

Revisando a classificação

Transcrição

Antes de discutirmos sobre os problemas de aprendizado não supervisionado, aprenderemos sobre o supervisionado e entender suas diferenças.

Relembremos a questão da empresa de cartões de crédito: Qual o risco de os clientes atrasarem o pagamento de faturas?

Pensaremos em um cliente em específico que chamaremos de Pedro Lauro e queremos classificá-lo em alguma categoria de risco, que são: baixo, alto ou médio.

Podemos utilizar uma técnica de aprendizado de máquina, em que a partir de uma base de dados de clientes que já conhecemos o risco, conseguiremos encaixar Pedro Lauro em alguma dessas categorias. Essa técnica é o **aprendizado supervisionado de classificação**. Mas como podemos validar se essa classificação está de fato correta?

Não podemos validar ainda a categoria do Pedro Lauro porque não temos dados o suficiente.

Vejamos um grupo de clientes em que já sabemos o risco de atrasarem a fatura. Maria da Silva, por exemplo, possui um risco baixo, essa informação é chamada de **rótulo** ou label. Teremos mais alguns clientes: João de Cruz, risco alto e Amadeu Romeu, risco médio.

Se utilizarmos a mesma técnica de aprendizado de máquina para descobrir o risco do Pedro Lauro, teríamos um resultado incorreto: Amadeu Romeu sairia como risco alto, e na verdade ele possui risco médio. A taxa de acerto foi de apenas 66%.

Passemos para outra situação a ser estudada: qual o comportamento dos clientes com o cartão de crédito, e isso inclui

- Valor gasto
- Número de transações
- Limite total
- Frequência de compras

No próximo vídeo analisaremos as respostas.