Biomatemática ***24*** *(2014), 101–118* ***ISSN 1679-365X***
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Uma visão geral de algumas abordagens para processamento de dados

Estevão L. Esmi[1], Marcos E. Valle[2]
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

Resumo. Nesse artigo, apresentamos três situações encontradas em muitas aplicações, incluindo a área de biomatemática, que requerem processamento de dados. Especificamente, discutimos algumas abordagens usadas em problemas de regressão, classificação e agrupamento. Não são apresentados detalhes das técnicas abordadas mas apenas uma visão geral de como os problemas são tratados. Com efeito, espera-se fornecer orientações gerais para um leitor da biomatemática leigo no assunto.

Palavras-chave*: Processamento de dados, aprendizado de máquina, regressão, classificação, agrupamento.*

1. Introdução

Na literatura encontramos uma vasta coleção de métodos para análise e processamento de dados distribuídos em diferentes áreas de pesquisas, tais como probabilidade, aprendizagem de máquina, inteligência computacional, mineração de dados, etc. De maneira geral, podemos dizer que a tarefa básica executada por estes métodos consiste em extrair informação ou padrões de um certo conjunto de dados que possam ser utilizados para inferir ou categorizar um determinado fenômeno ou evento. A formulação matemática para tais problemas depende tanto das características dos dados disponíveis quanto do tipo de sistema de informação que se deseja obter. Vejamos a seguir três situações que requerem uma abordagem de processamento de dados acompanhadas de exemplos que consideramos pertinentes para a área de biomatemática.

[1]eelaureano@gmail.com
[2]valle@ime.unicamp.br

Exemplo 1 (Crescimento Populacional Discreto com Atraso)**.** O crescimento de muitas espécies que requerem um tempo para maturidade sexual pode ser descrito em tempos discretos com atraso (Murray, 2002). Nesses casos, encontramos uma equação da forma

$$N_{t+1} = f(N_t, N_{t-\tau}), \tag{1.1}$$

em que N_t representa a densidade população no instante de tempo discreto t, τ representa o atraso (por exemplo, para a maturidade) e f é uma função geralmente não-linear. Do ponto de vista prático, se conhecemos a função f, podemos usar os valores $N_t, N_{t-\tau}$ para determinar N_{t+1}. Além disso, conhecendo $N_t, N_{t-1}, \ldots, N_{t-\tau}$, podemos também obter as populações subsequentes simplesmente aplicando (1.1) repetidas vezes. Em termos gerais, a tarefa de modelar consiste em determinar a forma de f que resume fatos sobre a espécie estudada e reflete observações prévias da densidade populacional da mesma. Contudo, podemos ter uma quantidade significativa de observações e pouca (ou nenhuma) informação sobre a espécie. Esse segundo caso pode ser visto como um exemplo da seguinte situação.

Situação 1 (Problema de Regressão)**.** *Dado um conjunto finito de pares ordenados* $\{(x_i, y_i) \in X \times Y : i = 1, \ldots, n\}$*, em que* X *e* Y *são dois universos arbitrários, deseja-se determinar uma função* $f : X \to Y$ *tal que* $f(x_i) \approx y_i$ *para todo* $i = 1, \ldots, n$*.*

A notação $f(x_i) \approx y_i$, que pode ser interpretada como $f(x_i)$ é próximo de y_i, é formalmente definida considerando um certo critério estabelecido sob o universo Y. Tal critério depende do contexto no qual a situação está inserida ou das hipóteses que são impostas sobre os dados $\{(x_i, y_i) \in X \times Y : i = 1, \ldots, n\}$.

Por exemplo, como a densidade populacional de uma determinada espécie em um instante t é dado por um número real não negativo, a função f que descreve o crescimento populacional discreto com atraso τ do Exemplo 1 corresponde a uma função de $X = \mathbb{R}^2_{\geq 0}$ para $Y = \mathbb{R}_{\geq 0}$, em que $\mathbb{R}_{\geq 0} = [0, +\infty)$ denota o conjunto dos números reais não-negativos. No contexto da Situação 1, com base num conjunto de observações $\{N_0, N_1, \ldots, N_n\}$, com $n > \tau$, deseja-se determinar uma função f tal que $f(N_t, N_{t-\tau}) \approx N_{t+1}$ para todo $t = \tau, \tau+1, \ldots, n-1$. A noção $f(N_t, N_{t-\tau}) \approx N_{t+1}$ pode ser representada, por exemplo, impondo que f minimiza o *erro quadrático médio* dado por

$$E(f) = \frac{1}{n-\tau} \sum_{t=\tau}^{n-1} \big(f(N_t, N_{t-\tau}) - N_{t+1}\big)^2. \tag{1.2}$$

O seguinte exemplo, que está disponível na base de dados de aprendizado de máquinas da *Universidade da Califórnia, Irvine - EUA* (Bache e Lichman, 2013), também refere-se à Situação 1.

Exemplo 2 (Estimativa do Número de Crimes Violentos)**.** Informações sócio-econômicas e policiais foram combinadas com o número de crimes por habitantes em 1994 municípios norte-americanos. Especificamente, foram consideradas 127 variáveis independentes contendo informações sociais, como número de habitantes na região urbana do município, econômicas, como a renda média familiar, e informações como número de policiais e o número de tipos diferentes de drogas apreendidos no município. Além disso, para cada município, foi coletado também o número de crimes violentos, incluindo assassinato, estrupo, roubo e assalto, por número de habitantes. As 127 variáveis independentes foram arranjadas em vetores coluna $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{127i}]^T \in \mathbb{R}^{127}$ e o número de crimes foi denotado por y_i, para cada $i = 1, \ldots, 1994$. O objetivo é determinar, com base nos conjuntos $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, 1994\}$ de dados amostrados, uma função $f : \mathbb{R}^{127} \to \mathbb{R}$ tal que $y_i \approx f(x_i)$, para todo $i = 1, \ldots, 1994$. Aqui, a notação "$\approx$" reflete o fato dos dados estarem sujeitos a erros. Com efeito, pode haver uma certa controvérsia com respeito à definição de crime violento. Portanto, o número de crimes violentos *per capita* podem conter erros.

