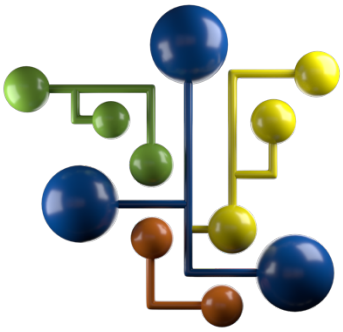


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Aprendizagem de Modelos Probabilísticos

Já sabemos sobre a prevalência da incerteza em ambientes reais. Os agentes podem manipular a incerteza usando os métodos de probabilidade e teoria da decisão, mas primeiro eles devem aprender suas teorias probabilísticas do mundo a partir da experiência. É possível fazê-lo, pela formulação da tarefa de aprendizagem em si como um processo de inferência probabilística. Veremos que uma visão bayesiana da aprendizagem é extremamente poderosa, fornecendo soluções gerais para os problemas de ruído, superadaptação e previsão ótima. Ele também leva em conta o fato de que um agente que não seja onisciente nunca poderá ter certeza sobre qual teoria do mundo é correta, mesmo que ainda tome decisões usando alguma teoria de mundo.

Os conceitos fundamentais aqui são dados e hipóteses (como em aprendizado de máquina em geral). Aqui, os dados são evidências, isto é, instanciações de algumas ou de todas as variáveis aleatórias que descrevem o domínio. As hipóteses são teorias probabilísticas de como o domínio funciona, incluindo teorias lógicas como caso especial.

Vamos considerar um exemplo simples. Nosso doce surpresa favorito tem dois sabores: cereja (hum) e lima (eca). O fabricante de doces tem um senso de humor peculiar e embrulha cada pedaço de doce na mesma embalagem opaca, independentemente do sabor. O doce é vendido em sacos muito grandes, dos quais existem cinco tipos conhecidos — novamente, indistinguíveis a partir do exterior:

h_1: 100% cereja
h_2: 75% cereja + 25% lima
h_3: 50% cereja + 50% lima
h_4: 25% cereja + 75% lima
h_5: 100% lima

Dado um novo saco de doces, a variável aleatória H (de hipótese) indica o tipo do saco, com valores possíveis h_1 até h_5. É evidente que H não é diretamente observável. À medida que os doces são abertos e inspecionados, são revelados os dados — D_1, D_2,…, D_n, onde cada D_i é uma variável aleatória com valores possíveis cereja e lima. A tarefa básica enfrentada pelo agente é prever o sabor do próximo doce. Apesar de sua trivialidade aparente, esse cenário serve para introduzir muitas questões importantes. O agente realmente precisa deduzir uma teoria de seu mundo, embora seja um mundo muito simples.

A aprendizagem bayesiana simplesmente calcula a probabilidade de cada hipótese, considerando-se os dados, e faz previsões de acordo com ela. Isto é, as previsões são feitas com o uso de todas as hipóteses, ponderadas por suas probabilidades, em vez de utilizar apenas uma única “melhor” hipótese. Desse modo, a aprendizagem é reduzida à inferência probabilística. Seja D a representação de todos os dados, com valor observado d; então, a probabilidade de cada hipótese é obtida pela regra de Bayes:

$$P(h_i \mid \mathbf{d}) = \alpha P(\mathbf{d} \mid h_i) P(h_i)$$

Agora, vamos supor que queremos fazer uma previsão sobre uma quantidade desconhecida X. Então, temos:

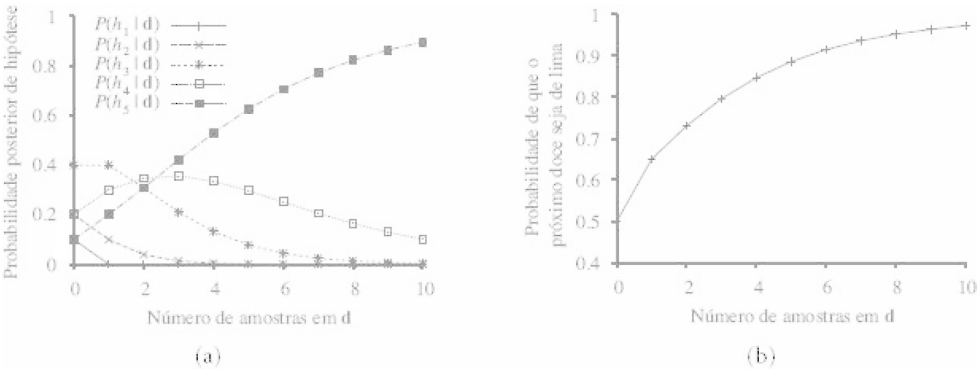
$$\mathbf{P}(X \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid \mathbf{d}, h_i)\mathbf{P}(h_i \mid \mathbf{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \mathbf{d})$$

onde pressupomos que cada hipótese determina uma distribuição de probabilidade sobre X. Essa equação mostra que as previsões são médias ponderadas sobre as previsões das hipóteses individuais. As hipóteses propriamente ditas são em essência "intermediários" entre os dados brutos e as previsões. As quantidades fundamentais na abordagem de Bayes são a hipótese a priori, $P(h_i)$ e a probabilidade dos dados sob cada hipótese, $P(d|h_i)$.

Para nosso exemplo dos doces, vamos supor por enquanto que a distribuição a priori sobre $h_1,\ldots, h_5$ seja dada por $\langle 0{,}1,\ 0{,}2,\ 0{,}4,\ 0{,}2,\ 0{,}1\rangle$, como anunciado pelo fabricante. A probabilidade dos dados é calculada sob a suposição de que as observações são i.i.d., de forma que:

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_i)$$

Por exemplo, suponha que o saco seja realmente um saco só com doces de lima (h_5) e que os primeiros 10 doces sejam todos de lima; então, $P(d \mid h_3)$ é $0{,}5_{10}$, porque metade dos doces em um saco do tipo h_3 é de doces de lima. A figura (a) abaixo mostra como as probabilidades posteriores das cinco hipóteses mudam à medida que a sequência de 10 doces de lima é observada. Note que as probabilidades começam com seus valores a priori e, portanto, h_3 é inicialmente a escolha mais provável e permanece assim depois que um doce de lima é desembrulhado. Depois de serem desembrulhados dois doces de lima, h_4 é mais provável; após três ou mais, h_5 (o temido saco só com doces de lima) é o mais provável. Depois de 10 doces seguidos, estamos bastante certos de nosso destino. A figura (b) abaixo mostra a probabilidade prevista de que o próximo doce seja de lima. Como seria de esperar, ela aumenta monotonicamente em direção a 1.



O exemplo mostra que a previsão bayesiana eventualmente concorda com a verdadeira hipótese. Isso é característico da aprendizagem bayesiana. Para qualquer probabilidade a priori fixa que não elimina a hipótese verdadeira, a probabilidade a posteriori de qualquer hipótese falsa vai, a partir de certo ponto, desaparecer sob determinadas condições técnicas. Isso simplesmente acontece porque a probabilidade de gerar dados "não característicos" indefinidamente é muitíssimo pequena. Mais importante ainda, a previsão bayesiana é ótima, quer o conjunto de dados seja pequeno, quer seja grande. Dada a hipótese a priori, qualquer outra previsão será correta com menor frequência.

É claro que o caráter ótimo da aprendizagem bayesiana tem um preço. Para problemas reais de aprendizagem, o espaço de hipóteses é em geral muito grande ou infinito. Em alguns casos, o somatório da equação (ou integração, no caso contínuo) pode ser executado de forma tratável, mas, na maioria dos casos, devemos recorrer a métodos aproximados ou simplificados.

Uma aproximação muito comum — habitualmente adotada na ciência — é fazer previsões com base em uma única hipótese mais provável, isto é, uma h_i que maximize $P(h_i \mid d)$. Com frequência, isso é chamado de hipótese de máximo a posteriori ou MAP. As previsões feitas de acordo com uma hipótese de MAP h_{MAP} são aproximadamente bayesianas até o ponto em que $P(X|d) \approx P(X|h_{MAP})$.

Em nosso exemplo de doces, $h_{MAP} = h_5$, após três doces de lima seguidos e, assim, o sistema de aprendizagem de MAP prevê que o quarto doce será de lima com probabilidade 1,0 — uma previsão muito mais perigosa que a previsão bayesiana de 0,8 mostrada na figura (b) acima. À medida que chegam mais dados, as previsões de MAP e Bayes ficam mais próximas porque os concorrentes da hipótese de MAP se tornam cada vez menos prováveis.

Embora nosso exemplo não mostre, a descoberta de hipóteses de MAP frequentemente é muito mais fácil que a aprendizagem bayesiana porque exige a resolução de um problema de otimização, em vez de um grande problema de somatório (ou integração).

Tanto na aprendizagem bayesiana quanto na aprendizagem de MAP, a hipótese a priori $P(h_i)$ desempenha uma função importante. A superadaptação (overfitting) pode ocorrer quando o espaço de hipóteses é muito expressivo, de forma que ele contenha muitas hipóteses que se

ajustam bem ao conjunto de dados. Em vez de impor um limite arbitrário sobre as hipóteses a serem consideradas, os métodos de aprendizagem bayesiana e MAP utilizam a hipótese a priori para penalizar a complexidade. Em geral, as hipóteses mais complexas têm uma probabilidade a priori mais baixa — em parte porque normalmente existem muito mais hipóteses complexas que hipóteses simples. Por outro lado, as hipóteses mais complexas têm capacidade maior de se adaptar aos dados (no caso extremo, uma tabela de busca pode reproduzir os dados exatamente com probabilidade 1). Consequentemente, a hipótese a priori incorpora um compromisso entre a complexidade de uma hipótese e seu grau de adaptação aos dados.

Uma simplificação final é fornecida supondo-se uma probabilidade a priori uniforme sobre o espaço de hipóteses. Nesse caso, a aprendizagem de MAP se reduz à escolha de um h_i que maximize $P(d|h_i)$. Isso é chamado de hipótese de máxima probabilidade (MP), h_{MP}. A aprendizagem de máxima probabilidade é muito comum em estatística, uma disciplina na qual muitos pesquisadores desconfiam da natureza subjetiva de hipóteses a priori. É uma abordagem razoável quando não existe nenhuma razão para preferir uma hipótese sobre outra a priori — por exemplo, quando todas as hipóteses são igualmente complexas. Ela proporciona uma boa aproximação para a aprendizagem bayesiana e de MAP quando o conjunto de dados é grande, porque os dados inundam a distribuição a priori sobre hipóteses, mas tem problemas com conjuntos de dados pequenos.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig