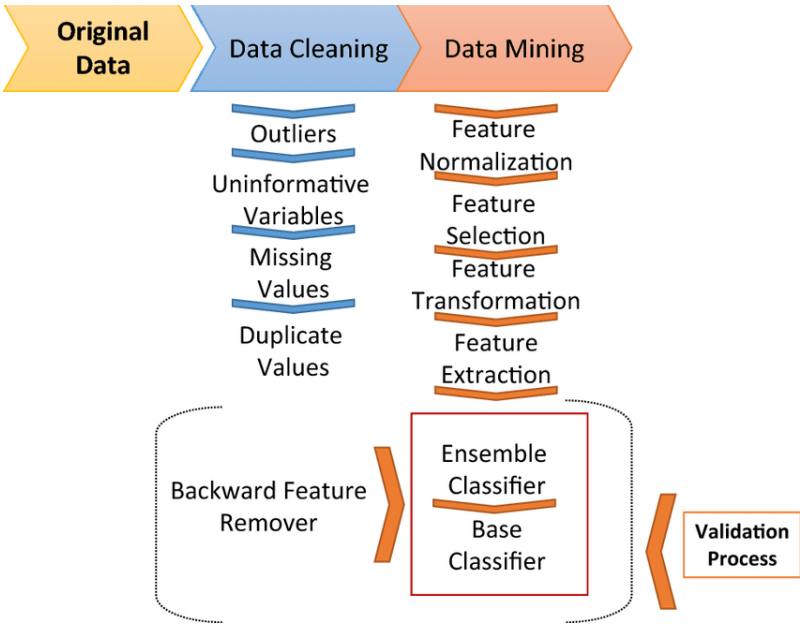
O valor de K deve ser escolhido cuidadosamente de forma que cada grupo de dados para treino e teste sejam grandes o suficiente para representarem estatisticamente o dataset original.



Data Cleaning and Normalization

Esse processo, similar ao processo de Transform do processo ETL, consiste em realizar a "limpeza" dos dados, que podem ser dos seguintes tipos:

- Outliers – Salário de bilionários entre cidadãos comuns;
- Dados faltantes, onde cabe ao cientista de dados decidir o que fazer com aquela determinada linha ou coluna de dados;
- Dados maliciosos, como ataques automatizados de bots.
- Dados incorretos, ou seja, dados que não fazem sentido, como letras aleatórias em locais de data, números em campos de email, dentre outros.
- Dados Irrelevantes, como colunas que não são importantes para a análise a ser feita.
- Dados inconsistentes, ou seja, campo de CPF com um CPF não válido, email sem o "@", dentre outros.
- Formatação, que é importantíssimo para gerar agrupamento nos dados.



Feature Engineering and Curse of Dimensionality

É o processo de manipular dados brutos para criar variáveis que ajudem modelos de machine learning a identificar padrões, ou seja, criar features que sejam relevantes para os modelos de machine learning, de forma a aumentar a precisão das previsões.

No entanto, criar features faz com que as dimensões a serem analisadas aumentem, tornando os dados mais dispersos e muito mais dados são criados. Logo, selecionar as features mais relevantes é importante nessa questão.

Existem algoritmos capazes de “selecionar” as features mais relevantes para o problema em questão, a técnica PCA (Principal Component Analysis) e a K-Means.

Imputation Techniques for Missing Data

Dropping

Pode-se levar em conta a opção de excluir linhas, caso não existam muitas linhas com dados faltantes e a exclusão destas não torne os dados tendenciosos.

Machine Learning

Utilizar o algoritmo KNN para encontrar os dados similares e aplicar a média para esses valores (leva em conta dados numéricos).

No caso de dados categóricos, uma opção é o Deep Learning, que é o mais complicado, mas consiste em utilizar (ou construir) um modelo de ML para inserir dados no seu próprio modelo de ML.

Existe também a Regressão, que encontrará relações entre valores lineares e não lineares para os dados faltantes.

Get More Data

Uma opção mais “fácil” é a de extrair mais dados, tornando os dados mais completos e possivelmente permitindo que você exclua linhas mais à frente.

Binning

Consiste em agrupar informações, sejam eles categóricos ou numéricos, em determinados intervalos para aumentar a precisão nas medições.

Existe também o Quantile Binning, que categoriza os dados de acordo com o seu lugar na distribuição dos dados, e separa-o em “baldes” de tamanho de amostras iguais.

Transforming

Como o próprio nome sugere, consiste em transformar um tipo de dado aplicando uma função matemática, tornando-o melhor para uma possível análise, visto que o custo computacional de processar uma função exponencial é muito maior que uma linear.

Encoding

É utilizado para transformar dados não manipuláveis por algoritmos em dados manipuláveis pelo modelo. O one-hot encoding é uma das técnicas para fazer esse tipo de manipulação.

A codificação one-hot é uma técnica para representar dados categóricos como vetores numéricos, em que cada categoria exclusiva é representada por uma coluna binária com um valor de 1 indicando sua presença e 0 indicando sua ausência.

Scaling / Normalization

O Scaling consiste em padronizar valores em um determinado intervalo e escalas, sendo útil para comparar variáveis diferentes em intervalos e escalas iguais.

A normalização altera o formato da distribuição dos dados, normalmente centrando o gráfico no valor (0,0), isso é útil ao aplica-los em modelos de redes neurais.

Shuffling

É utilizado para remover sinais da diferença entre os dados, no momento da coleta das informações, onde é feito o embaralhamento (similar a cartas) para que o treinamento não seja enviesado. Além disso, pode melhorar o desempenho de algoritmos.

Conclusões

A partir das aulas, foi possível compreender inúmeros aspectos, como o K-Nearest Neighbors, destacando sua aplicação em classificações baseadas nos “vizinhos”; a redução dimensional, que simplifica problemas de alta dimensionalidade mantendo a variância dos dados; o Data Warehousing, incluindo os processos ETL e ELT para gestão de dados estruturados e não estruturados. Também, o aprendizado por reforço, com foco no Q-Learning, e a matriz de confusão, essencial para avaliação de classificadores por métricas como recall, precisão e F1-Score.

Além disso, foram discutidos conceitos de Bias e Variance, estratégias de K-Fold Cross-Validation para evitar overfitting, e técnicas de data cleaning e normalization, fundamentais para garantir qualidade nos dados. Por fim, abordou-se a engenharia de features, a seleção de variáveis relevantes, e métodos de imputação, transformação, codificação, escalonamento e embaralhamento para otimização e análise eficazes dos dados.

Referencias

6. More Data Mining and Machine Learning Techniques;

7. Dealing with Real-World Data.