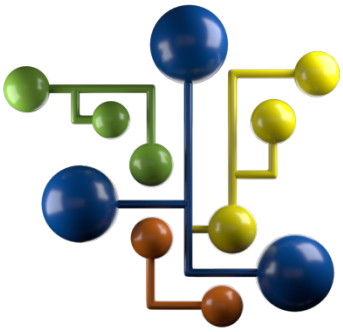


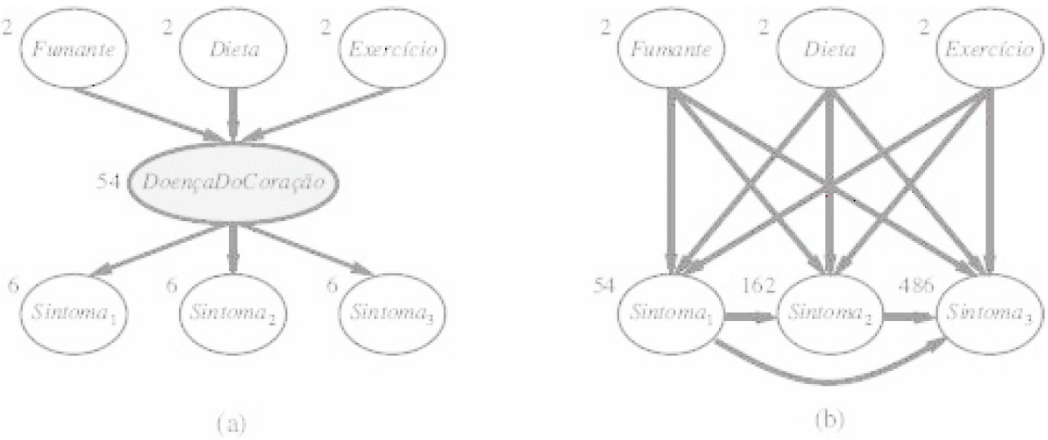
Data Science Academy

www.datascienceacademy.com.br

# Introdução à Inteligência Artificial

# Aprendizagem com Variáveis Ocultas
# Algoritmo Expectation Maximization

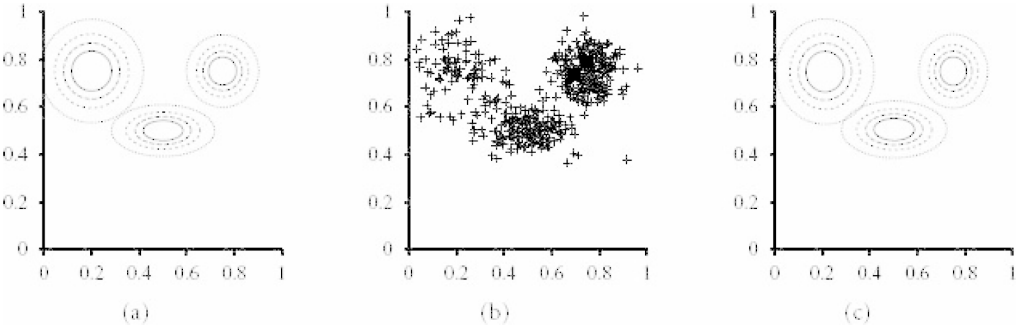
Muitos problemas reais têm variáveis ocultas (às vezes chamadas variáveis latentes) que não são observáveis nos dados disponíveis para aprendizagem. Por exemplo, registros médicos com frequência incluem os sintomas observados, o tratamento aplicado e, talvez, o resultado do tratamento, mas raramente contêm uma observação direta da própria doença! (Observe que o diagnóstico não é a doença; é uma consequência causal dos sintomas observados que, por sua vez, são causados pela doença.) Você poderia perguntar: “Se a doença não é observada, por que não construir um modelo sem ela?” A resposta aparece na figura abaixo, que mostra um pequeno modelo de diagnóstico fictício para doenças do coração. Existem três fatores de predisposição observáveis e três sintomas observáveis. Suponha que cada variável tenha três valores possíveis (por exemplo, nenhum, moderado e severo). A remoção da variável oculta a partir da rede em (a) produz a rede em (b); o número total de parâmetros aumenta de 78 para 708. Desse modo, variáveis latentes podem reduzir drasticamente o número de parâmetros exigidos para especificar uma rede bayesiana. Por sua vez, isso pode reduzir drasticamente a quantidade de dados necessários para se aprender os parâmetros.




As variáveis ocultas são importantes, mas complicam o problema de aprendizagem. Por exemplo, na figura (a), não é óbvia a maneira de aprender a distribuição condicional para DoençaDoCoração, dados seus pais, porque não conhecemos o valor de DoençaDoCoração em cada caso; o mesmo problema surge na aprendizagem das distribuições correspondentes aos sintomas. Esta seção descreve um algoritmo chamado maximização de expectativa (ou EM, Expectation Maximization), que resolve esse problema de modo muito geral. Mostraremos três exemplos e depois forneceremos uma descrição geral. A princípio, o algoritmo parece mágica; porém, uma vez desenvolvida a intuição, podemos encontrar aplicações para EM em enorme variedade de problemas de aprendizagem.

A formação não supervisionada de agrupamentos é o problema de distinguir várias categorias em uma coleção de objetos. O problema é não supervisionado porque os rótulos de categorias não são dados. Por exemplo, vamos supor que registramos o espectro de centenas de milhares de estrelas; existem diferentes tipos de estrelas revelados pelo espectro? Nesse caso, quantos tipos e quais são suas características? Todos nós estamos familiarizados com expressões como “gigante vermelha” e “anã branca”, mas as estrelas não têm esses rótulos para identificá-las — os astrônomos tiveram de executar a formação de agrupamentos não supervisionados para identificar essas categorias. Outros exemplos incluem a identificação de espécies, gêneros, ordens, e assim por diante, na taxonomia de organismos e a criação de espécies naturais para categorizar objetos comuns.

