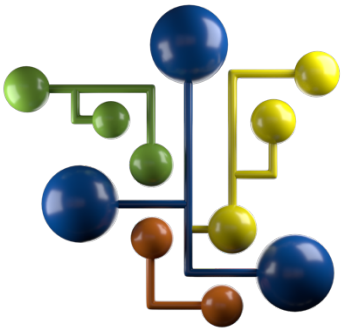


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Avaliação e Escolha da Melhor Hipótese

Queremos aprender uma hipótese que melhor se ajuste aos dados futuros. Para tornar isso preciso precisamos definir “dados futuros” e “melhor”. Façamos a suposição de estacionaridade: que há uma distribuição de probabilidade sobre exemplos que permanece estacionária ao longo do tempo. Cada exemplo de ponto de dados (antes de vê-lo) é uma variável aleatória Ej cujo valor observado ej = (xj, yj) é amostrado da distribuição e é independente dos exemplos anteriores:

$$\mathbf{P}(E_j|E_{j-1}, E_{j-2}, \ldots) = \mathbf{P}(E_j)$$

e cada exemplo tem uma distribuição de probabilidades anterior idêntica:

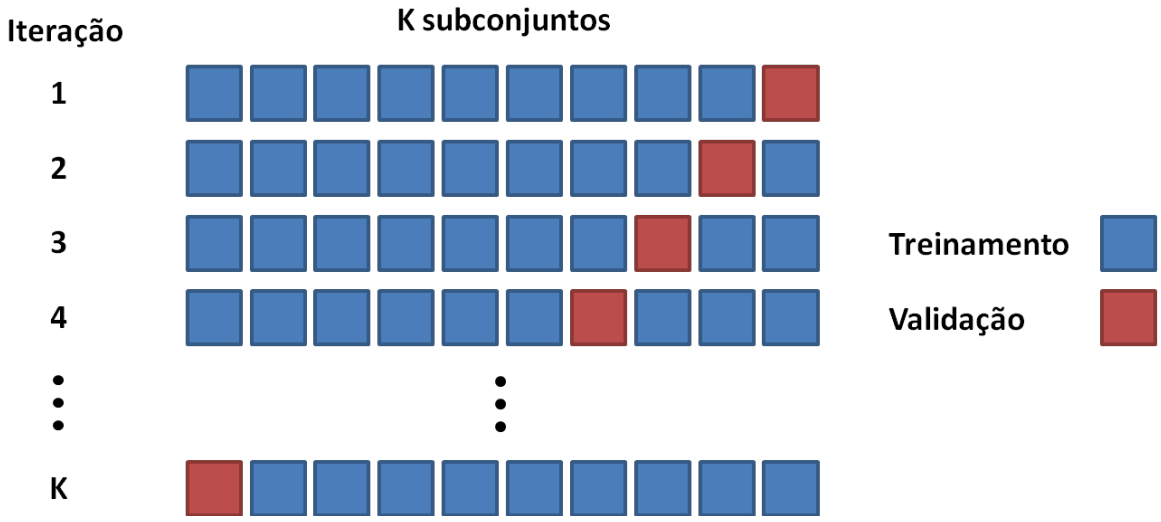
$$\mathbf{P}(E_j) = \mathbf{P}(E_{j-1}) = \mathbf{P}(E_{j-2}) = \cdots$$

Os exemplos que satisfazem essas suposições são chamados independentes e identicamente distribuídos. Essa suposição liga o passado ao futuro; sem tal conexão, todas as apostas estão fora — o futuro poderá ser qualquer coisa (veremos mais adiante que a aprendizagem ainda pode ocorrer se houver mudanças lentas na distribuição).

O próximo passo é definir o “melhor ajuste”. Definimos a taxa de erro de uma hipótese como a proporção de erros que ela comete — a proporção de vezes em que h(x) ≠ y para o exemplo (x, y). Agora, só porque uma hipótese h tem taxa de erro baixa no conjunto de treinamento não significa que ela generalize bem. Um professor sabe que um exame não vai avaliar com precisão os alunos se eles já viram as questões do exame. Da mesma forma, para obter avaliação precisa de uma hipótese, é preciso testá-la em um conjunto de exemplos que não tenham sido vistos ainda. A abordagem mais simples é a que já vimos: dividir aleatoriamente os dados disponíveis em um conjunto de treinamento a partir do qual o algoritmo de aprendizagem produz h e um conjunto de teste em que a precisão de h é avaliada. Esse método, chamado às vezes de validação cruzada por retenção, tem a desvantagem de não conseguir usar todos os dados disponíveis; se utilizarmos a metade dos dados para o conjunto de teste, estaremos treinando em apenas metade dos dados, e podemos ter uma hipótese mais fraca. Por outro lado, se reservamos apenas 10% dos dados para o conjunto de teste, podemos, por acaso estatístico, obter uma estimativa fraca da precisão real.

Podemos apertar mais os dados e ainda obter uma estimativa precisa usando uma técnica chamada validação cruzada com k-repetições. A ideia é que cada exemplo sirva duplamente — como dados de treinamento e dados de teste. Primeiro dividimos os dados em k subconjuntos iguais. Em seguida, realizamos k rodadas de aprendizagem; em cada rodada 1/k dos dados é retido como um conjunto de teste e os exemplos restantes são usados como dados

de treinamento. A pontuação média do conjunto de teste de k rodadas deve então ser uma estimativa melhor do que uma pontuação única. Os valores populares de k são 5 e 10 — o suficiente para dar uma estimativa que é estatisticamente provável que seja precisa, a um custo 5-10 vezes maior do tempo de computação. O extremo é k = n, também conhecido como validação cruzada com omissão de um.



Apesar dos melhores esforços dos metodólogos estatísticos, os usuários frequentemente invalidam seus resultados ao espreitar inadvertidamente os dados de teste. A espreita pode acontecer assim: um algoritmo de aprendizagem tem vários “botões” que podem ser fraudados para ajustar seu comportamento — por exemplo, vários critérios diferentes para escolher o próximo atributo de aprendizagem em árvore de decisão. O pesquisador gera hipóteses para várias configurações diferentes dos botões, as medidas de suas taxas de erro sobre o conjunto de teste e relatórios de taxa de erro das melhores hipóteses. Infelizmente, ocorreu a espreita! A razão é que a hipótese foi selecionada com base em sua taxa de erro de conjunto de teste; assim, a informação sobre o conjunto de teste vazou no algoritmo de aprendizagem.

Espreitar é uma consequência de uso do desempenho do conjunto de teste, tanto para escolher uma hipótese como para avaliá-la. A maneira de evitar isso é realmente reter o conjunto de teste — bloqueá-lo até que a aprendizagem esteja completa e se deseja simplesmente obter uma avaliação independente da hipótese final. (E, então, se você não gostar dos resultados… terá de obter, e bloquear, um conjunto de teste completamente novo se quiser voltar e encontrar uma hipótese melhor.) Se o conjunto de teste estiver bloqueado, mas você ainda quiser medir o desempenho dos dados não vistos, como forma de selecionar uma boa hipótese, divida os dados disponíveis (sem o conjunto de teste) em um conjunto de treinamento e em um conjunto de validação.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig