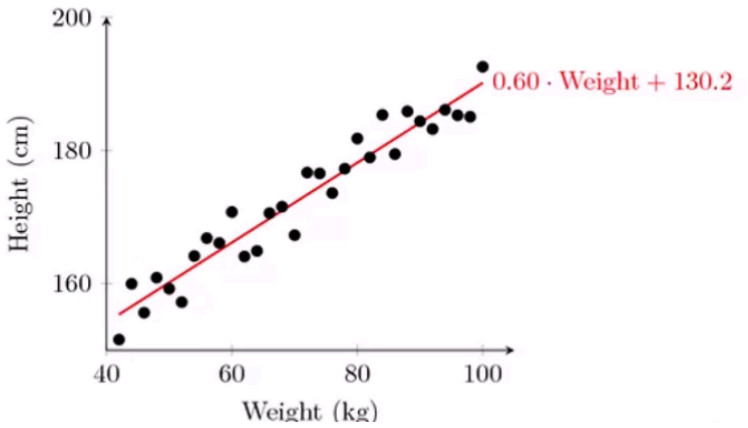
Relatório 11 - Prática: Predição e a Base de Aprendizado de Máquina (II)

Thassiana C. A. Muller

Regressão linear

A ideia é encontrar uma linha (ou hiperplano no caso de múltiplas variáveis independentes) que melhor se ajuste aos dados observados, minimizando a distância entre os valores observados e os valores previstos pela linha. Assim, obtém-se uma função de primeiro grau que relaciona uma variável dependente a uma ou mais variáveis independentes. A confiabilidade é dada pelo fator R^2 que vai de 0 a 1.

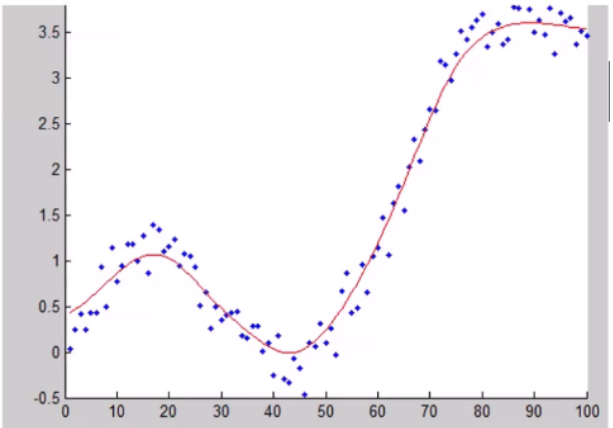
Exemplo de regressão linear:



Regressão polinomial

Semelhante a regressão linear, porém aqui a função que aproxima os dados é de ordem maior ou igual a 2.

Exemplo de regressão polinomial:



Modelos Multi-level

São uma classe de modelos estatísticos que permitem a análise de dados estruturados em diferentes níveis, são úteis quando os dados tem uma estrutura hierárquica ou aninhada, como alunos dentro de escolas ou pacientes dentro de hospitais.

Treinamento supervisionado

O modelo é treinado usando um conjunto de dados rotulados, ou seja, para cada entrada, o modelo conhece a saída correta (rótulo). O objetivo é aprender a mapear entradas para saídas para fazer previsões sobre novos dados.

Treinamento não supervisionado

O modelo tenta identificar padrões ou agrupamentos (clusters) nos dados por conta própria, sem conhecimento prévio das saídas corretas.

Conjunto de treinamento

É a porção dos dados usada para treinar o modelo. Ele ensina o modelo a reconhecer padrões ao ajustar seus parâmetros.

Conjunto de teste

É um conjunto separado de dados usado para avaliar o desempenho do modelo treinado, garantindo que ele se generalize bem para novos dados.

K-Means Clustering

É usado para descobrir padrões e grupos ocultos em seus dados. Ele funciona a partir de um aprendizado não supervisionado e cria um K número de grupos ou clusters a partir dos dados de treinamento buscando minimizar a distância entre os pontos de cada cluster, não sendo função do algoritmo definir rótulos ao grupos.

Exemplo do processo de clusterização:

Entropia

É uma medida do grau de desordem dos dados. A entropia é amplamente usada em algoritmos de classificação, especialmente em métodos de árvores de decisão. Esses algoritmos utilizam a entropia para decidir como dividir os dados em cada nó da árvore.

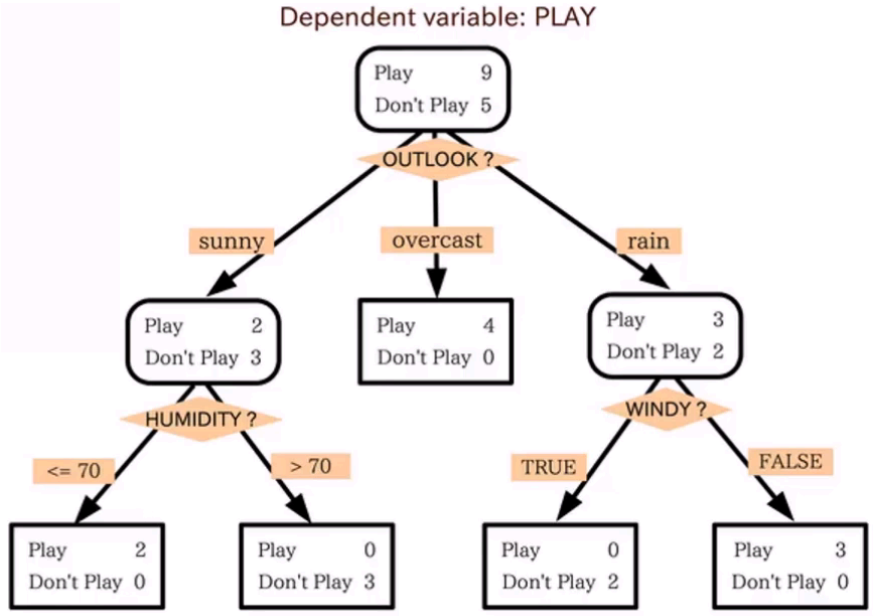
Árvores de decisão

Semelhantes a fluxogramas, as árvores de decisão ramificam ações de acordo com estados atuais. São usadas em computação para a tomada de decisões e até mesmo predição de dados.

Sua estrutura básica é geralmente composta por:

- Raiz (Root Node): O nó inicial da árvore, onde o primeiro atributo é avaliado.
- Nós Internos (Internal Nodes): Pontos de decisão na árvore que representam a escolha de um atributo a ser analisado.
- Ramos (Branches): Conexões entre os nós que mostram os resultados das decisões.
- Folhas (Leaf Nodes): Nós terminais que representam a classe final ou valor predito para um conjunto de entradas.

Exemplo de árvore de decisão

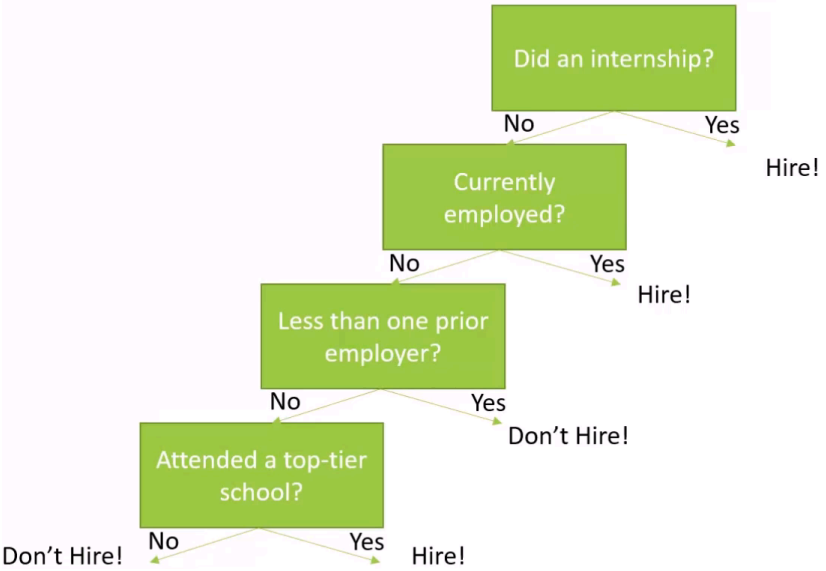


Exemplo de filtro automático de currículos a partir usando árvore de decisão.

Digamos que uma empresa queira filtrar os currículos mais relevantes, primeiro informa-se o histórico de contratações e alguns atributos dos currículos:

Candidate ID	Years Experience	Employed?	Previous employers	Level of Education	Top-tier school	Interned	Hired
0	10	1	4	0	0	0	1
1	0	0	0	0	1	1	1
2	7	0	6	0	0	0	0
3	2	1	1	1	1	0	1
4	20	0	2	2	1	0	0

Constrói-se a seguinte árvore de decisão



Primeiramente, é encontrado algum atributo que possa ser usado para separar o dataset de acordo com alguma métrica, nesse caso de forma a minimizar a entropia dos dados no próximo passo, esse método é chamado de ID3, não é o melhor método, mas funciona!

Random Forests

Para evitar o *overfitting* que pode surgir em árvores de decisão, criou-se as random forests que são o conjunto de múltiplas árvores de decisão, onde cada árvore é treinada em uma amostra diferente dos dados, ajudando a reduzir o overfitting e melhorar a precisão.

Ensemble Learning

É o uso de vários modelos de aprendizagem de máquina para resolução de um problema, como exemplo a Random Forest usa *bagging.*

Bagging é a criação de várias versões do mesmo modelo de aprendizado treinados em diferentes subconjuntos dos dados de treinamento, criados através de amostragem com

reposição (bootstrap). Cada modelo é treinado de forma independente, e as previsões finais são feitas pela média (para regressão) ou pela maioria dos votos (para classificação) das previsões de todos os modelos.

Outro tipo de ensemble learning é o boosting que treina vários modelos sequencialmente, onde cada modelo tenta corrigir os erros dos modelos anteriores. Em cada iteração, os exemplos mal classificados ganham mais peso, incentivando o próximo modelo a focar nesses exemplos. As previsões finais são uma combinação ponderada das previsões de todos os modelos.

SVM

É um algoritmo de aprendizado de máquina supervisionado usado para classificação e regressão. No entanto, é mais amplamente utilizado em problemas de classificação. O objetivo principal do SVM é encontrar o hiperplano que melhor separa as classes de dados no espaço de características (feature space).