Aprendizagem Supervisionada vs Não Supervisionada.

A análise de dados é uma disciplina muito variada, onde podemos encontrar muitos tipos de problemas, seja pelo assunto, pelos tipos de dados a trabalhar ou pela técnica a ser utilizada. Vamos nos concentrar nas técnicas.

Temos, por exemplo: regressão logística e random forest. Estas são técnicas de resolução de problemas do tipo de aprendizagem supervisionada, onde os dados-base já foram classificados anteriormente. Por outro lado, também temos k-means ou componentes principais. Essas são técnicas aplicadas para resolver problemas de aprendizado não supervisionado, onde os dados ainda não foram classificados.

Formalmente, temos 4 tipos de aprendizagem. Temos a supervisionada. É o tipo de problema em que sabemos a resposta. Quer dizer, há uma resposta certa e uma resposta errada. Existem vários modelos para atacar este tipo de problema.

Há os de classificação e os de regressão. Os de classificação têm o objetivo de localizar uma tag nos dados, por exemplo, para prever se um cliente vai ser inadimplente. Por outro lado, há os problemas de regressão, que são aqueles em que o resultado do modelo é um número. Por exemplo, prevendo a quantidade de chuva para os próximos 5 dias.

E temos a aprendizagem não supervisionada. Ao contrário da aprendizagem supervisionada, neste caso não podemos saber a resposta com antecedência, mesmo com informação histórica. Por exemplo, uma vez concluído um empréstimo, podemos avaliar se o cliente foi inadimplente ou pagou a tempo. O mesmo se aplica ao caso da chuva.

No entanto, no caso da aprendizagem supervisionada, não temos isso porque não há necessariamente respostas certas e erradas. Este tipo de problema tem normalmente como objetivo obter conhecimentos sobre o conjunto de dados, reduzindo a dimensionalidade, agrupando dados semelhantes, relacionando-os uns com os outros.

Alguns modelos utilizados neste tipo de aprendizagem são: K-Means, clusters hierárquicos e análise de componentes principais. Por exemplo, se formos confrontados com a necessidade de identificar quantos tipos de clientes temos, utilizaremos este tipo de modelo. O algoritmo analisará os dados procurando semelhanças e irá sugerir algumas classificações.

Entre supervisionada e não supervisionada, temos uma aprendizagem semi-supervisionada. Este é um caso muito particular, normalmente utilizado na medicina, onde os dados são muito limitados e só há conhecimento de parte das tags ou respostas corretas.

Portanto, eles devem conseguir tirar o máximo proveito da informação. Este tipo de aprendizagem consiste na utilização de uma combinação de aprendizagem não supervisionada com a informação das tags da aprendizagem supervisionada. Finalmente, temos a aprendizagem por reforço. É a mais diferente das anteriores, porque se baseia em tomar decisões e interagir continuamente com o ambiente. Esse tipo de aprendizagem é usada em robótica e videogame. Pode ser com modelo ou sem modelo.

Vamos revisar rapidamente os conceitos de supervisionada versus não supervisionada. Por exemplo, para identificar se o animal que está bebendo água na imagem é um cão ou um gato estamos diante de um problema de aprendizagem supervisionada, já que inequivocadamente, é um cão e não um gato. Novamente, se pedirmos para fazer uma previsão sobre o valor de fechamento do Dow Jones na semana, também estamos olhando para um problema de aprendizagem supervisionada. Neste caso, regressão.

Por outro lado, se quisermos saber quantos tipos de atacantes de futebol existem no mundo, considerando em conta gols e jogos disputados, enfrentamos um problema de aprendizagem não supervisionada. Resumindo, se um resultado é um número ou uma categoria conhecida como sim e não ou cão ou gato, utilizamos aprendizagem supervisionada.

Se temos um conjunto de dados que ainda não sabemos como classificar, utilizamos aprendizagem não supervisionada.