05

Double dipping

Se eu te mostro uma moeda e digo que vou jogá-la para cima, qual a chance de sair cara ou coroa? Assumimos que é 1/2 pois acreditamos que a moeda seja normal.

Mas se eu jogar a moeda 3 vezes para cima e todas as vezes der cara, você pode começar a suspeitar. Poderia até tentar argumentar: "suspeito que a moeda seja viciada para o lado da cara, ao invés de ser justa, olha só, os dados parecem indicar isso".

Na verdade 3 amostras é um número pequeno. Mas deixando esse ponto de lado, tem algo muito estranho ocorrendo. Olhamos os dados (3 vezes cara) e chegamos numa hipótese (ela está viciada para cara). Ai usamos **novamente** os dados das observações para testar nossa hipótese.

Esse ato é chamado de double dipping, quando você mergulha sua mão duas vezes no mesmo dado: uma para levantar uma hipótese e outra para validá-lo. E isso é perigoso, pois muitas vezes a hipótese vai se validar, uma vez que você usou os próprios dados para levantá-la.

Como evitar essa situação? Validamos a hipótese num conjunto diferente de dados de onde levantamos ela. Isto é, de antemão separamos nossos dados em mais de uma parte, levantamos hipóteses e modelos em uma parte e validamos em outra. Diversas dessas técnicas de validação são abordadas em nossos cursos de data science (regressão) e machine learning (validação).

Um [resumo do problema em outros contextos (https://en.wikipedia.org/wiki/Testing_hypotheses_suggested_by_the_data)](https://en.wikipedia.org/wiki/Testing_hypotheses_suggested_by_the_data) também pode ser encontrado na Wikipedia.