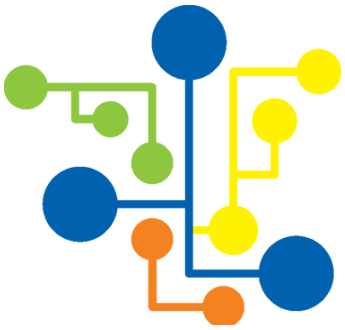

Data Science Academy

www.datascienceacademy.com.br

Matemática Para Machine Learning

Técnicas de Redução de Dimensionalidade

A redução da dimensão do espaço de recurso é chamada simplesmente de "redução de dimensionalidade". Existem várias maneiras de obter a redução de dimensionalidade, mas a maioria dessas técnicas se enquadra em uma das duas classes:

- Eliminação de recursos
- Extração de recursos

Eliminação de recursos é o que parece: reduzimos o espaço de recursos eliminando recursos (atributos ou variáveis). Em vez de considerar todas as variáveis, podemos descartar todas as variáveis, exceto as três que acreditamos prever melhor como será o produto interno bruto do Brasil, por exemplo. As vantagens dos métodos de eliminação de recursos incluem simplicidade e manutenção da interpretabilidade de suas variáveis.

Como desvantagem, no entanto, você não obtém informações dessas variáveis que eliminou. Se usarmos apenas o PIB do ano passado, a proporção da população em empregos na indústria e a taxa de desemprego para prever o PIB deste ano, estamos perdendo o que as variáveis descartadas possam contribuir para o nosso modelo. Ao eliminar os recursos, também eliminamos completamente todos os benefícios que essas variáveis descartadas trariam.

A extração de recursos, no entanto, não encontra esse problema. Digamos que temos dez variáveis independentes. Na extração de recursos, criamos dez variáveis independentes "novas", em que cada variável independente "nova" é uma combinação de cada uma das dez variáveis independentes "antigas". Entretanto, criamos essas novas variáveis independentes de uma maneira específica e ordenamos essas novas variáveis de acordo com o quão bem elas predizem nossa variável dependente.

Você pode dizer: “Onde a redução da dimensionalidade entra em jogo?” Bem, mantemos quantas das novas variáveis independentes desejamos, mas deixamos de lado as “menos importantes”. Como ordenamos as novas variáveis de acordo com a medida em que elas ajudam a prever nossa variável dependente, sabemos qual variável é a mais importante e a menos importante. Mas - e aqui está o pulo do gato - porque essas novas variáveis independentes são combinações das nossas antigas, ainda mantemos as partes mais valiosas de nossas variáveis antigas, mesmo quando descartamos uma ou mais dessas "novas" variáveis!

A análise de componentes principais é uma técnica para a extração de recursos - para combinar nossas variáveis de entrada de uma maneira específica, podemos eliminar as variáveis “menos importantes”, mantendo as partes mais valiosas de todas as variáveis! Como um benefício adicional, cada uma das "novas" variáveis após o PCA é independente uma da outra. Isso é um benefício, porque as suposições de um modelo linear exigem que nossas variáveis independentes sejam independentes uma da outra. Se decidirmos ajustar um modelo de regressão linear com essas "novas" variáveis, essa suposição será necessariamente satisfeita.