O seguinte exemplo, que pode ser encontrado em (Bache e Lichman, 2013; Rocha Neto et al., 2011), ilustra outra situação frequente:

Exemplo 3 (Classificação da Coluna Vertebral)**.** A coluna vertebral é composta por um grupo de vértebras, discos, nervos, músculos, medula e junções. Além da proteção de órgãos internos, a coluna vertebral possui papeis importantes como suporte e movimento do corpo humano. Disfunções na coluna podem causar dores de diferentes intensidades. Hérnia de disco e espondilolistese, que refere-se ao deslocamento de uma vértebra com relação à vértebra anterior ou à coluna, são exemplos de patologias que causam dores intensas. Com objetivo de auxiliar no diagnóstico médico dessas duas disfunções da coluna vertebral, foram coletados de 310 pacientes seis atributos como ângulo da lordose lombar, inclinação sacral e ângulo de incidência pélvica (Rocha Neto et al., 2011). Além dos 6 atributos biomecânicos, cada um dos 310 pacientes foi diagnosticado como tendo hérnia de disco, espondilolistese ou nenhuma patologia. O objetivo é encontrar um sistema capaz identificar se um novo paciente possui ou não hérnia de disco ou espondilolistese com base nos seis atributos

biomecânicos e nos dados coletados anteriormente. De um modo geral, tem-se a seguinte situação.

Situação 2 (Problema de Classificação)**.** *Dado um conjunto* $\{(x_i, y_i) \in X \times Y : i = 1, \ldots, n\}$*, em que* X *é um universo arbitrário e* $Y = \{\ell_1, \ldots, \ell_L\}$ *é um conjunto finito, deseja-se determinar* $f : X \to Y$ *tal que* $f(x_i) = y_i, \forall i = 1, \ldots, n$*.*

A função $f : X \to Y$ apresentada na Situação 2 é chamada *classificador*, X é o *conjunto das características* e $Y = \{\ell_1, \ldots, \ell_L\}$ é referido como *conjunto de rótulos de classe*. No Exemplo 3, o conjunto das características é associado à $X = \mathbb{R}^6$ e $Y = \{\text{hérnia de disco}, \text{espondilolistese}, \text{nenhuma patologia}\}$, em que os rótulos em Y refletem a veracidade da proposição "o paciente i, cujos dados coletados são $\mathbf{x}_i = [x_1, \ldots, x_6]^T \in X$ foi diagnosticado com $y_i \in Y$".

O próximo problema, cujos detalhes podem ser encontrados em (Ripley, 1996), ilustra uma última situação que pode ser encontrada na prática.

Exemplo 4 (Grupos de Vírus)**.** Foram coletados 61 vírus que afetam lavouras como de tomate, pepino, etc. e, para cada vírus, foram extraídas 18 medidas. A questão de interesse é determinar se existem grupos distintos de vírus. De um modo mais geral, tem-se a seguinte situação:

Situação 3 (Problema de Agrupamento)**.** *Dado um conjunto finito* $\{x_i \in X : i = 1, \ldots, n\}$*, em que* X *é um universo arbitrário, pede-se determinar uma função* $f : X \to Y$*, com* $Y = \{\ell_1, \ldots, \ell_L\}$*.*

O conjunto $\mathcal{C}_i = \{x \in X : f(x) = \ell_i\}$, que contém os elementos de X que compartilham a propriedade ℓ_i, é geralmente referido como um *agrupamento* no universo X.

Nas próximas seções, iremos referir aos problemas descritos pelas Situações 1, 2 e 3 como *problema de regressão, classificação e agrupamento*, respectivamente. Observe que um problema de classificação pode ser visto como um problema de regressão em que o conjunto Y é finito e deseja-se a igualdade $f(x_i) = y_i$ para todo $i = 1, \ldots, n$. Um problema de agrupamento, por sua vez, é semelhante ao problema de classificação em que os valores $\ell_1, \ldots, \ell_L$ não são fornecidos. É comum na literatura distinguir as duas situações como aprendizado supervisionado e não-supervisionado (Haykin, 2009; Mehrotra et al., 1997; Ripley, 1996).

Primeiramente, o termo "aprendizado" pode ser interpretado como a tarefa de determinar a função $f : X \to Y$. Com efeito, uma vez conhecida a

função f, sabemos (ou aprendemos) como os dados x_i e y_i estão relacionados. Além disso, esse termo é particularmente usado no contexto de redes neurais artificiais em que a tarefa de determinar a função corresponde ao ajuste das conexões sinápticas dos neurônios (Haykin, 2009; Hecht-Nielsen, 1989). No *aprendizado supervisionado*, também referido como *aprendizado com professor*, são fornecidos pares (x_i, y_i) para $i = 1, \ldots, n$. Intuitivamente, pensamos que um professor informou qual deve ser a resposta desejada y_i para cada observação x_i. Ambas Situações 1 e 2 são tratadas no contexto do aprendizado supervisionado. Na Situação 3, não se tem, *a priori*, os valores y_i. Em outras palavras, não há um professor que informa qual valor f deve assumir em x_i. Nesse caso, a função $f : X \to Y$ é determinada usando uma técnica referida como *aprendizado não-supervisionado* ou *sem professor*.

2. Abordagens para o Problema de Regressão

Um problema de ajuste de curva é dado em sua forma mais simples como um caso particular da Situação 1. De acordo com nossa apresentação, um problema de interpolação também pode ser visto como um caso particular da Situação 1 no qual deseja-se a igualdade $f(x_i) = y_i$ para todo $i = 1, \ldots, n$. Alguns autores, porém, podem distinguir entre problemas de ajuste de curvas e interpolação.

Em termos gerais, os métodos de regressão podem diferir entre si, entre outras coisas, pelas hipóteses adicionais que cada um impõe aos conjuntos X e Y e a forma de f. Por exemplo, um método poderia assumir que X e Y são conjuntos compactos e f é uma função contínua, enquanto que para outro método tais hipóteses não são necessárias. Repare ainda que certas imposições sobre a função f requerem automaticamente que X e Y sejam equipados com certas estruturas especiais. Por exemplo, só podemos falar da continuidade de uma função $f : X \to Y$ se X e Y forem espaços topológicos. No decorrer deste texto, a menos que seja necessário, omitiremos algumas hipóteses sobre X e Y se as mesmas estiverem claras de acordo com o contexto, tal como no caso do exemplo anterior.

Os métodos de regressão estatísticos assumem que existe uma dependência estocástica entre os valores $x \in X$ e $y \in Y$ representada pela probabilidade condicional de y dado x, $p(y|x)$. Assume-se também que a ocorrência aleatória de um elemento x é dada segundo uma distribuição de probabilidade

$p(x)$. Baseado nestas premissas, defini-se um operador $f : X \to Y$ chamado *regressão*, que fornece, para cada $x \in X$, a respectiva esperança matemática condicional de $y = f(x)$. Adicionalmente, considera-se que o conjunto de pares $(x_1, y_1), \ldots, (x_n, y_n)$ foram observados de forma aleatória e independente segundo uma função de distribuição conjunta $p(x, y) = p(x)p(y|x)$. Sob algumas condições, o problema de estimar f^* tal que f^* aproxima-se da função desconhecida f, pode ser reduzido à um problema de minimização da função de risco dada por uma esperança matemática com respeito a medida de probabilidade associada a função de distribuição conjunta $p(x, y)$. O leitor interessado pode obter mais informações em (Vapnik, 1998).

Uma das possíveis abordagens à problemas como os da Situação 1 é assumir que f pertence a uma certa classe de funções F, em geral parametrizada. Especificamente, assumimos que F é o conjunto das funções $\mathbf{f}(\,\cdot\,, \lambda_1, \ldots, \lambda_p)$ em que $\lambda_1, \ldots, \lambda_p$ são parâmetros[1]. Nesse caso, o problema de determinar $\mathbf{f} \in F$ resume-se em encontrar parâmetros $\lambda_1, \ldots, \lambda_p$ tais que $\mathbf{f}(x_i, \lambda_1, \ldots, \lambda_p) \approx y_i$, para todo $i = 1, \ldots, n$. Dizemos que o problema de regressão é *linear* se a função $\mathbf{f}$ for linear com respeito aos parâmetros $\lambda_1, \ldots, \lambda_p$. Caso contrário, dizemos que ele é *não linear*.

No caso linear, o conjunto F é completamente caracterizado pela combinação linear de funções $f_1, \ldots, f_p$, ou seja, podemos expressar $\mathbf{f} \in F$ como

$$\mathbf{f}(x, \lambda_1, \ldots, \lambda_p) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \ldots + \lambda_p f_p(x), \tag{2.3}$$

para todo $x \in X$. Note que não impomos nenhuma restrição sobre as funções $f_1, \ldots, f_p$ tal como continuidade, linearidade ou diferenciabilidade. Uma vez caracterizado F, o problema de regressão corresponde a resolução de um sistema de equações lineares

$$\lambda_1 f_1(x_i) + \lambda_2 f_2(x_i) + \ldots + \lambda_p f_p(x_i) = y_i, \ \forall i = 1, \ldots, n, \tag{2.4}$$

que pode ser resolvido, por exemplo, com o método convencional dos *quadrados mínimos* (Bjorck, 1996; Burden e Faires, 2004; Dahlquist e Bjorck, 2008; Trefethen e Bau III, 1997).

Exemplo 5 (Polinômios Trigonométricos)**.** No caso em que $f : X \to Y$ é uma função periódica com período T, podemos considerar a família de funções

[1]Nesse artigo usamos negrito para enfatizar que $\mathbf{f}(x, \lambda_1, \ldots, \lambda_p)$, além da variável independente x, depende também dos parâmetros $\lambda_1, \ldots, \lambda_p$.

paramétricas F descrita pela seguinte equação para parâmetros $a_0, a_1, \ldots, a_m$ e $b_1, \ldots, b_m$:

$$\mathbf{f}(x, a_0, \ldots, a_m, b_1, \ldots, b_m) = \frac{a_0}{2} + \sum_{i=1}^{m} \left(a_i \cos\left(\frac{2\pi i x}{T}\right) + b_i \operatorname{sen}\left(\frac{2\pi i x}{T}\right)\right). \tag{2.5}$$

Uma função $\mathbf{f}$ que satisfaz (2.5) é linear nos parâmetros $a_0, \ldots, a_m, b_1, \ldots, b_m$ e, portanto, é um exemplo linear para abordagem da Situação 1 (Stoer e Bulirsch, 1980). Os parâmetros $a_0, \ldots, a_m, b_1, \ldots, b_m$ em (2.5) são chamados *coeficientes de Fourier* (Dahlquist e Bjorck, 2008). No contexto da Situação 1, se os dados $x_1, \ldots, x_n$ são igualmente espaçados e há uma concordância entre n e o número de parâmetros de modo ser possível obter a igualdade $y_i = f(x_i), \forall i = 1, \ldots, n$, então, os parâmetros em (2.5) podem ser determinados de forma eficiente utilizando a chamada *transformada rápida de Fourier* (FFT, acrônimo do termo inglês *fast Fourier Transform*).

É importante observar que, ao mesmo tempo que a escolha de F simplifica o problema, também pode limitar a capacidade de "aderência" de $\mathbf{f}$ ao conjunto de dados $\{(x_i, y_i) : i = 1, \ldots, n\}$ se a escolha de F não for apropriada. As *redes neurais artificiais* (RNA), que podem ser vistas como uma abordagem não-linear para a Situação 1, tem se mostrado eficientes em problemas de regressão devido à sua capacidade de aderência ao conjunto de dados (Hassoun, 1995; Haykin, 2009; Hecht-Nielsen, 1989).

Exemplo 6 (Redes Neurais Artificiais)**.** Uma RNA é um modelo matemático inspirado no funcionamento do cérebro, cujas unidades de processamento são os *neurônios*. O projeto de uma RNA envolve duas etapas:

1. Determinar a topologia da rede, ou seja, como os neurônios estão organizados e interagem entre si.

2. Definir a regra de aprendizado, ou seja, como são ajustados o parâmetros, referidos como *pesos sinápticos*.

No nosso contexto, a primeira etapa corresponde a escolha da função $\mathbf{f}$. Por exemplo, um caso particular de uma rede MLP (acrônimo do termo inglês *multilayer perceptron*) estabelece um mapeamento $\mathbf{f}(\cdot, w_{11}, \ldots, w_{Ld}, m_1, \ldots, m_L) : \mathbb{R}^d \to \mathbb{R}$ dado pela seguinte equação para todo vetor $\mathbf{x} = [x_1, \ldots, x_d]^T \in \mathbb{R}^d$:

$$\mathbf{f}(\mathbf{x}, w_{11}, \ldots, w_{Ld}, m_1, \ldots, m_L) = \sum_{l=1}^{L} m_l \varphi \left(\sum_{j=1}^{d} w_{lj} x_j \right), \tag{2.6}$$

em que φ denota uma função diferenciável, monótona e limitada (Haykin, 2009; Vapnik, 1998) tal como *tanh* e a função logística dada por $\mathcal{S}(t) = 1/(1+e^{-t})$. Os parâmetros $w_{11}, \ldots, w_{Ld}, m_1, \ldots, m_L$ são chamados *pesos sinápticos* da RNA. A segunda etapa consiste na aplicação de um método de ajuste não linear tal como a classe dos *algoritmos de retropropagação* (Hassoun, 1995; Haykin, 2009; Hecht-Nielsen, 1989). Por exemplo, na classe dos algoritmos de retropropagação, os parâmetros $w_{11}, \ldots, w_{Ld}, m_1, \ldots, m_L$ da rede MLP descrita em (2.6) são determinados minimizando o erro quadrático médio, denotado por $E := E(w_{11}, \ldots, w_{Ld}, m_1, \ldots, m_L)$, sobre o conjunto dos dados de treinamento $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$, com $\mathbf{x}_i = [x_{1i}, \ldots, x_{di}]^T \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{L} m_l \varphi\left(\sum_{j=1}^{d} w_{lj} x_{ji}\right)\right)^2. \tag{2.7}$$

O leitor interessado na teoria e aplicações de RNAs pode consultar (Hassoun, 1995; Haykin, 2009; Mehrotra et al., 1997).

3. Abordagens para o Problema de Classificação

A Situação 2 representa a formulação matemática de um problema de classificação convencional e é um caso particular da Situação 1, distinguindo-se pelas restrição adicionais de Y ser um conjunto finito não vazio e que a função f a ser determinada seja capaz de reproduzir ao máximo os pares ordenados dados, isto é, a equação[2] $f(x_i) = y_i$ deve valer para todo $i = 1, \ldots, n$. Tais restrições podem ser impeditivas para a aplicação direta de um método de regressão qualquer. Por exemplo, é sabido que um método de interpolação polinomial produz um polinômio p que interpola perfeitamente os pontos (x_i, y_i), isto é, $p(x_i) = y_i$ para todo $i = 1, \ldots, n$. Porém, a menos que o polinômio seja de grau 0, seu conjunto imagem não é finito. Além disso, se restringirmos a escolha de f à um conjunto de funções F, então, dependendo da escolha de F, pode ocorrer que não exita $f \in F$ tal que $f(x_i) = y_i$ para todo $i = 1, \ldots, n$.

Exemplo 7 (Máquinas de Vetores de Suporte)**.** As máquinas de vetores de suporte (SVM, acrônimo do termo inglês *support vector machines*) compreende

[2]Gostaríamos de observar que a igualdade $f(x_i) = y_i, \forall i = 1, \ldots, n$ pode ser relaxada em algumas aplicações, por exemplo, no caso em que os dados apresentados $\{(x_i, y_i) : i = 1, \ldots, n\}$ estão sujeitos a erros.

uma classe importante de classificadores com significativo destaque tanto pela eficiência em situações práticas como pela elegância teórica (Haykin, 2009; Vapnik, 1998, 1999). Embora existam na literatura versões mais gerais, uma SVM é inicialmente projetada para resolver um caso particular da Situação 2 em que Y contém apenas dois rótulos, ou seja, um problema de classificação binário. Nesse contexto, considere um conjunto de dados $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ em que $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{di}]^T \in \mathbb{R}^d$ é um vetor coluna e $y_i \in \{-1, +1\}$ para todo $i = 1, \ldots, n$. No caso mais simples, referido como *SVM linear*, deseja-se encontrar um hiperplano que divide as classes $\mathcal{C}_1 = \{\mathbf{x}_i : y_i = +1\}$ e $\mathcal{C}_0 = \{\mathbf{x}_i : y_i = -1\}$. Em outras palavras, deseja-se encontrar um vetor de parâmetros $\boldsymbol{\lambda} \in \mathbb{R}^d$ e um parâmetro θ tais que

$$\begin{cases} \boldsymbol{\lambda}^T \mathbf{x}_i - \theta \geq \rho, & \text{se } \ y_i = +1, \\ \boldsymbol{\lambda}^T \mathbf{x}_i - \theta \leq \rho, & \text{se } \ y_i = -1, \end{cases} \tag{3.8}$$

para todo $i = 1, \ldots, n$. Aqui, o valor $\rho > 0$ denota a margem de separação entre as duas classes. O melhor classificador em (3.8) é obtido determinando $\boldsymbol{\lambda}$ e θ que maximizam ρ. Com intuito de determinar os parâmetros ótimos, primeiramente combinamos ambas inequações em (3.8) como

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i = 1, \ldots, n, \tag{3.9}$$

em que $\mathbf{w} = \boldsymbol{\lambda}/\rho$ e $b = -\theta/\rho$. Agora, pode-se mostrar que maximizar ρ é equivalente a minimizar a norma-2 de $\mathbf{w}$. Logo, em teoria, o melhor classificador é determinado resolvendo o seguinte problema de otimização quadrática

$$\underset{\mathbf{w}}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T \mathbf{w}, \tag{3.10}$$

$$\text{sujeito à} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad \forall i = 1, \ldots, n. \tag{3.11}$$

Contudo, o problema em (3.10) não admite soluções em muitas situações práticas. Com efeito, (3.10) possui solução somente se existir um hiperplano que separa as classes $\mathcal{C}_1$ e $\mathcal{C}_0$ dos pontos associados aos valores $+1$ e -1, respectivamente. Em vista desse fato, considera-se a seguinte modificação de (3.10) na qual são adicionadas variáveis de folga $s_1, \ldots, s_n$ que flexibilizam as restrições do pro-

blema original:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} s_i, \tag{3.12}$$

$$\text{sujeito à} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - s_i, \quad \forall i = 1, \ldots, n, \tag{3.13}$$

$$s_i \geq 0, \quad \forall i = 1, \ldots, n. \tag{3.14}$$

O parâmetro C em (3.12) controla tanto a complexidade da SVM linear como o número de pontos não-separáveis pelo hiperplano. Com efeito, por um lado, grandes valores de C resultam valores pequenos de s_i. Consequentemente, a inequações em (3.13) são quase todas satisfeitas. Em outras palavras, há uma certa confiança na qualidade do conjunto de dados $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$. Por outro lado, valores pequenos de C são recomendados quando os dados estão sujeitos a erros e um menor peso deve ser creditado a eles. Por fim, de um modo intuitivo, uma SVM não-linear é obtida substituindo em (3.12) $\mathbf{x}_i$ por $\Phi(\mathbf{x}_i)$, em que Φ é uma função vetorial (Haykin, 2009; Vapnik, 1998). Na prática, porém, não é necessário conhecer Φ, mas apenas uma função $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, chamada *kernel*, que satisfaz $k(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1)^T\Phi(\mathbf{x}_2)$ para todo $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$.

Apesar das observações apresentadas no início da seção, um regressor pode ser aplicado em um problema de classificação da seguinte forma. Dado um conjunto de pares $\{(x_i, y_i) : i = 1, \ldots, n\}$, suponha que conhecemos um regressor $r : X \to Z$, tal que $r(x_i) \approx y_i$ para todo $i = 1, \ldots, n$, em que $Y = \{\ell_1, \ell_2, \ldots, \ell_L\} \subseteq Z$. Suponha também que Z é equipado com uma medida de distância σ, ou seja, $\sigma(z_1, z_2) \leq \sigma(z_1, z_3)$ sempre que z_1 for mais "próximo" de z_2 que z_3. Por exemplo, podemos considerar $\sigma(z_1, z_2) = \|z_1 - z_2\|_2$ quando Z for um espaço vetorial Euclidiano. Agora, obtemos um classificador $f : X \to Y$ compondo r com uma função do tipo "argmin", que retorna o argumento que minimiza uma certa função. Precisamente, definimos a função f que associa x à classe $\ell_i \in Y$ da seguinte forma:

$$f(x) = \underset{\ell_i \in Y}{\operatorname{argmin}}\, \sigma(r(x), \ell_i). \tag{3.15}$$

Repare em (3.15) que há uma competição entre os rótulos de classe $\ell_1, \ldots, \ell_L$ e o vencedor é aquele mais próximo de $f(x)$ em termos da medida σ. Resumindo, primeiro obtém-se uma regra funcional f para os pares (x_i, y_i) cuja imagem não é necessariamente restrita a conjunto Y. Posteriormente, verifica-se para cada ponto da imagem de f qual é o respectivo elemento de Y mais próximo segundo a medida σ.

Outra possível abordagem consiste em atribuir um rótulo de classe a um elemento $x \in X$ baseado nos rótulos das classes dos pontos x_i e sua respectiva proximidade à x. A hipótese central desta abordagem é que elementos suficientemente próximos em X possuem o mesmo rótulo de classe. Em contraste com a abordagem anterior, a competição ocorre no conjunto das características e não num conjunto Z que contém Y.

Exemplo 8 (k-Vizinhos Mais Próximos)**.** Um dos classificadores mais conhecidos da literatura que utiliza a similaridade no conjunto das características é o *k-vizinhos mais próximos* (kNN, acrônimo do inglês *k-nearest neighbors*) (Duda et al., 2001). Dado uma certa medida de distância σ em X e um inteiro positivo $k \leq n$, o kNN associa à x o rótulo da classe que mais ocorre entre os k pontos $x_{\mu_1}, \ldots, x_{\mu_k}$ pertencentes a vizinhança de x limitada pela k-ésima menor distancia entre x e x_j, para $j = 1, \ldots, n$, isto é, os pontos x_{μ_i} são tais que $\sigma(x, x_{\mu_i}) \leq \sigma(x, x_j)$ para todo j tal que $j \neq \mu_i$, para $i = 1, \ldots, k$.

Observação 1. O kNN pode depender criticamente da distância adotada para comparar os elementos de X. Além disso, como esse método requer a comparação da entrada x com todos os dados $x_1, \ldots, x_n$, sua aplicação torna-se lenta quando n é grande e quando o classificador é utilizado repetidas vezes para diferentes entradas. Nessas situações, um classificador através de (3.15) pode ser mais indicado.

Por fim, a maioria dos classificadores com fundamentos estatísticos executam a classificação de um elemento x baseado no cálculo da probabilidade de atribuir à classe $\ell \in Y$ dado $x \in X$, ou seja, baseado no cálculo da função de distribuição condicionada $p(\ell|x)$ (Vapnik, 1998). Por exemplo, o classificar de Bayes, também conhecido como regra de decisão posteriori máxima, associa x a ℓ_i tal que $p(\ell_i|x) \geq p(\ell_j|x)$ para todo $\ell_j \in Y$ (Ripley, 1996). Neste contexto, o teorema de Bayes torna-se uma ferramenta útil para estimar $p(\ell|x)$ via a seguinte equação

$$p(\ell|x) = \frac{p(\ell)p(x|\ell)}{p(x)}. \tag{3.16}$$

Tal como os métodos de regressão, os métodos para se estimar $p(x|\ell)$ se dividem em paramétricos e não paramétricos. Os métodos paramétricos assumem que os valores $p(x|\ell_i)$ seguem uma função de distribuição de probabilidade parametrizada conhecida, tal como a distribuição normal. Assim, o problema torna-se o de estimar os parâmetros da distribuições assumidas. Por fim, os modelos não

paramétricos, tal como o nome sugere, são aqueles que não impõem restrições à forma da função de distribuição condicionada $p(x|\ell)$. Neste caso, métodos de estimação de funções de densidade não paramétricos, tal como *janela de Parzen* (Duda e Hart, 1973; Duda et al., 2001), são utilizados.

4. Abordagens para o Problema de Agrupamento

No problema de agrupamento, dado apenas um conjunto finito $\{x_i \in X : i = 1, \ldots, n\}$, procuramos uma função $f : X \to Y$, em que $Y = \{\ell_1, \ldots, \ell_L\}$ é um conjunto com L rótulos. Tal como no problema de classificação, o universo X também é chamado *conjunto das características*. Porém, ao contrário do problema de classificação, o conjunto dos rótulos é desconhecido *a priori* no problema de agrupamento. Portanto, embora semelhantes, as abordagens empregadas nas Situações 2 e 3 são conceitualmente diferentes. Com efeito, essa interpretação para o problema de agrupamento é também referida na literatura como *problema de classificação não-supervisionado* (Xu e Wunsch, 2005).

O problema de agrupamento também pode ser formulado como um problema de particionar o conjunto das características. Com efeito, tomando $\mathcal{C}_i = \{x \in X : f(x) = \ell_i\}$ para $i = 1, \ldots, L$, obtemos subconjuntos disjuntos $\mathcal{C}_1, \ldots, \mathcal{C}_L$ que cobrem X, isto é,

$$\bigcup_{i=1}^{L} \mathcal{C}_i = X \quad \text{com} \quad \mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \forall i \neq j. \tag{4.17}$$

Reciprocamente, dados conjuntos $\mathcal{C}_1, \ldots, \mathcal{C}_L$ disjuntos que cobrem X, podemos definir a função $f : X \to Y$ através da equivalência $f(x) = \ell_i \iff x \in \mathcal{C}_i$. Observe que não foi imposto nenhuma restrição aos subconjuntos $\mathcal{C}_i$, tal como convexidade, ou sobre o valor de L.

Observação 2. Embora semelhantes, a formulação do problema de agrupamento como um problema de particionamento de X permite o desenvolvimento de técnicas supervisionadas. Por exemplo, os subconjuntos disjuntos que particionam X podem ser determinados com base em algumas amostras de $\mathcal{C}_j$, para $j = 1, \ldots, L$, além do conjunto finito $\{x_i : i = 1, \ldots, n\}$ (Finley e Joachims, 2005; Marcu e Daume, 2005). Nesse artigo, porém, o problema de agrupamento é considerado apenas conforme exposto na Situação 3.

Uma abordagem relativamente simples para um problema de agrupamento não-supervisionado é formulada assumindo que o conjunto das carac-

terísticas X é equipado com uma medida de distância σ e que o conjunto dado $\mathcal{D} = \{x_1, \ldots, x_n\}$ é representativo o suficiente para se obter um particionamento adequado de X. Suponha que existam elementos $c_1, \ldots, c_L \in X$, chamados *centros dos grupos*, tais que

$$\mathcal{C}_i = \{x \in X : \sigma(c_i, x) \leq \sigma(c_j, x),\ \forall j \neq i\}, \tag{4.18}$$

para cada $i = 1, \ldots, L$. Em outras palavras, o grupo $\mathcal{C}_i$ é formado por todos os elementos de X que estão mais próximos de c_i que de c_j, para $i \neq j$. Nesse caso, o problema de agrupamento resume-se em encontrar os centros dos grupos $c_1, \ldots, c_L$. Formalmente, os centros dos grupos são definidos como os elementos que minimizam a qualidade do particionamento de X considerando o subconjunto $\mathcal{D} = \{x_1, \ldots, x_n\} \subseteq X$ dado, ou seja, os centros dos grupos são os elementos $c_1^*, \ldots, c_L^*$ que resolvem o problema de minimização

$$\min_{c_1,\ldots,c_L} \sum_{i=1}^{L} \left(\sum_{x_j \in \mathcal{D} \cap \mathcal{C}_i} \sigma(c_i, x_j) \right). \tag{4.19}$$

Note que $x_j \in \mathcal{D} \cap \mathcal{C}_i$ se e somente se o dado $x_j \in \mathcal{D}$ pertence ao grupo $\mathcal{C}_i$. Logo, o termo $\sum_{x_j \in \mathcal{D} \cap \mathcal{C}_i} \sigma(c_i, x_j)$ em (4.19) fornece a soma das distância entre todos os elementos x_j que pertencem ao grupo $\mathcal{C}_i$. Por fim, esta estratégia produz uma função $f : X \to Y = \{\ell_1, \ldots, \ell_L\}$ dada por

$$f(x) = \ell_{i^*} \quad \text{tal que} \quad \sigma(x, c_{i^*}) \leq \sigma(x, c_j),\ \forall j = 1, \ldots, L.$$

Exemplo 9 (k-médias)**.** Considere $X = \mathbb{R}^d$ equipado com a medida de distância $\sigma(\mathbf{x}, \mathbf{z}) = ||\mathbf{x} - \mathbf{z}||_2^2$ para todo $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$. O algoritmo *k-médias* (do inglês *k-means*) resolve o problema de minimização em (4.19) com L fixo. Especificamente, o problema em (4.19) pode ser escrito como o seguinte problema de otimização restrita em que $u_{ij} \in \{0, 1\}$ indica a pertinência de x_j em $\mathcal{C}_i$, isto é, $u_{ij} = 1 \iff x_j \in \mathcal{C}_i$ (Oliveira e Pedrycz, 2007):

$$\text{minimize} \sum_{i=1}^{L} \sum_{j=1}^{n} u_{ij} ||\mathbf{c}_i - \mathbf{x}_j||_2^2, \tag{4.20}$$

$$\text{sujeito à} \sum_{i=1}^{L} u_{ij} = 1, \quad \forall j = 1, \ldots, n, \tag{4.21}$$

$$\sum_{j=1}^{n} u_{ij} \geq 1, \quad \forall i = 1, \ldots, L, \tag{4.22}$$

$$u_{ij} \in \{0, 1\}, \quad \forall i = 1, \ldots, L \ \text{ e } \ \forall j = 1, \ldots, n. \tag{4.23}$$

Em termos gerais, a restrição (4.21) assegura que cada $\mathbf{x}_j$ pertence à um único grupo $\mathcal{C}_i$. A restrição (4.22) garante que cada grupo possui pelo menos um elemento do conjunto de dados $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Dado uma aproximação inicial $\{\mathbf{c}_1, \ldots, \mathbf{c}_L\} \in \mathbb{R}^d$, um mínimo local do problema (4.20)-(4.22) pode ser determinado como segue: Enquanto não houver mudanças significativas em $\mathbf{c}_1, \ldots, \mathbf{c}_L$, calcule

$$u_{ij} = \begin{cases} 1, & i = \underset{l=1,\ldots,L}{\operatorname{argmin}} \|\mathbf{c}_l - \mathbf{x}_j\|_2, \\ 0, & \text{caso contrário,} \end{cases} \tag{4.24}$$

para cada índice i e j e atualize os centros dos grupos através da equação

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} u_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}}, \quad \forall i = 1, \ldots, L. \tag{4.25}$$

Por fim, variações do k-médias inclui, por exemplo, o *fuzzy* k-médias (no inglês, *fuzzy k-means*), em que os conjuntos $\mathcal{C}_i$ são *fuzzy* para todo $i = 1, \ldots, L$ (Bezdek, 1981; Oliveira e Pedrycz, 2007). Outras técnicas de agrupamento incluem os mapas auto-organizáveis de Kohonen (2001) e agrupamentos hierárquicos (Duda et al., 2001).

Observação 3. Note que a cardinalidade do contra-domínio Y, isto é, $|Y| = L$, deve ser especificada *a priori* no k-médias. Por conta disso, alguns pesquisadores entendem que um *professor* deve fornecer a quantidade de grupos existentes e, consequentemente, classificam o k-médias como uma técnica de aprendizagem supervisionado. Em oposição à tais pequisadores, acreditamos que a definição do contra-domínio Y não é suficiente para classificar o método como supervisionado. Com efeito, diferente do problema de regressão e classificação, nada se sabe sobre os valores $f(x_i)$ além de que eles pertencem ao conjunto Y.

5. Considerações Finais

Neste artigo apresentamos algumas ideias gerais que estão presentes em diferentes abordagens para três tipos de situações das quais frequentemente nos deparamos ao lidar com um conjunto de dados. São estes: problemas de regressão, classificação e agrupamento.

Na teoria de *aprendizagem de máquina* (**?**Haykin, 2009), problemas de regressão e classificação surgem no contexto de aprendizagem supervisionado.

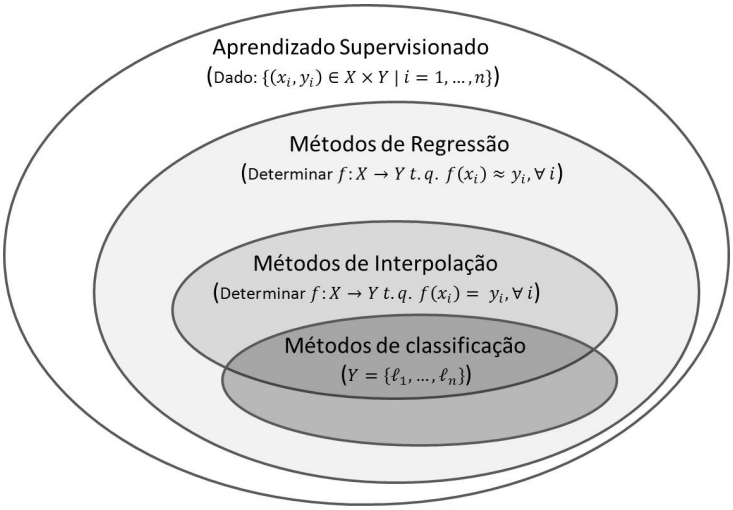

Figura 1: Métodos para aprendizagem supervisionada

A tarefa de classificação pode ser vista como um caso especial de interpolação cujo o contra-domínio Y é um conjunto finito se a igualdade $f(x_i) = y_i$ não for relaxada, isto é, se for não permitido haver erros de classificação nos elementos $x_1, \ldots, x_n$ dados. A interpolação, por sua vez, consiste de um caso particular de regressão. A Figura 1 ilustra a disposição hierárquica dos métodos para aprendizagem supervisionada discutidos neste artigo.

Em muitos casos o problema de regressão é formulado em termos de uma família F de funções paramétricas $\mathbf{f}(\cdot\, \lambda_1, \ldots, \lambda_p)$. O caso linear pode ser resolvido usando, por exemplo, quadrados mínimos (Bjorck, 1996; Burden e Faires, 2004; Dahlquist e Bjorck, 2008). Para o caso não linear citamos como o exemplo a rede neural MLP (Hassoun, 1995; Haykin, 2009; Hecht-Nielsen, 1989). Aparte desta, encontramos na literatura muitos outros modelos não lineares, tais como a *rede de base radial* (Haykin, 2009), modelos híbridos que combinam técnicas de aprendizagem de máquina com conjuntos fuzzy (Pedrycz e Gomide, 2007; Jang, 1993; Esmi et al., 2012), etc.

Os métodos de classificação, embora visto com um caso particular dos métodos de regressão na Figura 1, possuem um caráter distinto, pois o conjunto Y assume valores categóricos. Com efeito, revimos a SVM que converte o problema de classificação binário para um problema de otimização quadrática restrita (Vapnik, 1998). Outras abordagens envolvem competições ou em X ou

Figura 2: Métodos para aprendizagem não supervisionada

em Y. Por exemplo, o kNN envolve uma competição em X (Duda et al., 2001). No segundo, por exemplo, surge da composição de um regressor com a função argmin conforme (3.15).

Em contraste, a tarefa de agrupamento, apesar de parecida com a de classificação em termos de formulação, geralmente representa uma abordagem à apredizagem não-supersionado (Xu e Wunsch, 2005). A Figura 2 apresenta os métodos de agrupamento no contexto do aprendizado não-supervisionado. O problema de agrupamento é equivalente a encontrar uma partição $\mathcal{C}_1, \ldots, \mathcal{C}_L$ de X. Uma abordagem relativamente simples, que inclui o método k-médias, é descrito por (4.19) supondo uma medida de distância σ dada. Por fim, variações do k-médias inclui, por exemplo, o *fuzzy* k-médias (no inglês, *fuzzy k-means*), em que os conjuntos $\mathcal{C}_i$ são *fuzzy* para todo $i = 1, \ldots, L$ (Bezdek, 1981; Oliveira e Pedrycz, 2007). Outras técnicas de agrupamento incluem os mapas auto-organizáveis de Kohonen (2001) e agrupamentos hierárquicos (Duda et al., 2001).

Agradecimentos

Este trabalho foi parcialmente apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pelo Fundo de Apoio ao Ensino, à Pesquisa e à Extensão (FAEPEX) da Unicamp sob os processos 2013/12310-4, 304240/2011-7 e 519.292, respectivamente.

Referências

Bache, K. e Lichman, M. (2013). UCI machine learning repository. Disponível em: `http://archive.ics.uci.edu/ml`. Acesso em Maio, 2014.

Bezdek, J. C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, N. York.

Bjorck, A. (1996). *Numerical Methods for Least Squares Problems*. SIAM Publications, Philadelphia, 1^a edição.

Burden, R. L. e Faires, J. D. (2004). *Numerical Analysis*. Brooks Cole, 8^a edição.

Dahlquist, G. e Bjorck, A. (2008). *Numerical Methods in Scientific Computing*, volume 1. SIAM Publications, Philadelphia.

Duda, R. e Hart, P. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons, N. York.

Duda, R. O., Hart, P. E., e Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, N. York, 2^a edição.

Esmi, E. L., Sussner, P., Valle, M. E., Sakuray, F., e Barros, L. C. (2012). Fuzzy Associative Memories Based on Subsethood and Similarity Measures with Applications to Speaker Identification. In *Lecture Notes in Computer Science: International Conference on Hybrid Artificial Intelligence Systems (HAIS 2012)*, pag. 479–490. Springer, Berlin.

Finley, T. e Joachims, T. (2005). Supervised clustering with support vector machines. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pag. 217–224, N. York. ACM.

Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. MIT Press, Cambridge.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice-Hall, Upper Saddle River, NJ, 3^a edição.

Hecht-Nielsen, R. (1989). *Neurocomputing*. Addison-Wesley, Reading.

Jang, J.-S. R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685.

Kohonen, T. (2001). *Self-Organizing Maps*, volume 30 de *Springer Series in Information Sciences*. Springer, 3^a edição.

Marcu, D. e Daume, H. (2005). A bayesian model for supervised clustering with the dirichlet process prior. *Journal of Machine Learning Research*, 6:1551–1577.

Mehrotra, K., Mohan, C. K., e Ranka, S. (1997). *Elements of Artificial Neural Networks*. MIT Press, Cambridge.

Murray, J. D. (2002). *Mathematical Biology: I. An Introduction*. Interdisciplinary Applied Mathematics. Volume 1 of Mathematical Biology. Springer, 3^a edição.

Oliveira, J. V. e Pedrycz, W., editors (2007). *Advances in Fuzzy Clustering and its Applications*. John Wiley & Sons, Chichester.

Pedrycz, W. e Gomide, F. (2007). *Fuzzy Systems Engineering: Toward Human-Centric Computing*. Wiley-IEEE Press, N. York.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rocha Neto, A. R., Sousa, R., Barreto, G. A., e Cardoso, J. S. (2011). Diagnostic of pathology on the vertebral column with embedded reject option. In Vitrià, J., Sanches, J. M., e Hernández, M., editors, *Pattern Recognition and Image Analysis*, volume 6669 de *Lecture Notes in Computer Science*, pag. 588–595. Springer Berlin.

Stoer, J. e Bulirsch, R. (1980). *Introduction to Numerical Analysis*. Springer, N. York.

Trefethen, L. N. e Bau III, D. (1997). *Numerical Linear Algebra*. SIAM Publications, Philadelphia.

Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, N. York.

Vapnik, V. N. (1999). *The Nature of Statistical Learning Theory*. Springer, 2 edição.

Xu, R. e Wunsch, D., I. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.