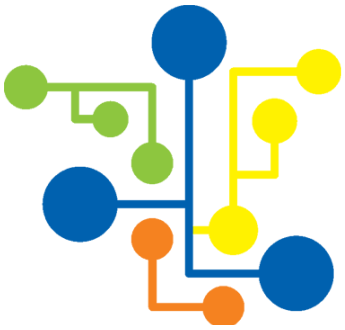

Data Science Academy

www.datascienceacademy.com.br

Matemática Para Machine Learning

PCA - Reduzir a Dimensão do Espaço de Recursos

Digamos que você queira prever qual será o produto interno bruto (PIB) do Brasil em um determinado ano. Você tem muitas informações disponíveis: o PIB do Brasil no primeiro trimestre, o PIB durante todo os anos anteriores e assim por diante. Você tem qualquer indicador econômico disponível ao público, como taxa de desemprego, taxa de inflação, etc...

Você tem dados do Censo do Brasil de 2010 estimando quantos brasileiros trabalham em cada setor e dados auxiliares atualizando essas estimativas entre cada censo. Você sabe quantos membros da Câmara e do Senado pertencem a cada partido político. Você pode reunir dados de preços de ações, o número de lançamentos de novas ações ocorridos na bolsa em um ano e a taxa de crescimento de cada empresa listada na bolsa. Apesar de ser um número esmagador de variáveis a serem consideradas, isso apenas arranha a superfície.

Se você já trabalhou com muitas variáveis antes, sabe que isso pode apresentar problemas. Você entende os relacionamentos entre cada variável? Você tem tantas variáveis que corre o risco de ajustar demais o modelo aos seus dados ou pode estar violando suposições de qualquer tática de modelagem que esteja usando?

Você pode fazer a pergunta: "Com todas as variáveis que coletei como concentro-me em apenas algumas delas?" Em termos técnicos, você deseja "reduzir a dimensão do espaço de recursos". Ao reduzir a dimensão do seu espaço de recursos, você tem menos relacionamentos entre variáveis a considerar e é menos provável que super ajuste seu modelo. Isso não significa imediatamente que o ajuste excessivo não é mais uma preocupação, mas estamos seguindo na direção certa! E seguimos essa direção usando PCA – Análise de Componentes Principais.

Continue acompanhando!