DCE 22 [redacted] ___/___/___

⇒ 08h11
⇒ 09h11
⇒ 12h11

Aprendizado de Máquina Não-Supervisionado

1) Introdução

A distinção entre aprendizado de máquina com e sem supervisão, é feita pela ausência ou presença da possibilidade de que durante o processo de treinamento de um dado modelo, esse possa receber um sinal que supervisiona ou "guia" esse treinamento. Sem ainda entrar em detalhes técnicos, de forma genérica, essa supervisão pode ser a existência de classes ou labels associado a cada amostra, e o uso dessas classes durante o treinamento de um modelo (nesse caso, um modelo supervisionado). Por exemplo, em um problema de reconhecimento facial, para cada imagem do domínio (por exemplo, fotos no formato de documentos de identidade) teríamos um [redacted] identificador da classe (podendo ser um identificador granular, como o nome da [redacted] pessoa, ou algo mais agrupado, como se está sorrindo ou outras classes refletindo emoções).

Na ausência da informação necessária para essa supervisão, automaticamente [redacted] se cria uma limitação no tipo de objetivos que podemos

induzir o modelo a seguir no processo de treinamento. De modo que poderemos explorar aspectos como redução de dimensionalidade, agrupamento de dados (ou clustering), estimação de densidade (ou representações da distribuição empírica dos dados) e aprendizado ou estimação de representações simplificadas ou ajustadas para diferentes tarefas do conjunto de dados ou de uma estrutura latente ou não-observada (que se assume existir nesse conjunto de dados).

De certa forma, a definição dessa família de métodos e modelos de aprendizado de máquina pela "ausência" de supervisão gera uma grande diversidade de metodologias e algoritmos, que será impossível cobrir de forma extensiva. Na nossa tentativa iremos seguir a seguinte estratégia:

a) Definir objetivos genéricos presentes nas diversas metologias de aprendizado não-supervisionado; Relacionando com certas tarefas práticas.

b) Descrever e listar de forma breve os métodos mais comuns para certos objetivos e tarefas.

c) Descrever com detalhes a modelagem e algoritmos de treinamento dos seguintes métodos:

2) Objetivos e tarefas associadas ao Aprendizado de Máquina não-supervisionado.

Antes de descrever esses objetivos e tarefas, faremos um caveat lector, certos objetivos são em certa medida equivalente, mas foram desenvolvidos de forma independente por sub-comunidades distintas de pesquisa dentro da computação, estatística, engenharia, física ou matemática. Mantendo dessa forma um grau de distinção devido as especificidades dos problemas que cada comunidade procura resolver. Por exemplo, redução de dimensionalidade e compressão são objetivos similares quando tratamos de dados numéricos multidimensionais, apesar de compressão ser um objetivo também válido para dados não numéricos, como texto ou árvores.

2.a) Agrupamento ou clustering.

Se tratando de métodos não-supervisionados um dos grandes objetivos que coincide com uma tarefa prática é agrupar o conjunto de dados, sem ter uma informação inicial de que grupos existem nos dados.

Esse objetivo aparece de forma prática como tarefa ou sub-tarefa na resolução de diversos problemas, desde visão computacional e processamento de imagens, a processamento de linguagem natural ou aplicações em ciência, engenharia, finanças ou saúde. De forma analógica, é também uma

(04) DCE22

__/__/__

função cognitiva presente no pensamento e capacidade humana e, por inferência e ciência empírica, possivelmente em diversos animais. De forma sucinta, agrupar dados é um "módulo" sem dúvida essencial em sistemas ou seres com algum grau de inteligência.

Portanto, e talvez por ser tão essencial, muitas vezes aprendizado não-supervisionado é visto como sinônimo de agrupamento de dados ou clustering. Não adotamos esse nível de reducionismo, e vamos listar outros objetivos e tarefas, embora assinale uma conexão entre todos métodos e clustering (por exemplo, certos métodos de redução de dimensionalidade podem ser vistos como soft clustering, ou agrupamento com pertencimento parcial).

Matematicamente dado um conjunto $X = \{x_1, \ldots, x_n\}$ de dados, assumimos a existência de grupos C, e uma função $\pi: X \mapsto C$, que atribua para cada ponto $x_i \in X$ qual grupo $c_j \in C$, essa amostra dos dados pertence, usando apenas a informação presente no conjunto de dados X. O conjunto dos grupos C é no geral contável, mas tipicamente é finito e tem menor tamanho que o conjunto de amostras.

Um exemplo de tarefa seria dado o conjunto de imagens de dígitos do MNIST, sem utilizar a informação que associa cada imagem a um dígito de 0 a 9, procurar agrupar as imagens em 9 conjuntos distintos. Aplicando um dom método

(Continuação de 01-04) ___/___/___

de agrupamento de dados nessa tarefa poderia resultar em grupos que coincidem mais ou menos com os dígitos corretos (ou alternativamente, imagens de um mesmo dígito tendem a ser parecidas e serem associadas ao mesmo grupo), embora, devido à não utilização da informação de classe (label/rótulo do dígito), não podemos esperar o mesmo nível de acurácia de um aprendizado supervisionado, nem que o método saiba qual dígito associar a cada grupo.

Dentro desse mesmo objetivo, existe também grande variedade de metodologias e algoritmos de aprendizado conforme adicionarmos mais hipóteses sobre correlação e dependência entre os grupos, bem como adicionarmos outras hipóteses sobre tipo de domínio dos dados e a estrutura de dependência e correlação entre eles.

2.a.1) Alguns métodos:

- K-médias ou K-means: procuramos nesse métodos encontrar K grupos no conjunto de dados numéricos (e vetoriais). Para cada grupo assumimos que existe uma representação ótima dado pela média dos pontos pertencentes a esse grupo.

- K-médoides (incluindo K-medianas e K-moda): generalização do K-médias para casos onde a soma de elementos do domínio não seja definida, mas existe noções possíveis de frequência ou probabilidade de ocorrência, e uma métrica

(06) DCE 22

__/__/__

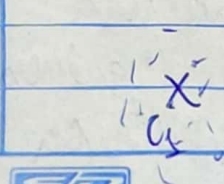
ou distância definida entre amostras ou pontos do domínio. De modo que é possível iterativamente

a) Calcular e re-calcular um "medoide", sendo ponto representativo de um dado sub-conjunto, utilizando critérios como mediana ou moda estimados. Ou seja, $\forall c_j \in C : \exists\, m(c_j) \in X$ ($m(c_j)$ é o medoide)

b) Atribuir pontos a um "medoide" via minimização de distância: $\underset{m(c_j) \in C}{\operatorname{argmin}} \sum d(x_i, m(c_j))$

- Modelos de tópicos, em particular o LDA (Latent Dirichlet Allocation): específico para texto (ou outros dados categóricos), esse método trata uma amostra como um conjunto finito de palavras, onde cada palavra é escolhida de um dado tópico. É um método probabilístico, que consiste em tratar uma amostra para obter estimativa de probabilidades para cada tópico. Um tópico, nada mais é que uma distribuição de probabilidade sobre palavras, ou seja, um grupo de probabilidades. O modelo de tópico aprende esses grupos de distribuições. Por se tratar de grupos probabilísticos pode ser entendido como soft-clustering.

- Modelos de misturas (Mixture de Experts, mistura de Gaussianas, etc): Modelos probabilísticos que assumem que cada observação é uma soma ponderada (ou outras formas de agregação) de "observações latentes" obtidas de K distribuições.



Por exemplo, na figura ao lado diríamos que $X_1 = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3$, com C_1, C_2, C_3 obtidos de uma Gaussiana.

Usando a notação $\mathcal{C}$ (conjunto de grupos), $P_{\mathcal{C}}$ para distribuição de grupos, e $m(c_j)$ ($c_j \in \mathcal{C}$) sendo um ponto central do conjunto de grupos, e $\sim$ para representar geração de amostras de uma distribuição, temos:

$$\mathcal{C} \sim P_{\mathcal{C}}$$

$$\Leftrightarrow \{c_1, \ldots, c_M\} \sim P_{\mathcal{C}}$$

$$\Rightarrow x_i = \sum_{j=1}^{M} \alpha_{ij}\, m(c_j) \in X$$

Tipicamente o treinamento se trata de algoritmos para estimar α_{ij}, $P_{\mathcal{C}}$ e os pontos representativos dos grupos, no caso de mistura de gaussianas, para cada grupo uma média μ_k e matriz de correlação Σ_k.

- Aprendizado de estruturas hierárquicas: alguns métodos assumem uma estrutura hierárquica ou de árvore entre as variáveis ou entre os clusteres (agrupamentos). Em bio-informática os métodos de inferência de estrutura filogenética não-supervisionado são exemplos dessa categoria. No geral nesse caso, partindo de sequências de RNA (strings com 4 letras mas tamanho variável e longo), se estima uma possível estrutura de antepassados comuns via trechos compartilhados de sub-strings, gerando uma árvore de grupos e sub-grupos. Para dados numéricos, métodos que assumem um espaço hiperbólico (Embeddings de Poincaré) podem atingir resultados similares de aprender uma hierarquia de grupos.

(08) DCE 22

/ /

2. b) Redução de dimensionalidade e compressão

Um segundo objetivo comum em aprendizado não-supervisionado é a redução de dimensionalidade e compressão (no sentido de teoria de informação, diminuindo a quantidade de bits no dataset original).

No caso de redução de dimensionalidade, majoritariamente aparece na formulação do conjunto de dados como solução de um sistema de equações (lineares, não lineares, diferenciais ordinárias etc) com certo grau de sobredeterminação, sugerindo que o conjunto, e a modelagem do problema, tem mais graus de liberdade que necessário. Isso acontece nas versões mais simples de métodos de componentes principais e de valor singular (PCA e SVD), que são comumente utilizados também para redução de ruídos num conjunto de dados. No caso do SVD cada amostra é representada por soma ponderada de vetor ortonormais e um valor singular σ_k, e selecionando os K maiores valores singulares reduzimos um conjunto de dimensão d para K, com $K < d$. Esse processo é ótimo no sentido de minimizar o erro introduzido por essa redução em termos da norma de Frobenius. Ao aplicar esse método estamos também comprimindo a amostra por precisar de menos números para representar o mesmo conjunto ($N \times d$ para $2N \times K + K^2$).

De forma mais genérica, e inspirado pela literatura de teoria de informação, podemos fazer uma simplificação a mais

Os métodos de redução de dimensionalidade podem ser vistos pela combinação de um codificador (encoder) e decodificador (decoder), com a restrição que o domínio onde se faz o encoding ou codificação das amostras iniciais é mais simples, tem menor dimensionalidade, maior compressão, ou ganho de explicabilidade em algum grau. Esses encoder ϕ_{enc} e decoder ϕ_{dec} são treináveis e o geral se minimiza algum erro de reconstrução. Formalmente temos dados iniciais $X = \{x_1, \ldots, x_n\}$, codificador $\phi_{enc}: X \mapsto Y$, decodificador $\phi_{dec}: Y \mapsto X$ e objetivo

$$\min_{\phi_{enc}, \phi_{dec}} L(X) = \sum_{x \in X} erro(x, \underbrace{\phi_{dec}(\phi_{enc}(x))}_{\tilde{x} \Rightarrow \text{reconstrução de } x})$$

Por exemplo, no PCA, assumindo de forma matricial $\tilde{X} = \bar{\Phi} X$, temos $\bar{\Phi}$ uma matriz, portanto um encoder e decoder linear. A solução com a pseudo-inversa $(X^T)(XX^T)^{-1}X$ explicita essa composição.

Um caso mais moderno são os denoising auto-encoder e variational auto-encoders que representam métodos usando essa mesma formulação, mas com redes neurais fazendo o papel de encoder e decoder. O treinamento nesses casos é feito por descida de gradiente sobre os parâmetros da rede neural e induz termos de regularização e objetivos mais complexos como a verissimilhança dos dados.

(19) DCE22

___/___/___

2.g) Estimação de densidade

Outra grande categoria de tarefas e objetivo do aprendizado não-supervisionado é a estimação de densidade da distribuição empírica dos dados. Dependendo da aplicação dessa densidade estimada, a ênfase do treinamento pode ser aprender representações boas para a distribuição, ou capturar a diversidade de variação nos dados.

Temos nessa categoria métodos com redes generativas adversais (GAN), auto-encoders variacionais (VAE), modelos de difusão (Diffusion models), redes de fluxor normalizantes (normalizing flows) e redes generativas de fluxos (Generative Normalizing flows).

3) Breve descrição de alguns métodos e modelos

3.a) Notação

- Dados: $X = \{x_1, \dots, x_n\}$
- Grupos: $C = \{C_1, \dots, C_K\}$
- Centróides ou representação do grupo: $m: C \mapsto X$
- Função de pertencimento: $\pi: X \mapsto C$
- Variável Indicadora de grupo

(1) $z_{ij} = \begin{cases} 1 & \text{se } \pi(x_i) = C_j, \text{ ponto } x_i \text{ pertence a grupo } C_j \\ 0 & \text{caso contrário} \end{cases}$

~~Objetivo matemático na modelagem~~

3. b) K-means ou K-média

- Modelagem: chamamos de μ_j cada centroide $m(C_j)$, e assumimos que a média dos elementos desse grupo

$$\Rightarrow m(C_j) = \mu_j = \frac{1}{k_j} \sum_{x_i \in C_j} x_i = \frac{\sum_{i=1}^{n} z_{ij} x_i}{\sum_{i=1}^{n} z_{ij}}$$

- Cada ponto é representado pelo seu centroide, de modo que queremos atribuir a cada ponto o centroide com menor distância desse ponto

$$\underset{\mu_j}{\operatorname{argmin}} \sum_{\substack{c_j \in C \\ x \in X}} d(\mu_j, x)$$

- Combinando ambos obtemos o objetivo

$$L = L(\mu_1, \dots, \mu_K, z_{11}, \dots, z_{nK}) = \sum_{\substack{c_j \in C \\ x_i \in X}} z_{ij} \sum_{k=1}^{d} (x_{ik} - \mu_{jk})^2$$

$$\Rightarrow \frac{\partial L}{\partial \mu_{jk'}} = -2 \sum_{x_i \in X} z_{ij} (x_{ik'} - \mu_{jk'}) = 0$$

$$\Rightarrow \sum_{x_i \in X} z_{ij} x_{ik'} = \left(\sum_{x_i \in X} z_{ij} \right) \mu_{jk'}$$

(12) DCE22

__/__/____

Que demonstra que

(2) $\mu_j = \frac{\sum_{i=1}^{n} z_{ij} x_i}{\sum_{i=1}^{n} z_{ij}}$ são ótimos em relação ao objetivo.

Por outro lado, ao recalcular z_{ij} usando

(3) $z_{ij} = \begin{cases} 1 \text{ se } \mu_j = \underset{\mu_1, \mu_2, \ldots, \mu_K}{\operatorname{argmin}} d(\mu_j, x_i) \\ 0 \text{ caso contrário.} \end{cases}$

Induzimos possíveis mudanças no conjunto de centróides, o Algoritmo de treinamento portanto é

- Passo 1: Inicializa os centroide $\{\mu_1 \ldots \mu_K\}$ (escolha randômica, ou algo mais avançado com Kmeans++)
- Passo 2: Até convergência (ou seja até nada mudar mais) Repita: Eq (2) e Eq (3)