A formação de agrupamentos não supervisionados começa com dados. A figura (b) abaixo mostra 500 pontos de dados, cada um dos quais especifica os valores de dois atributos contínuos.




Os pontos de dados poderiam corresponder a estrelas, e os atributos poderiam corresponder a intensidades espectrais em duas frequências específicas. Em seguida, precisamos compreender que espécie de distribuição de probabilidade poderia ter gerado os dados. A formação de agrupamentos pressupõe que os dados são gerados a partir de uma distribuição de mistura P. Tal distribuição tem k componentes, cada um dos quais é por si só uma distribuição. Um ponto de dados é gerado escolhendo-se primeiro um componente e depois gerando-se uma amostra a partir desse componente. Seja a variável aleatória C que indica o componente, com valores 1,…, k; então, a distribuição de mistura é dada por:

$$P(\mathbf{x}) = \sum_{i=1}^{k} P(C=i)\, P(\mathbf{x} \mid C=i)$$

onde x se refere aos valores dos atributos para um ponto de dados. No caso de dados contínuos, uma escolha natural para as distribuições de componentes é a gaussiana multivariada, que fornece a família de distribuições chamada mistura de distribuições gaussianas. Os parâmetros de uma mistura de distribuições gaussianas são $w_i = P(C = i)$ (o peso de cada componente), $\mu_i$ (a média de cada componente) e $\Sigma_i$ (a covariância de cada componente). A figura (a) mostra uma mistura de três

gaussianos; essa mistura é de fato a origem dos dados contidos em (b), assim como o modelo mostrado na figura (a).

Então, o problema de formação não supervisionada de agrupamentos consiste em recuperar um modelo de mistura como o da figura (b) a partir de dados brutos como os da figura (a). É claro que, se soubéssemos que componente gerou cada ponto de dados, seria fácil recuperar os componentes gaussianos: poderíamos simplesmente selecionar todos os pontos de dados a partir de um dado componente e depois aplicar (em uma versão multivariada) a equação para ajustar os parâmetros de um gaussiano a um conjunto de dados. Por outro lado, se os parâmetros de cada componente fossem conhecidos, poderíamos, pelo menos em um sentido probabilístico, atribuir cada ponto de dados a um componente. O problema é que não conhecemos nem as atribuições nem os parâmetros.

A ideia básica de EM nesse contexto é fingir que conhecemos os parâmetros do modelo e depois deduzir a probabilidade de cada ponto de dados pertencer a cada componente. Depois disso, readaptamos os componentes aos dados, onde cada componente é ajustado ao conjunto de dados inteiro, com cada ponto ponderado pela probabilidade de pertencer a esse componente. O processo iterage até a convergência. Essencialmente, estamos “completando” os dados, deduzindo distribuições de probabilidades sobre as variáveis ocultas — o componente ao qual pertence cada ponto de dados — com base no modelo atual. Para a mistura de distribuições gaussianas, inicializamos arbitrariamente os parâmetros do modelo de mistura e depois repetimos as duas etapas a seguir:

1. Etapa E (Expectation): Calcular as probabilidades $p_{ij} = P(C= i \mid x_j)$, a probabilidade de que o dado $x_j$ tenha sido gerado pelo componente i. Pela regra de Bayes, temos $p_{ij} = \alpha P(x_j \mid C = i)P(C = i)$. O termo $P(x_j \mid C = i)$ é simplesmente a probabilidade em $x_j$ do i-ésimo gaussiano, e o termo $P(C =i)$ é o parâmetro que representa o peso para o i-ésimo gaussiano. Definir $n_i = \Sigma_j p_{ij}$.

2. Etapa M (Maximization): Calcular a nova média, a covariância e os pesos de componentes, usando as etapas seguintes em sequência:

$$\mu_i \leftarrow \sum_j p_{ij}\mathbf{x}_j / n_i$$
$$\Sigma_i \leftarrow \sum_j p_{ij}(\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^\top / n_i$$
$$w_i \leftarrow n_i / N$$

onde N é o número total de pontos de dados. A etapa E, pode ser visualizada como o cálculo dos valores esperados $p_{ij}$ das variáveis indicadoras ocultas $Z_{ij}$, onde $Z_{ij}$ é 1 se o dado $x_j$ foi gerado pelo i-ésimo componente e 0 em caso contrário. A etapa M, ou etapa de maximização, encontra os novos valores dos parâmetros que maximizam a probabilidade logarítimica dos dados, dados os valores esperados das variáveis indicadoras ocultas.

Existem dois pontos a serem observados. Primeiro, a probabilidade logarítmica para o modelo aprendido final excede ligeiramente a do modelo original, a partir do qual os dados foram gerados. Isso poderia parecer surpreendente, mas, na verdade, simplesmente reflete o fato de que os dados foram gerados ao acaso e não poderiam fornecer um reflexo exato do modelo subjacente. O segundo ponto é que EM aumenta a probabilidade logarítmica dos dados em cada iteração. Esse fato pode ser provado no caso geral. Além disso, sob certas condições (válidas na maioria dos casos), pode-se provar que EM alcança um máximo local de probabilidade (em casos raros, ele pode alcançar um ponto de sela ou até um mínimo local). Nesse sentido, EM é semelhante a um algoritmo de subida de encosta baseado em gradiente, mas que não tem nenhum parâmetro "tamanho do passo".

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig