



Big Data Real-Time Analytics Com Python e Spark 3.0

Big Data Real-Time Analytics Com Python e Spark Versão 3.0

O Que é Feature Selection?

A seleção de recursos (feature selection) é o processo de isolar os recursos mais consistentes, não redundantes e relevantes a serem usados na construção de um modelo. Reduzir metódicamente o tamanho dos conjuntos de dados é importante, pois o tamanho e a variedade dos conjuntos de dados continuam a crescer. O principal objetivo da seleção de recursos é melhorar o desempenho de um modelo preditivo e reduzir o custo computacional da modelagem.

A seleção de recursos, um dos principais componentes da engenharia de recursos, é o processo de seleção dos recursos mais importantes a serem inseridos em algoritmos de aprendizado de máquina. Técnicas de seleção de recursos são empregadas para reduzir o número de variáveis de entrada, eliminando recursos redundantes ou irrelevantes e estreitando o conjunto de recursos para aqueles mais relevantes para o modelo de aprendizado de máquina.

Os principais benefícios de realizar a seleção de recursos com antecedência, em vez de deixar o modelo de aprendizado de máquina descobrir quais recursos são mais importantes, incluem:

- **Modelos mais simples:** modelos simples são fáceis de explicar - um modelo muito complexo e inexplicável não é valioso.
- **Tempos de treinamento mais curtos:** um subconjunto mais preciso de recursos diminui a quantidade de tempo necessária para treinar um modelo.
- **Redução de variância:** aumenta a precisão das estimativas que podem ser obtidas para uma determinada simulação
- **Evitar a “maldição” da alta dimensionalidade:** à medida que a dimensionalidade e o número de recursos aumentam, o volume de espaço aumenta tão rapidamente que os dados disponíveis se tornam limitados. Selecionar e reduzir o número de recursos evita que não tenhamos dados suficientes para o treinamento o modelo.