Aplicação de Transformações Lineares: Sistema de Auxílio ao Diagnóstico Médico

Guilherme de Alencar Barreto

gbarreto@ufc.br

Departamento de Engenharia de Teleinformática (DETI)
Engenharias de Computação, Telecomunicações e Teleinformática
Universidade Federal do Ceará – UFC
www.researchgate.net/profile/Guilherme_Barreto2/

Conteúdo dos Slides

Transformadas Matriciais

Para cada $\mathbf{x} \in \mathbb{R}^n$, uma transformada T é linear (ou matricial) quando for escrita como

$$\mathbf{y} = T(\mathbf{x}) = \mathbf{W}\mathbf{x} \quad (\text{ou} \quad \mathbf{W}\mathbf{x} = \mathbf{y}), \tag{1}$$

em que $\mathbf{W}$ é uma matriz $m \times n$.

- Para simplificar, muitas vezes denotamos essa transformação matricial por

$$\boxed{\mathbf{x} \mapsto \mathbf{W}\mathbf{x}} \tag{2}$$

- Note que o domínio de T é o $\mathbb{R}^n$ quando $\mathbf{W}$ tem n colunas, e o contra-domínio de T é o $\mathbb{R}^m$ quando cada coluna de $\mathbf{W}$ tem m elementos.

Transformadas Matriciais

- A transformação linear $\mathbf{y} = \mathbf{W}\mathbf{x}$ produz um vetor de saída $\mathbf{y}$ a partir de um vetor de entrada $\mathbf{x}$ por meio da matriz $\mathbf{W}$.
- Etmologicamente, o termo *vetor* é um substantivo masculino que significa condutor ou portador. Do latim "vectore".
- Em reconhecimento de padrões, um vetor é o portador de informação sobre o objeto a ser, por exemplo, classificado.
- Etmologicamente, o termo *matriz* é um substantivo feminino que significa útero. Portanto, dá a entender como aquilo que gera, determina, algum resultado.

Diagrama de Blocos

Ajuda muito no entendimento de uma transformação linear se representarmos a relações $\mathbf{y} = \mathbf{W}\mathbf{x}$ na forma de um diagrama de blocos do tipo entrada-saída.

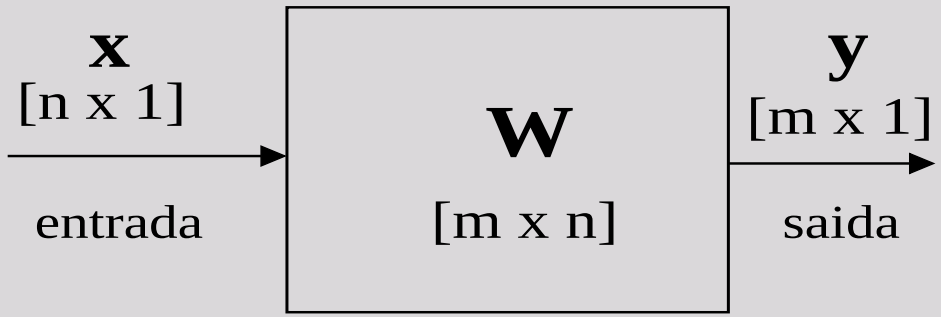


Diagrama de Blocos

Em função das dimensões n e m dos vetores de entrada e saída, respectivamente, temos os seguintes tipos de sistemas:

- $n > 1$ e $m > 1$: Sistemas **MIMO** (multi-input, multi-output).
- $n > 1$ e $m = 1$: Sistemas **MISO** (multi-input, single-output).
- $n = 1$ e $m = 1$: Sistemas **SISO** (single-input, single-output).
- $n = 1$ e $m > 1$: Sistemas **SIMO** (single-input, multi-output).

