Capítulo 6

Aprendizado de Máquina

6.1 Introdução

6.2 Motivação

O *aprendizado de máquina* é parte da área da Ciência da Computação conhecida como inteligência artificial. Embora tenha se popularizado, especialmente nos últimos anos, a inteligência artificial, incluído o aprendizado de máquina, já era considerada no passado entre as áreas essenciais para a computação. Por exemplo, nos anos 60, já havia departamentos inteiros em universidades, dedicados quase que integralmente a estas áreas. Depois de uma certa estagnação nos resultados apresentados, essas áreas perderam um pouco sua proeminência, embora permanecessem ativas ao longo dos anos.

Esse panorama sofreu enorme mudança a partir do início deste século. A inteligência artificial, em particular o aprendizado de máquina, tem colhido resultados concretos e relevantes, em diferentes áreas de aplicação. Os departamentos especializados em inteligência artificial estão ressurgindo nas universidades, motivados pelo sucesso dos resultados da área. Métodos baseados em aprendizado de máquina, apresentam importantes aplicações em atividades tão distintas, como por exemplo, mercado financeiro: para avaliar grau de risco na concessão de empréstimos; clínica médica: para elaborar possíveis diagnósticos médicos de pacientes; análise de conteúdo de correio eletrônico: para identificar mensagens espúrias; análise de jogos: para elaborar estratégias vencedoras de jogos discretos e muitos outros. Um exemplo elucidativo é o jogo de xadrez. Um programa de computador, baseado em aprendizado de máquina, recentemente derrotou no jogo de xadrez, um outro pro-

grama, que utilizava técnica diferente, o qual havia derrotado o enxadrista campeão mundial. Há poucos anos, seria impensável imaginar que um algoritmo pudesse derrotar o campeão mundial de xadrez.

6.3 Problema Inicial

Nesta seção, descrevemos um exemplo de problema que pode ser tratado através de aprendizado de máquina.

O problema consiste em elaborar um método geral para realizar diagnósticos médicos. Para tal, recorremos ao profissional especializado, no caso, um médico. Obtemos, então, uma lista de possíveis doenças, como por exemplo, gripe, gastrite, apendicite, bronquite, covid-19 etc. Em paralelo, compilamos uma lista dos elementos que serão usados para elaborar o diagnóstico: sintomas, temperatura do paciente, resultados do hemograma, dificuldade de locomoção do paciente, etc. O conjunto das doenças corresponde à informação principal que desejamos prever, representado por um vetor. Para cada doença, isto é, elemento do vetor, são considerados todos os componentes do diagnóstico com a informação da sua relevância, para a doença em questão. Em princípio, iremos considerar os componentes do diagnóstico como informação binária, se relevante ou não. Por outro lado, utilizamos um conjunto de pesos para ponderar cada sintoma, em relação às doenças. O peso representará a importância relativa do sintoma em questão, em relação à doença considerada. Os pesos são representados por números reais, positivos ou negativos. Por exemplo, os sintomas "dor de garganta" e "falta de ar" terão pesos maiores para a doença "gripe" do que para "gastrite". No caso sintoma "falta de ar", um peso ainda maior para a doença covid-19.

O conjunto dos elementos que desejamos identificar e classificar que, no exemplo acima, corresponde ao conjunto das doenças, é denominado *conjunto universo*, com n elementos. Cada um destes elementos consiste de um vetor de dimensão d quantificado pelos seus atributos. Cada atributo corresponde a um possível sintoma com a sua ponderação. Esses atributos correspondem ao conjunto de sintomas, denominado *conjunto de predicados*. O conjunto dos pesos corresponde às incógnitas a serem determinadas.

Assim sendo, os dados de entrada correspondem a um vetor, o conjunto universo, d-dimensional e com n elementos. Além disso, o conjunto de predicados é conhecido, e corresponde a cada dimensão do vetor. Há um conjunto de pesos, a serem determinados, com o objetivo de ponderar cada predicado de cada elemento do conjunto universo.

Com o auxílio destes pesos será calculado um valor numérico, para cada elemento do conjunto universo. Finalmente, há também um número real b, a ser determinado, o qual é denominado *limiar*. Se o valor numérico do elemento em questão for maior do

que o limiar, a saída para o elemento correspondente será SIM, isto é, o diagnóstico é positivo para a doença; caso contrário, NÃO. (Colocar expressão do somatório (com o b) separando as regiões na Figura 6.1(a). Na Figura (b), tornar a reta invisível.)

Nesse contexto, o nosso problema corresponde a escolher, apropriadamente, os d pesos para cada elemento do conjunto universo, bem como o valor limiar, de modo a permitir a classificação, SIM ou NÃO, para cada elemento.

Como obter uma solução para o nosso problema?

A idéia geral é utilizar o método do aprendizado. Isto é, partimos, inicialmente, de um subconjunto do conjunto universo, que já se encontra classificado, para o qual supõe-se conhecido um limiar apropriado b, bem como os pesos para cada elemento do subconjunto. Com estas informações, os elementos deste subconjunto já terão o seu valor numérico associado, o qual corresponde a um vetor d-dimensional. Observando o conjunto de valores numéricos, eles deverão permitir uma separação através de um hiperplano, que divida o espaço d-dimensional em duas partes compatíveis com os valores numéricos obtidos: o semi-espaço SIM e o semi-espaço NÃO.

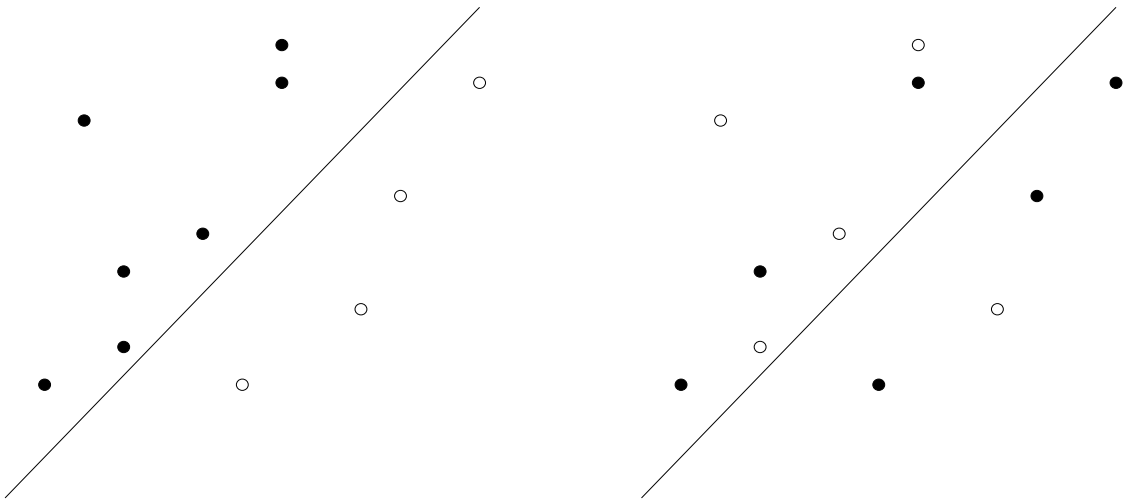
No aprendizado de máquina, os valores atribuídos a este subconjunto inicial, são obtidos seguindo as informações fornecidas pelo especialista em questão, denominado *professor*. A ideia seria aprender, de certa forma, como estes valores foram atribuídos, de modo a automatizar o processo. As atribuições obtidas do professor são denominadas *exemplos de aprendizado*, ou simplesmente *exemplos*.

Uma ilustração de uma possível separação em dois semi-espaços é apresentada na Figura 6.1(a), no qual os valores atribuídos aos pesos e ao limiar, permitiriam esta separação. No exemplo da Figura 6.1(b), no entanto, esta separação não é possível.

Ambos os exemplos são de dimensão 2, cujos hiperplanos consistem de retas no plano. Os círculos vazados representam elementos que deveriam ser classificados como SIM, e os cheios classificados como NÃO. Devemos ressaltar que o nosso objetivo é definir os pesos de modo que os elementos do conjunto universo possam ser separados linearmente, isto é, segundo um hiperplano. Caso isto não seja possível, o processo se torna mais complexo.

De acordo com a descrição acima, o valor de cada elemento do conjunto universo é obtido através de uma soma dos pesos atribuídos a cada predicado do elemento. Se esta soma for superior ao valor limiar, a resposta para o referido elemento será rotulado com SIM, caso contrário NÃO. Iremos utilizar os valores $+1$ e -1, respectivamente, para designar SIM e NÃO.

O conjunto dos dados será denotado por X, com elementos $X_1, X_2, \ldots, X_n$, onde cada X_i é um vetor no espaço d-dimensional. Cada X_i, por sua vez, é ponderado por um vetor W, também no espaço d-dimensional.



(a) (b)

Figura 6.1: Elementos do conjunto universo, classificado segundo os pesos e o limiar. Pontos em negrito representam itens com $y'_i = 1$ e os vazados itens com $y'_i = -1$.

6.3.1 Descrição do Problema

Os dados do problema correspondem a uma tripla (X, W, b), onde $X = \{X_1, X_2, \ldots, X_n\}$ é o conjunto universo de itens, $W = \{w_1, w_2, \ldots, w_d\}$ é um conjunto de pesos a serem aplicados sobre um conjunto P de d predicados. Assim, cada item $X_i \in X$ por sua vez é um vetor $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, onde a cada um de seus elementos x_{ij} é aplicado ao peso w_j do predicado correspondente. A variável b se refere ao limiar. Cada um dos valores x_{ij}, w_j e b é um número real.

O objetivo é determinar um valor y_i igual a $+1$ ou -1, denominado *resultado* de x_i para cada conjunto X_i de entrada. O valor y_i é determinado em função de X_i, W, e b, da seguinte maneira:

$$\sum_{j=1}^{d} x_{ij} w_j + b \begin{cases} > 0 \Rightarrow y_i = +1 \\ < 0 \Rightarrow y_i = -1 \\ = 0 \Rightarrow y_i = 0 \end{cases}$$

No exemplo considerado da aplicação em diagnóstico médico, cada x_i é uma doença possível a ser considerada, cada $x_{ij} \in X_i$ é um valor atribuído ao sintoma relativo ao predicado P_j correspondente, e w_j o seu peso. O objetivo é determinar um resultado y_i, referente à doença X_i, tal que $y_i = +1$ indique que o paciente está acometido da doença X_i, enquanto $y_i = -1$ no caso em que o paciente não esteja acometido da doença. Este objetivo é alcançado se o conjunto X for linearmente

separável. Para tal, iremos determinar um hiperplano

$$\sum_{j=1}^{d} x_{ij} w_j + b = 0$$

o qual fará a separação linear do conjunto entre valores $y_i = +1$ e $y_i = -1$. Os pontos que satisfazem esta equação estão localizados sobre o hiperplano, e possuem $y_i = 0$, o que indica que o hiperplano não realizou a separação do conjunto.

Observe que se não houvesse o limiar b, a origem sempre seria um ponto do hiperplano. E por consequência, a origem não poderia ser classificada como $+1$ ou -1 conforme a conveniência do problema em questão.

Para tornar a notação ligeiramente mais concisa, iremos eliminar do cálculo o limiar b, acrescentando em seu lugar um predicado P_0, cujos valores relativos aos itens X_i são $x_{1,0} = x_{2,0} = \ldots = x_{n,0} = b$.
Adicionando também um peso inicial $w_0 = 1$, o cálculo se torna:

$$\sum_{j=0}^{d} x_{ij} w_j \begin{cases} > 0 \Rightarrow y_i = +1 \\ < 0 \Rightarrow y_i = -1 \\ = 0 \Rightarrow y_i = 0 \end{cases}$$

Finalmente, podemos escrever o resultado calculado y_i simplesmente como

$$y_i = \text{sinal}\left(\sum_{j=0}^{d} x_{ij} w_j\right), \text{ onde}$$

$$\text{sinal}(z) = \begin{cases} +1, & \text{se } z > 0 \\ -1, & \text{se } z < 0 \\ 0, & \text{se } z = 0 \end{cases}$$

Daqui em diante, o hiperplano $\sum_{j=0}^{d} x_{ij} w_j = 0$ é aquele procurado para separar linearmente os pontos da entrada.

Além do resultado y_i, para cada $X_i \in X$, há um valor y'_i, denominado *resultado previsto*. O resultado previsto é fornecido juntamente com a entrada, em contrapartida ao resultado calculado y_i. O valor y'_i, dado, é sempre $y'_i = +1$ ou $y'_i = -1$, não se admitindo o valor $y'_i = 0$.

Os valores de y_i podem então ser divididos em dois subconjuntos, segundo o valor seja igual a $+1$ ou -1. Esses valores constituem pontos no espaço de d dimensões. O método que descrevemos neste capítulo se aplica quando estes subconjuntos forem *linearmente separáveis*. Isto é, existe um hiperplano que separa os pontos dos subconjuntos com $y_i = +1$ e $y_i = -1$. No exemplo da Figura 6.1(a), os pontos em negrito representam aqueles com $y'_i = +1$, os quais podem ser separados dos pontos

vazados, com $y'_i = -1$, mediante o traçado de um hiperplano. Representam, pois, um conjunto linearmente separável. A Figura 6.1(b) ilustra um caso de um conjunto que não é linearmente separável.

Assim sendo, a ideia de utilizar deste método de aprendizado para a solução de problemas é a seguinte: são dados os valores reais dos itens X_i isto é, x_{ij}, $i = 1, 2, \ldots, n$, e $j = 1, 2, \ldots, d$, o valor do limiar b, os valores (iniciais) dos pesos w_j, $j = 1, 2, \ldots, d$, bem como o valor de cada resultado previsto y'_i, $+1$ ou -1. O objetivo é determinar o resultado y_i, $i = 1, 2, \ldots, n$. Em relação ao resultado previsto, este deve ser fornecido por um especialista ou professor. No caso da aplicação em diagnóstico médico, segundo sua avaliação de que o paciente esteja acometido ou não, de cada doença em questão, considerando os sintomas apresentados.

A dinâmica da utilização do método está abaixo detalhada. Os valores dos resultados de y_i, $i = 1, 2, \ldots, n$, são determinados utilizando os dados da entrada. Se o resultado y_i é igual ao resultado previsto y'_i, para todo $X_i \in X$, o processo termina. Caso contrário, escolhe-se um item $X_i \in X$, tal que $y_i \neq y'_i$ isto é, sinal$(\sum_j x_{ij} w_j) \neq y'_i$. Neste caso, os pesos w_j são atualizados, visando corrigir a divergência.

Os pesos serão atualizados segundo as seguintes expressões:

$$w_j \leftarrow w_j + y'_i x_{ij}, \text{ para } 0 \leq j \leq d \tag{6.1}$$

onde x_{ij} é um elemento do item X_i em que

$$\text{sinal}\left(\sum_{j=0}^{d} x_{ij} w_j\right) \neq y'_i$$

Os pesos, portanto sofrerão atualizações em iterações distintas do cálculo. Seja $w_j(t)$ o peso do predicado P_j na iteração t do processo.

O lema seguinte garante que após a atualização de w_j, mediante a expressão acima, o valor do resultado obtido sinal$(\sum_{j=0}^{d} x_{ij} w_j)$ se torna mais próximo do previsto y'_i.

Lema 6.1. *Seja x_i um item de entrada tal que o resultado previsto y'_i difere do esperado y_i, após a iteração t do processo. Suponha que na iteração $t + 1$ os pesos serão atualizados mediante a aplicação da Equação 6.1, relativa ao item x_i. Então*

$$\sum_j x_{ij} w_j(t+1) > \sum_j x_{ij} w_j(t)$$

se e somente se $y'_i = +1$.

Demonstração. Como $y'_i \neq y_i$ na iteração t, sabemos que

$$y'_i \neq \text{sinal}\left(\sum_{j=0}^{d} x_{ij} w_j(t)\right)$$

Na iteração $t+1$, cada peso $w_j(t)$ é atualizado para $w_j(t+1)$, cujos valores são

$$w_j(t+1) = w_j(t) + y'_i x_{ij}, \text{ para todo } 0 \leq j \leq d.$$

Vamos comparar $\sum_j x_{ij} w_j(t+1)$ com $\sum_j x_{ij} w_j(t)$. Assim,

$$\sum_j x_{ij} w_j(t+1) = \sum_j x_{ij} \left[w_j(t) + y'_i x_{ij}\right] =$$

$$= \sum_j \left(x_{ij} w_j(t) + y'_i x_{ij}^2\right).$$

Da última igualdade, decorre que

$$\sum_j x_{ij} w_j(t+1) > \sum_j x_{ij} w_j(t),$$

se e somente se $y'_i > 0$, isto é, $y'_i = +1$. ■

Decorre do lema anterior que, após a atualização de $w_j(t)$, o valor do resultado calculado y_i, ou se tornou igual ao previsto y'_i, ou caso contrário, $\sum_j x_{ij} w_j(t+1)$ se tornou mais próximo de zero, isto é, mais próximo do ponto de mudança de sinal, onde alcançaria o valor de y'_i.

Com efeito, suponha inicialmente $y'_i = +1$. Então, $y_i = -1$. Isto é $\sum_j x_{ij} w_j < 0$. Do Lema 6.1, sabemos que

$$\sum_j x_{ij} w_j(t+1) > \sum_j x_{ij} w_j(t).$$

Neste caso, se $\sum_j x_{ij} w_j(t+1) > 0$, então $y_i = +1$. Ou seja, a aplicação da atualização de $w_j(t)$ produziu a igualdade desejada $y_i = y'_i$. Mas caso contrário, $\sum_j x_{ij} w_j(t+1)$ decresceu. Ou seja, se aproximou do valor que tornaria iguais y_i e y'_i.

O argumento para $y'_i = -1$ é similar.

Uma observação importante é que, apesar da atualização dos pesos w_j conduzir a uma aproximação do valor $\sum_j x_{ij} w_j(t+1)$ ao valor de y'_i, não há qualquer garantia em relação aos valores de $\sum_j x_{\ell j} w_j(t+1)$, para $\ell \neq i$. De fato, a aplicação da atualização dos pesos w_j, considerando um certo item x_i, em que havia a discordância $y_i \neq y'_i$

na iteração t, pode produzir uma discordância dos valores $y_\ell \neq y'_\ell$, para um certo item x_ℓ, em que havia a igualdade $y_\ell = y'_\ell$ na iteração t, isto é, antes da atualização dos pesos.

6.4 O Algoritmo do Perceptron

O Algoritmo do Perceptron pode ser agora formulado. A entrada consiste de um conjunto universo X, composto de itens X_i, $i = 1, \ldots, n$, onde cada X_i é um vetor de tamanho d, com elementos x_{ij}, $1 \leq i \leq n$ e $1 \leq j \leq d$. Há um conjunto de pesos $W = w_1, \ldots, w_d$, onde cada w_j é aplicado a um predicado P_j. Para cada item $X_i \in X$, é ainda fornecido o resultado previsto y'_i. Finalmente, há o valor limiar b, também fornecido de entrada. Os valores dos elementos x_{ij}, pesos w_j e limiar b são todos reais, enquanto o valor previsto y'_i é igual a ± 1.

O objetivo é determinar o resultado calculado y_i, para cada item X_i. O valor de tal resultado é também igual a ± 1. A finalidade é conseguir a igualdade $y_i = y'_i$ para todo $1 \leq i \leq n$. Enquanto tal igualdade não for atingida, procede-se a atualização dos pesos w_j. O procedimento termina quando a igualdade $y_i = y'_i$ é alcançada para todo $1 \leq i \leq n$.

A formulação seguinte descreve o processo.

Algumas informações importantes em relação ao Algoritmo do Perceptron são consideradas a seguir.

O Algoritmo do Perceptron geralmente apresenta um ótimo desempenho na prática, desde que o conjunto de dados seja linearmente separável. Se este não for linearmente separável, o algoritmo não termina, pois sempre haverá algum item X_i, tal que $y_i \neq y'_i$, após uma certa iteração.

Se o conjunto de dados for linearmente separável, é possível provar que o algoritmo converge para uma solução. Embora, na prática, ele termine rapidamente, não há qualquer garantia neste sentido. Na realidade, o algoritmo pode terminar em um número exponencial de passos. A velocidade da convergência na direção de uma solução pode ser avaliada através do conceito de margem, descrito a seguir.

Seja X um conjunto de dados linearmente separável, com itens X_i e y_i um resultado, para cada X_i, tal que $y_i = y'_i$. A *margem* do separador linear é a menor distância do ponto correspondente ao resultado y_i de algum item X_i até o hiperplano que separa as soluções $y_i = +1$ das soluções $y_i = -1$. Ver Figura 6.2. O interesse é determinar um separador linear $\sum_{j=0}^{d} x_{ij} w_j = 0$ que apresente margem máxima em relação aos demais.

O conceito de margem tem algumas consequências para o desempenho e estabilidade do Algoritmo do Perceptron. Pode ser provado que o número de passos requerido pelo Algoritmo do Perceptron é, no pior caso, inversamente proporcional

Algoritmo 6.1 Algoritmo do Perceptron

Dados: $X = \{X_1, \ldots, X_n\}$, $X_i = \{x_{i1}, \ldots, x_{id}\}$, $W = \{w_1, \ldots, w_d\}$, $Y' = \{y'_1, \ldots, y'_n\}$, b, onde $x_{ij}, w_j, b \in \mathbb{R}$ e $y'_i \in \{-1, +1\}$
  para $i \leftarrow 1, 2, \ldots, n$ **:**
    $x_{i,0} \leftarrow -b$; $X_i \leftarrow X_i \cup \{x_{i,0}\}$
  $w_0 \leftarrow 1$; $W \leftarrow W \cup \{w_0\}$
  $\ell \leftarrow 0$
  enquanto $\ell = 0$ **:**
    para $i \leftarrow 1, 2, \ldots, n$ **:**
      $soma \leftarrow \sum_{j=0}^{d} x_{ij} w_j$
      $y_i \leftarrow \text{sinal}(soma)$
    $i \leftarrow 1$; $\ell \leftarrow 1$
    repetir
      se $y_i \neq y'_i$ **então**
        para $j \leftarrow 0, 1, \ldots, d$ **:**
          $w_j \leftarrow w_j + y'_i x_{ij}$
        $\ell \leftarrow 0$
      senão
        $i \leftarrow i + 1$
    até que $i > n$ ou $\ell = 0$

ao quadrado da margem da solução. Note que uma margem maior implica em duas vantagens:

1. A primeira é que uma solução é obtida em um menor número de iterações. Intuitivamente, pode ser observado que se os elementos a serem separados já estão distantes uns dos outros, a tarefa fica facilitada.

2. A segunda é que o sistema fica menos susceptível a alterações na classificação, resultante da introdução de uma pequena variação nos predicados dos itens. Observe que pontos "próximos" aos pontos de entrada tendem a ser classificados no mesmo conjunto.

O exemplo seguinte ilustra o aprendizado da função booleana OU por um Perceptron. Os dados de entrada correspondem a $n = 4$ itens de dimensão $d = 2$. Os resultados previstos $\{y'_1, y'_2, y'_3, y'_4\}$ correspondem aos valores da função OU, com $y'_1 = -1$ e $y'_1 = +1$ refletindo os resultados de 0 e 1 do OU lógico, respectivamente.

Modificar a tabela e colocá-la no lugar correto. Modificar o texto descritivo da tabela. Colocar b com valor -1.

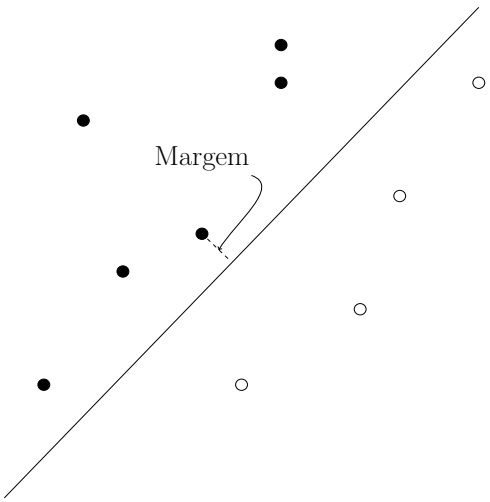


Figura 6.2: Margem de uma solução de um conjunto linearmente separável. Pontos em negrito representam pesos com $y'_i = 1$ e os vazados pesos com $y'_i = -1$.

A Tabela **??** detalha os cálculos iterativos, desde a primeira iteração até a última, quando os resultados calculados são iguais aos previstos para todos os itens da entrada. As colunas de 1 a 4 contém os detalhes dos cálculos da soma $\sum_{j=0}^{2} x_{ij}w_j = x_{i,0}w_0 + x_{i,1}w_1 + x_{i,2}w_2$ para todos os itens de entrada, ou seja, para $i = 1, 2, 3, 4$. As colunas de 5 a 8 contém os resultados calculados y_1, y_2, y_3, y_4, respectivamente, mediante a aplicação da equação

$$y_i = \text{sinal}\left(\sum_{j=0}^{d} x_{ij}w_j\right)$$

Quando o resultado calculado y_i difere do previsto y'_i, o fato é indicado por um círculo, em torno do primeiro. As colunas 3 a 12 contém a atualização dos pesos w_j relativo a um índice i escolhido, segundo o critério de que o resultado calculado y_i diferiu do previsto y'_i. O valor do índice i escolhido aparece na coluna 9, e os valores w_0, w_1, w_2 atualizados estão nas colunas 10 a 12, respectivamente. Estes valores dos pesos atualizados serão aplicados na próxima iteração.

O processo se repete, iterativamente, enquanto perdurar a desigualdade $y_i \neq y'_i$. No exemplo, foram necessárias 7 iterações para atingir a igualdade. O desempenho do Algoritmo do Perceptron é considerado bastante satisfatório para efeitos práticos. Geralmente, produz resultados em relativamente poucas iterações. Contudo, não há garantia de um bom desempenho. Com efeito, o número de iterações necessárias para obter a solução final não é limitado polinomialmente. Além disso, o algoritmo pode não terminar para entradas cujos resultados não admitam uma separação linear.

Deve ser observado que o problema geral resolvido pelo Algoritmo do Perceptron

pode ser resolvido por programação linear e, portanto, admite algoritmo polinomial. Abaixo, encontra-se a formulação de um programa linear que pode ser usado como solução geral.

São dados os valores constantes n, d, bem como os conjuntos $X_i = \{x_{i,1}, \ldots, x_{i,d}\}$ para cada $1 \leq i \leq n$. Além disso, o vetor de resultados previstos $Y' = \{y'_1, \ldots, y'_n\}$, onde cada $y'_i \in \{-1, +1\}$. O problema consiste em determinar o conjunto de pesos $W = \{w_1, \ldots, w_d\}$ e um valor limiar b tal que

$$\begin{array}{ll} \sum_{j=1}^{d} x_{ij} w_j + b > 0, & \text{se } y'_i = +1 \\ \sum_{j=1}^{d} x_{ij} w_j + b < 0, & \text{se } y'_i = -1 \end{array}$$

Este problema pode ser resolvido pela seguinte formulação de programação linear, onde se postulam os pesos $w_1, \ldots, w_d$ e limiar b de soma mínima,

$$\begin{array}{ll} \text{minimizar} & \left(\sum_{i=1}^{d} w_i\right) + b \\ \text{sujeito a} & \\ y'_i \left(\sum_{j=1}^{d} x_{ij} w_j + b\right) > 0 & \qquad i = 1, \ldots, n \end{array}$$

Pode ser mostrado (Exercício ???) que a formulação acima obtém uma solução para o problema mencionado, caso exista. Além disso, segundo condições estabelecidas, resolve o problema considerado pelo Algoritmo do Perceptron.

6.5 A Dimensão de Vapnik-Chervonenkis

O modelo de aprendizado de máquina visto até o momento, por exemplo o perceptron, se baseia no princípio de que há um conjunto de dados de uma certa aplicação, cujas variáveis se destinam a determinar uma resposta para cada exemplo. Supondo que a resposta seja binária, seria do tipo SIM ou NÃO. Estes dados se destinam ao treinamento do sistema. Isto é, para cada exemplo, já haveria uma resposta SIM ou NÃO, previamente conhecida. Mediante a aplicação da técnica de aprendizado de máquina, o objetivo é computar uma resposta calculada. Esta pode ser comparada com a resposta correta, que já seria conhecida, para aferir a qualidade do método de aprendizado.

Obviamente, a finalidade última seria não somente treinar o método de aprendizado, mas capacitá-lo a produzir respostas corretas.

Uma questão natural seria conhecer se o objetivo de capacitar o método de aprendizado para produzir as respostas corretas é realizável. Este objetivo pode ser alcançado, dentro de limites probabilísticos pré-estabelecidos. Para tal, utilizamos a Teoria de Vapnik-Chervonenkis, cujos princípios serão expostos nas subseções seguintes. Pela sua importância, essa teoria é considerada por muitos como a parte

central do aprendizado de máquina, a qual culmina com o Teorema de Vapnik-Chervonenkis.

6.5.1 A dimensão Vapnik-Chervonenkis

Um *sistema de conjuntos* $(U, \mathcal{S})$ consiste de um conjunto U, denominado *universo*, juntamente com uma família $\mathcal{S}$ de subconjuntos de U. Um subconjunto $C \subseteq U$ é dito *aniquilado* se cada subconjunto C' de C pode ser expresso como a interseção de C com algum subconjunto U' de $\mathcal{S}$:

$$C' = C \cap U'$$

Uma definição alternativa para conjuntos aniquilados é a seguinte. Em um sistema de conjuntos $(U, \mathcal{S})$, seja $C \subseteq U$. Considere a coleção formada pelas interseções de C com os subconjuntos de $\mathcal{S}$, denotada por:

$$\mathcal{S} \cap C = \{U' \cap C \mid U' \in \mathcal{S}\}$$

isto é, $\mathcal{S} \cap C$ é a coleção cujos elementos correspondem às interseções de C, com cada um dos subconjuntos $U' \in \mathcal{S}$. Nesse caso, C é um conjunto aniquilado se $\mathcal{S} \cap C$ contém precisamente todas as interseções de C com os subconjuntos de $\mathcal{S}$. Então,

$$|\mathcal{S} \cap C| = 2^{|C|}$$

A *dimensão Vapnik-Chervonenkis*, ou *dimensão VC*, é a cardinalidade do maior subconjunto de U que está aniquilado. Se esta cardinalidade não for limitada, a dimensão correspondente é ∞.

Em seguida, descrevemos alguns exemplos de sistemas de conjuntos e sua dimensão VC.

(i) Seja $U = \{1, 2, 3\}$ e $\mathcal{S} = \{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$.

Considere os seguintes subconjuntos C de U:

$C = \{3\} \quad \Rightarrow \quad C \cap \mathcal{S} = \{\varnothing, \{3\}\} \quad \Rightarrow \{3\}$ é aniquilado.

$C = \{2, 3\} \quad \Rightarrow \quad C \cap \mathcal{S} = \{\varnothing, \{2\}, \{3\}, \{2, 3\}\} \quad \Rightarrow \{2, 3\}$ é aniquilado.

$C = \{1, 2\} \quad \Rightarrow \quad C \cap \mathcal{S} = \{\{1\}, \{2\}, \{1, 2\}\} \quad \Rightarrow \{1, 2\}$ não é aniquilado.

$C = \{1, 2, 3\} \quad \Rightarrow \quad C \cap \mathcal{S} = \{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\} \quad \Rightarrow \{1, 2, 3\}$ não é aniquilado.

Observe que a cardinalidade do maior subconjunto de U que é aniquilado é 2. Logo, a dimensão VC do sistema $(\{1, 2, 3\}, \{\{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\})$ é 2.

(ii) Seja U o conjunto dos pontos de uma reta real, e $\mathcal{S}$, o conjunto de intervalos da reta.

Seja $C = \{a, b \,|a, b \in \mathbb{R}, a \neq b\}$.

C está aniquilado, pois existem intervalos distintos I_1, I_2, I_3, I_4, tais que:

$$I_1 \cap C = \varnothing, \quad I_2 \cap C = \{a\}, \quad I_3 \cap C = \{b\}, \quad I_4 \cap C = \{a, b\}$$

Por outro lado, qualquer conjunto C de 3 pontos $\{a, b, c\}$, com $a < b < c$, não pode ser aniquilado, pois qualquer intervalo que contenha, simultaneamente, a e c, necessariamente contém b.

Assim sendo, a dimensão VC de um conjunto de intervalos de uma reta é 2.

(iii) Seja U o conjunto dos pontos do plano cartesiano, $U = \mathbb{R}^2$, e $\mathcal{S}$ a família dos semi-planos $S \subseteq U$, isto é, delimitados por alguma reta em U.

Seja $C = \{a, b, c\} \subseteq U$, um conjunto de 3 pontos não colineares em U. Através do triângulo a, b, c, é fácil escolher semi-planos $S', S_a, S_b, S_c, S_{ab}, S_{ac}, S_{bc}, S_{abc}$ tais que

$$S' \cap C = \varnothing, \quad S_a \cap C = \{a\}, \quad S_b \cap C = \{b\}, \quad S_c \cap C = \{c\},$$

$$S_{ab} \cap C = \{a, b\}, \quad S_{ac} \cap C = \{a, c\}, \quad S_{bc} \cap C = \{b, c\}, \quad S_{abc} \cap C = \{a, b, c\}$$

Nesse caso, C é um conjunto aniquilado, pois

$$|C \cap \mathcal{S}| = 2^3 = 8$$

Examinemos, em seguida, um conjunto C, com pelo menos 4 pontos. Se C contém 3 pontos colineares, então C não pode ser aniquilado, pois qualquer semi-plano que contenha os pontos extremos conterá também o ponto central desses 3 pontos colineares. Suponha, agora, que não existam 3 pontos colineares. Escolher 3 pontos arbitrários de C, a, b, c, e formar um triângulo com esses pontos. Seja $d \neq a, b, c$ um outro ponto arbitrário de C. Se d for interior a esse triângulo, então C não poderá ser aniquilado, pois qualquer semi-plano que contenha d necessariamente conterá a, b ou c. Suponha, então, que a, b, c, d formem um quadrilátero convexo. Sejam b, c os vértices adjacentes a d, neste quadrilátero. Então qualquer semi-plano que contenha b, c, deve conter necessariamente a ou d. Logo, C não pode ser aniquilado. Isto é, a dimensão VC de semi-planos no $\mathbb{R}^2$ é < 4. Logo, é igual a 3.

(iv) Seja U o conjunto dos pontos de um círculo em um plano. Seja $\mathcal{S}$ a família dos polígonos convexos do plano formados por pontos de U. Qualquer subconjunto

dos pontos de U define um polígono convexo, basta considerá-los em ordem consecutiva no círculo. Seja C um polígono convexo formado por n pontos de U, e C' um polígono convexo formado por um subconjunto desses n pontos. Existe um polígono convexo $U' \in \mathcal{S}$, tal que os pontos de C' podem ser obtidos exatamente como interseção de U' e C. Consequentemente,

$$|\mathcal{S} \cap C| = 2^{|C|}$$

Logo, a dimensão VC de $(U, \mathcal{S})$ é infinita.

6.5.2 A Função de Aniquilamento

A *função de aniquilamento* de um sistema $(U, \mathcal{S})$ de conjuntos é a função que relaciona um inteiro n à quantidade máxima de subconjuntos de $C \subset U$, $|C| = n$, da forma $S \cap C$, para $S \in \mathcal{S}$. Ou seja, de acordo com a definição da dimensão VC, sabemos que se $C \subset U$ é um conjunto aniquilado, então $\mathcal{S} \cap C$ contém todas as intersecções de C com os subconjuntos de $\mathcal{S}$. Isto é, $|\mathcal{S} \cap C| = 2^{|C|}$. Então, a dimensão VC de $(U, \mathcal{S})$ é a máxima cardinalidade de algum subconjunto $C \subseteq U$, tal que $|\mathcal{S} \cap C| = 2^{|C|}$. Consideramos, agora, uma variável n para exprimir a cardinalidade de subconjuntos $C \subseteq U$, isto é, $|C| = n$. Seja d o valor da dimensão VC de $(U, \mathcal{S})$. Então, para $n = |C| \leq d$, o número máximo de subconjuntos de C que podem ser aniquilados é 2^n. Pela definição de dimensão VC, não há subconjunto C de U que possa ser aniquilado quando $|C| > d$. Nesse caso, o interesse seria determinar, para cada n, o número máximo de subconjuntos de um conjunto $C \subseteq U$ que podem ser aniquilados. Sabemos que esse valor é $\leq 2^n$. A função de aniquilamento visa estudar a relação entre n e este número máximo. Sabemos que na faixa $0 \leq n \leq d$, o número máximo de subconjuntos que podem ser aniquilados é exponencial em n, da forma 2^n. Pode ser provado, contudo, que, a partir do valor $n > d$, este número máximo passa a ser um polinômio de grau d, em n, isto é, a função de aniquilamento pode ser representada por um gráfico da forma da Figura ??? (fazer).

Assim, a função de aniquilamento $\Pi_{\mathcal{S}}(n)$ de um sistema de conjuntos $(U, \mathcal{S})$ é o valor máximo de subconjuntos de algum conjunto $C \subseteq U$, $|C| = n$, que podem ser aniquilados, isto é, que correspondam a interseções de subconjuntos de $\mathcal{S}$. Ou seja,

$$\Pi_{\mathcal{S}}(n) = \max_{\substack{C \subseteq U \\ |C| = n}} \left| \{ C \cap S \mid S \in \mathcal{S} \} \right|$$

Como exemplo, já foi observado que a dimensão VC de um sistema $(U, \mathcal{S})$, onde $U = \mathbb{R}^2$ e $\mathcal{S}$ é a família dos semi-planos $S \subseteq U$, é igual a 3. Nesse caso,

$$\Pi_{\mathcal{S}}(n) = 2^n, \quad \text{para} \quad n \leq 3,$$

e, a partir de $n > 3$, o crescimento de $\Pi_{\mathcal{S}}(n)$ é mais lento.

Para sistemas de conjuntos $(U, \mathcal{S})$, cuja dimensão d seja limitada, vale o seguinte resultado.

Lema 6.2. *Para qualquer sistema de conjuntos $(U, \mathcal{S})$, de dimensão VC d limitada, para todo n vale*

$$\Pi_{\mathcal{S}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq 2n^d$$

O lema anterior implica que, para valores $n > d$, o crescimento de $\Pi_{\mathcal{S}}(n)$ é limitado por um polinômio em n, de grau d.

Naturalmente, se d for infinito, o valor de $\Pi_{\mathcal{S}}(n)$ é igual a 2^n, para todo n.

Em seguida, examinaremos a função de aniquilamento de interseções de sistemas de conjuntos.

6.5.3 Interseção de Sistemas de Conjuntos

Sejam os sistemas de conjuntos $(U, \mathcal{S}_1)$, $(U, \mathcal{S}_2)$, que compartilham o mesmo conjunto universo U. O *sistema de interseção* $(U, \mathcal{S}_1 \cap \mathcal{S}_2)$ é definido como aquele em que os subconjuntos são formados pelas interseções de todos os pares de subconjuntos $S_1 \cap S_2$, $S_1 \in \mathcal{S}_1$ e $S_2 \in \mathcal{S}_2$, isto é:

$$\mathcal{S}_1 \cap \mathcal{S}_2 = \left\{ S_1 \cap S_2 \;\mid\; S_1 \in \mathcal{S}_1, S_2 \in \mathcal{S}_2 \right\}$$

O seguinte teorema relaciona as funções de aniquilamento de dois sistemas de conjuntos, com a função de aniquilamento do respectivo sistema de interseção.

Teorema 6.3. *Sejam $(U, \mathcal{S}_1)$, $(U, \mathcal{S}_2)$ dois sistemas de conjuntos com o mesmo conjunto universo. Então*

$$\Pi_{\mathcal{S}_1 \cap \mathcal{S}_2}(n) \leq \Pi_{\mathcal{S}_1}(n) \Pi_{\mathcal{S}_2}(n)$$

Demonstração. (Revisar) Seja um subconjunto $C \subseteq U$. Suponha que C seja aniquilado por $(U, \mathcal{S}_1 \cap \mathcal{S}_2)$. Então, para cada subconjunto $C' \subseteq C$, existe um subconjunto $U' \in \mathcal{S}_1 \cap \mathcal{S}_2$, tal que $C' = C \cap U'$. Isto implica que $U' \in \mathcal{S}_1$ e $U' \in \mathcal{S}_2$. Então C será também aniquilado por ambos $(U, \mathcal{S}_1)$ e $(U, \mathcal{S}_2)$. Suponha, agora, que C seja aniquilado por $(U, \mathcal{S}_1)$, mas não por $(U, \mathcal{S}_2)$. Seja $C' \subseteq C$. Então existe subconjunto $U' \in \mathcal{S}_1 \setminus \mathcal{S}_2$, tal que $C' = C \cap U'$. Isto implica que C não será aniquilado por $(U, \mathcal{S}_1 \cap \mathcal{S}_2)$. Logo, $\Pi_{\mathcal{S}_1 \cap \mathcal{S}_2}(n) \leq \Pi_{\mathcal{S}_1}(n) \Pi_{\mathcal{S}_2}(n)$. ■

6.6 O Teorema de Vapnik-Chervonenkis

Dado um sistema de subconjuntos $(U, \mathcal{S})$, o Teorema de Vapnik-Chervonenkis (Teorema VC) visa determinar a quantidade mínima de amostras necessárias a serem selecionadas de U, com uma dada probabilidade p de tal forma que, para cada $S \in \mathcal{S}$, a fração de amostras em S seja aproximadamente $p(S)$.

Teorema 6.4. *Seja $(U, \mathcal{S})$ um sistema de subconjuntos de dimensão VC igual a d. Considere uma probabilidade arbitrária p de distribuição dos objetos de U, e $U' \subseteq U$ um conjunto de amostras retiradas de U, de acordo com a probabilidade p. Para qualquer $\epsilon \in (0,1)$, se $n = \Omega(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon})$, então*

$$P\left(\exists S \in \mathcal{S} \mid \left|\frac{|S \cap U'|}{n} - p(S)\right| > \epsilon\right) \leq 2\mathrm{e}^{-\epsilon^2 n/6}$$

Em seguida, detalhamos o significado do teorema.

Consideramos um conjunto universo U e uma família $\mathcal{S}$ de subconjuntos de U. Seja p uma probabilidade arbitrária dos objetos de U, isto é, ao selecionarmos objetos de U, a seleção será realizada segundo a probabilidade p. Cada subconjunto $S \in \mathcal{S}$ possui, então, probabilidade $p(S)$. Escolhemos um subconjunto $U' \subseteq U$ de amostras. A quantidade destas amostras presentes em cada $S \in \mathcal{S}$ é, pois, $|S \cap U'|$. Para uma aproximação desejada $\epsilon \in (0,1)$ e um valor $n = \Omega(\frac{d}{\epsilon^2} \log \frac{d}{\epsilon})$, admitimos que o valor $\left|\frac{|S \cap U'|}{n} - p(S)\right|$ possa ser superior a ϵ, mas nesse caso, a probabilidade de que isso ocorra seja pequena, limitada a $2\mathrm{e}^{-\epsilon^2 n/6}$.

6.7 Regressão Linear

Consideremos um conjunto de dados da forma (x_i, y_i), $1 \leq i \leq n$, que podem corresponder a um conjunto de observações ou medidas. Por exemplo, o valor x_i pode corresponder ao consumo de carnes pelo indivíduo i e y_i ao nível de colesterol deste indivíduo. O nosso objetivo é determinar uma relação entre o valor x_i e o valor y_i. Através dessa relação, o valor y_i pode ser determinado em função do valor de x_i, o que possibilitaria sua previsão. Esta técnica se constitui, portanto, em mais um instrumento de aprendizagem de máquina.

A relação mais simples entre os valores

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

é a relação linear. Nesse caso, a tentativa é que cada par (x_i, y_i) corresponda às coordenadas de um ponto de uma reta. Então, a relação procurada será expressa por uma equação do tipo

$$y = ax + b$$

A questão, pois, será determinar os parâmetros a e b, de tal forma que os valores (x_i, y_i) possam satisfazer

$$y_i = ax_i + b$$

Naturalmente, será aceitável que os pontos (x_i, y_i) não estejam exatamente sobre a reta obtida $y = ax + b$. Mas, nesse caso, deverão estar próximos à reta. Ou seja, os parâmetros a, b a serem calculados introduzem um erro, que será denotado por $E(a, b)$.

O nosso objetivo é o de determinar os parâmetros a, b, a partir dos pares (x_i, y_i), com $1 \leq i \leq n$, de tal forma que o erro $E(a, b)$, introduzido ao se considerar a reta $y = ax + b$ como representação da relação entre os pares (x_i, y_i), seja minimizado de alguma forma.

A questão, agora, se torna a de escolher, exatamente, a função mais adequada à minimização.

Note que, o erro $E_i(a, b)$ cometido ao assumir para o valor y_i aquele obtido pela reta $y = ax + b$ é igual a

$$E_i(a, b) = y_i - (ax_i + b)$$

Portanto, o erro total $E(a, b)$ é igual a

$$E(a, b) = \sum_{i=1}^{n} [y_i - (ax_i + b)]$$

Contudo, a minimização da função $E(a, b)$ anterior é inadequada para os nossos propósitos. Uma razão importante para não adotar esta expressão como a função objetivo a ser minimizada é que os erros em cada ponto são grandezas que admitem sinal, positivo ou negativo. Assim, erros negativos poderiam compensar erros positivos, o que poderia tornar o resultado final falso. Uma ideia, então, seria a de minimizar o módulo do erro de cada ponto. Nesse caso, a função de minimização seria:

$$E(a, b) = \sum_{i=1}^{n} |y_i - (ax_i + b)|$$

A alternativa de minimizar esta última função é também inadequada. O motivo é que tornaria o cálculo da minimização mais complicado, pois o módulo de uma função a torna não diferenciável, o que praticamente elimina a possibilidade de utilizar o cálculo diferencial para a determinação do valor mínimo de $E(a, b)$.

Finalmente, vamos considerar a alternativa de minimizar a soma dos quadrados dos erros de cada ponto. Esses valores são todos positivos, e a função a ser minimizada é diferenciável. Essa técnica é bastante difundida e recebe o nome de Método dos Mínimos Quadrados. Será examinada na próxima subseção.

6.7.1 O Método dos Mínimos Quadrados

Seja um conjunto de dados da forma

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

onde os valores x_i são designados como variáveis independentes e os valores y_i, variáveis dependentes. O objetivo é ajustar uma relação entre essas variáveis. No caso, esta relação corresponde a uma função linear da forma

$$y = ax + b$$

ou seja, uma reta no $\mathbb{R}^2$. Para tal, necessitamos determinar os parâmetros a e b. De um modo geral, os dados (x_i, y_i), $1 \leq i \leq n$ correspondem a pontos no $\mathbb{R}^2$, e esses pontos não se distribuem exatamente em uma reta. Então, o objetivo é o de encontrar a "melhor" reta, ou seja, determinar os parâmetros a e b que conduzem à reta que representará o conjunto de dados (x_i, y_i). Naturalmente, a reta ajustada será uma aproximação, em princípio não passará exatamente sobre os pontos (x_i, y_i). Assim, em lugar de passar pelo ponto (x_i, y_i), a reta passará pelo ponto $(x_i, ax_i + b)$. O erro cometido, em relação ao ponto é da forma $E_i(a, b) = y_i - (ax_i + b)$.

Conforme mencionado na subseção anterior, o objetivo é encontrar a reta que minimize a soma dos quadrados dos erros. Isto é, a e b serão determinados de modo a minimizar

$$E(a, b) = \sum_{i=1}^{n} E_i^2(a, b) = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

Para encontrar os valores de a e b que minimizam a função anterior, empregamos o cálculo diferencial. Para tal, devemos encontrar a derivada parcial $\frac{\partial f}{\partial a} E(a, b)$, relativa a a, bem como a derivada parcial $\frac{\partial f}{\partial b} E(a, b)$, relativa a b. Os pontos minimizantes são exatamente aqueles que anulam essas derivadas parciais.

Os valores das derivadas parciais são:

$$\frac{\partial E}{\partial a} = 2 \sum_{i=1}^{n} [y_i - (ax_i + b)].(-x_i)$$

$$\frac{\partial E}{\partial b} = 2 \sum_{i=1}^{n} [y_i - (ax_i + b)].(-1)$$

Igualando as derivadas a zero, obtemos

$$\frac{\partial E}{\partial a} = 0 \quad \Rightarrow \sum_{i=1}^{n} [y_i - (ax_i + b)]x_i = 0$$

$$\frac{\partial E}{\partial b} = 0 \quad \Rightarrow \sum_{i=1}^{n} [y_i - (ax_i + b)] = 0$$

Considerando, inicialmente, a equação $\frac{\partial E}{\partial b} = 0$,

$$\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

Dividindo por n

$$\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{a}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} b = 0$$

Representando por $\overline{x}$ e $\overline{y}$, respectivamente, as médias aritméticas dos valores x_i e y_i, isto é,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{e} \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

obtemos

$$\overline{y} - a\overline{x} - b = 0$$

Logo,

$$b = \overline{y} - a\overline{x}$$

Resta determinar o valor do parâmetro a, da reta. Substituindo o resultado de b, na equação $\frac{\partial E}{\partial a} = 0$, obtemos

$$\sum_{i=1}^{n} [y_i - (ax_i + \overline{y} - a\overline{x})]x_i = 0$$

$$\sum_{i=1}^{n} [x_i(y_i - \overline{y}) + ax_i(\overline{x} - x_i)] = 0$$

$$\sum_{i=1}^{n} [x_i(y_i - \overline{y})] + a \sum_{i=1}^{n} x_i(\overline{x} - x_i) = 0$$

Logo,

$$a = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}$$

Com os valores de a e b calculados podemos determinar a reta $y = ax + b$, que minimiza a soma dos quadrados dos erros, conforme desejado.

Observamos que a expressão anterior somente é válida se $\sum_{i=1}^{n} x_i(x_i - \overline{x})$. Isto equivale à condição de que os valores x_i não sejam todos idênticos, isto é, $x_i \neq x_j$, para algum par i, j, com $1 \leq i, j \leq n$.

O método descrito conduz ao seguinte algoritmo para determinar a reta que minimiza a soma dos quadrados dos erros, em relação aos dados fornecidos.

Os dados compreendem os vetores $X = \{x_i\}$ e $Y = \{y_i\}$, $1 \leq i \leq n$, com a condição de que $x_i \neq x_j$ para algum par i, j. O algoritmo para ajustar uma reta $y = ax + b$ aos dados correspondentes é o seguinte, que retorna os valores de a e b calculados.

Algoritmo 6.2 Regressão linear

Dados: vetores $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$

$\overline{x} \leftarrow \left(\sum_{i=1}^{n} x_i\right)/n$; $\overline{y} \leftarrow \left(\sum_{i=1}^{n} y_i\right)/n$

$a \leftarrow \left(\sum_{i=1}^{n} x_i(y_i - \overline{y})\right) / \left(\sum_{i=1}^{n} x_i(x_i - \overline{x})\right)$

$b \leftarrow \overline{y} - a\overline{x}$

retornar a, b

É imediato verificar que o algoritmo possui complexidade $O(n)$. A sua correção também é imediata, pois o algoritmo se resume a calcular os valores de a e b, segundo a fórmula descrita.

Como exemplo, suponha que se deseja avaliar a temperatura y de uma certa massa x. Sabe-se que a temperatura depende dessa massa e varia diretamente com ela. Para tal, são efetuadas medições, cujos valores da temperatura y_i e da massa correspondente foram os ilustrados na Tabela 6.1.

i	x_i	y_i
1	5	-2
2	8	0
3	10	3
4	12	5
5	13	7
6	15	8
7	20	10
8	25	13
9	30	15
10	34	18

Tabela 6.1: Valores da temperatura y_i em função da massa x_i

Para ajustar uma reta correspondente a esses pontos, aplicamos a técnica dos mínimos quadrados. A equação da reta é $y = ax + b$, onde

$$a = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} \quad \text{e} \quad b = \overline{y} - a\overline{x},$$

e as médias $\overline{x}$ e $\overline{y}$, respectivamente das medidas x_i e y_i são iguais a

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 17{,}2 \qquad \overline{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 7{,}7$$

Logo,

$$a = \frac{553{,}6}{849{,}6} = 0{,}65 \qquad b = 7{,}7 - 0{,}65 \cdot 17{,}2 = -3{,}51$$

A equação $y = ax + b$ da reta ajustada é $y = 0{,}65x - 3{,}51$.

Veja a Figura ? (fazer figura; checar com dados novos)

A regressão linear no $\mathbb{R}^2$, conforme abordada nesta subseção é conhecida pela denominação *regressão linear simples*. Quando a quantidade de variáveis independentes é maior do que 1, a denominação passa a ser *regressão linear múltipla*. Nesse caso, ao invés de tentar ajustar uma reta, a tentativa é de ajuste de um hiperplano, de mesma dimensão que a quantidade de variáveis independentes. Pode-se aplicar técnica similar à utilizada para a regressão linear simples.

6.8 Exercícios

6.1 Considere o problema de, em um conjunto de pacientes, separar aqueles que uma certa doença têm daqueles que não têm. Modelar este problema de maneira similar àquela apresentada na introdução do capítulo.

6.2 Generalizar o resultado de que a dimensão VC de semi-planos no $\mathbb{R}^2$ é igual a 3. Provar, então, que a dimensão VC de semi-hiperplanos no $\mathbb{R}^d$ é $d + 1$.

6.3 Seja U o conjunto dos pontos em um plano, e $\mathcal{S}$ o conjunto de retângulos deste plano de lados paralelos aos eixos. Mostrar que a dimensão VC de $(U, \mathcal{S})$ é igual a 4.

6.4 Seja U o conjunto de pontos de um plano, e $\mathcal{S}$ o conjunto de todos os subconjuntos finitos de U. Mostrar que a dimensão VC de $(U, \mathcal{S})$ é igual a ∞.

6.5 Seja U a circunferência de um círculo, e $\mathcal{S}$ o conjunto de k arcos deste círculo em U. Demonstrar que a dimensão VC de $(U, \mathcal{S})$ é igual a $2(k + 1)$.

6.6 Sejam $(U, \mathcal{S}_1)$ e $(U, \mathcal{S}_2)$ dois sistemas de subconjuntos, partilhando o mesmo conjunto universo U. Provar ou dar contra-exemplo para a afirmação:

"(dimensão VC de $\mathcal{S}_1 \leq$ dimensão VC de $\mathcal{S}_2$) $\Leftrightarrow \mathcal{S}_1 \subseteq \mathcal{S}_2$".

6.7 Obter a reta correspondente à regressão linear considerando apenas as 5 primeiras medições da Tabela 6.1.

6.8 Mostrar que a expressão

$$a = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}$$

utilizada na determinação da reta de ajuste na regressão linear é válida se e somente se os valores $x_1, \ldots, x_n$ não são todos idênticos.

6.9 O método de regressão linear foi apresentado para o caso no $\mathbb{R}^2$. Generalizar para k dimensões, em que o vetor X se encontra no $\mathbb{R}^k$.

6.10 Dados os vetores $X = \{x_i\}$ e $Y = \{y_i\}, 1 \leq i \leq n$, ajustar entre esses dados uma equação do segundo grau do tipo

$$y = ax^2 + bx + c,$$

utilizando uma técnica similar à de minimização da soma dos quadrados dos erros cometidos.