Transformadas Matriciais

- Um vetor é um segmento de reta que tem comprimento (norma) e orientação (ângulo com a horizontal).
- Dado um vetor $\mathbf{x} = [x_1\ x_2\ \cdots\ x_n]^T$, sua norma é dada por

$$r = \|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{3}$$

- Para calcular o ângulo θ com a horizontal precisamos da definição de produto escalar entre 2 vetores:

$$\cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|} \quad \Rightarrow \quad \theta = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|}\right) \tag{4}$$

em que $\mathbf{v} = [v_1\ v_2\ \cdots\ v_n]^T$ é um vetor de referência e $\mathbf{x} \cdot \mathbf{v} = \sum_{i=1}^{n} x_i v_i$.

Transformadas Matriciais

- Por exemplo, o ângulo entre o vetor $\mathbf{x} = [1\ 1]^T$ e o eixo horizontal pode ser calculado como

$$\theta = \arccos\left(\frac{(1).(1) + (1).(0)}{\sqrt{2} \times (1)}\right) = \arccos\left(\frac{1}{\sqrt{2}}\right) = 45^o \quad (5)$$

em que $\mathbf{v} = [1\ 0]^T$ é o vetor de referência escolhido.

Transformadas Matriciais

- Qual o ângulo entre o vetor $\mathbf{x} = [1\ 2\ 3]^T$ e o plano horizontal?
- Solução: Neste caso, o vetor de referência é $\mathbf{v} = [1\ 2\ 0]^T$. Assim, temos que

$$\theta = \arccos\left(\frac{(1).(1)+(2).(2)+(3).(0)}{\sqrt{14}\times\sqrt{5}}\right) \quad (6)$$

$$= \arccos\left(\frac{5}{\sqrt{70}}\right) \quad (7)$$

$$\approx 53^o \quad (8)$$

Transformadas Matriciais

- Transformações lineares alteram a norma e/ou a orientação de vetores.
- Se a matriz $\mathbf{W}$ é uma matriz identidade, nem a norma nem a orientação são alteradas.
- Por exemplo, para $\mathbf{x} = [1\ 2\ 3]^T$, temos que

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \mathbf{x} \tag{9}$$

Transformadas Matriciais

- Se a matriz $\mathbf{W}$ é múltipla da matriz identidade, i.e. $\mathbf{W} = \alpha\mathbf{I}$, $\alpha \in \mathbb{R}$, só a norma é alterada.
- Por exemplo, para $\mathbf{x} = [1\ 2\ 3]^T$, temos que

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \alpha\mathbf{I}\mathbf{x} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \alpha \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} \alpha \\ 2\alpha \\ 3\alpha \end{bmatrix} = \alpha\mathbf{x} \qquad (10)$$

daí norma($\mathbf{y}$)=$\|\mathbf{y}\| = \|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$.

Transformadas Matriciais

- As chamadas matrizes de rotação, promovem mudanças apenas na orientação do vetor, sem alterar sua norma.
- Por exemplo, para $\mathbf{x} = [1\ 0]^T$, temos que

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \tag{11}$$

- Essa matriz faz com que o vetor $[1\ \ 0]^T$ girasse no sentido anti-horário de 45^o. As normas de $\mathbf{x}$ e $\mathbf{y}$ são iguais a 1.
- Qual é a matriz de rotação que gira um vetor no sentido anti-horário em um ângulo de θ graus no plano?

Transformadas Matriciais

- A seguinte matriz $\mathbf{W}$ transforma o vetor $\mathbf{x}$ de forma a rotacioná-lo de 45^o e alterar a sua norma por um fator α.

$$\mathbf{W} = \begin{bmatrix} \alpha\frac{\sqrt{2}}{2} & 0 \\ \alpha\frac{\sqrt{2}}{2} & \alpha \end{bmatrix} \tag{12}$$

- Uma transformação linear pode ser decomposta em outras duas, uma de rotação $\mathbf{W}_1$ e outra de mudança da norma $\mathbf{W}_2$, e aplicada sequencialmente e em qualquer ordem.

$$\mathbf{W} = \begin{bmatrix} \alpha\frac{\sqrt{2}}{2} & 0 \\ \alpha\frac{\sqrt{2}}{2} & \alpha \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{2}}{2} & 1 \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} = \mathbf{W}_1\mathbf{W}_2 = \mathbf{W}_2\mathbf{W}_1 \tag{13}$$

Definição do Problema

- Considere que um médico tem que diagnosticar a doença de pele de um certo paciente com base em
  - **Informações clínicas**: informações coletadas pelo médico durante a *anamnese* e inspeção visual da pele no consultório.
  - **Informações histopatológicas**: normalmente resultam de uma biópsia, ou seja, da análise do tecido em um laboratório de patologia.

Doenças de Pele Envolvidas no Problema

Após um período, tal médico coletou tais informações sobre 358 pacientes e suas respectivas patologias.

Doença (número de pacientes)
Psoríase(111)
Dermatite seborréica(60)
Líquen plano(71)
Pitiríase rósea(48)
Dermatite crônica(48)
Pitiríase rubra pilar(20)

Chute Aleatório vs. Chute Informado

- Falando em tomada de decisão, é importante destacar duas formas bem simples de se tomar uma decisão.
- São elas: o chute aleatório (*random guess*) e o chute informado (*informed guess*), este também chamado de *educated guess*
- Um chute aleatório envolve a escolha aleatória de qualquer uma das classes do problema para alocar um novo objeto.
- Nada mais é do que "jogar uma moeda justa" (duas classes) ou um dado honesto (Mais que 2 classes).
- Para o problema ora tratado, corresponde a selecionar uma das 6 classes aleatoriamente.
- Neste caso, a chance de acertar é de $1/6 = 16{,}67\%$.

Chute Informado (*Informed/Educated Guess*)

- Um chute informado envolve sempre a escolha da classe com maior probabilidade *a priori* dentre as classes existentes.
- Para um problema com C classes, a probabilidade a priori da i-ésima classe é dada por $P(\omega_i) = N_i/N$, $i = 1, \ldots, C$, em que N_i é o número de exemplos da classe ω_i e N é o número total de exemplos.
- Para o problema ora tratado, a classe com maior probabilidade a priori é a classe $\omega_1 =$ Psoríase, ou seja

$$P(\omega_1) = \frac{111}{358} = 0,31 \;\; (31\%) \tag{14}$$

- Isto significa que a cada 100 chutes, irá acertar 31 casos.

Informação Advinda dos Atributos

- Tanto o chute aleatório quanto o informado só utilizam os rótulos para construir uma estratégia de tomada de decisão.
- Um classificador mais elaborado deve também utilizar informação provida pelas variáveis de entrada, chamadas de atributos, no contexto de classificação de padrões.
- No escopo do problema ora tratado, os atributos carregam informação sobre o estado do paciente.

Informações de Natureza Clínica

Clìnicos
1: eritema
2: escala
3: bordas definidas
4: coceira
5: fenômeno de Koebner
6: pápulas poligonais
7: pápulas foliculares
8: envolvimento da mucosa oral
9: envolvimento do joelho e do cotovelo
10: envolvimento do escalpo
11: histórico familiar
34: idade

Informações de Natureza Histopatológica

Histopatológicos
12: incontinência de melanina
13: eosinófilos no infiltrado
14: infiltrado PNL
15: fibrose na derme papilar
16: exocitose
17: acantose
18: hiperceratose
19: paraceratose
20: dilatação em clava dos cones epiteliais
21: alongamento dos cones epiteliais da epiderme
22: estreitamento da epiderme suprapapilar
23: pústulas espongiformes
24: microabscesso de Munro
25: hipergranulose focal
26: ausência da camada granulosa
27: vacuolização e destruição da camada basal
28: espongiose
29: aspecto "dente de serra" das cristas interpapilares
30: tampões cárneos foliculares
31: paraceratose perifolicular
32: infiltrado inflamatório mononuclear
33: infiltrado em banda

Banco de Dados dos Pacientes

- Cada medida clínica ou histopatológica pode ser entendida como uma variável que o médico usa para guiar sua decisão (diagnóstico).
- O médico organiza em um computador as informações de cada um dos 358 pacientes e o valor numérico correspondente de cada medida clínica ou histopatológica.
- De posse deste banco de dados, usando a Álgebra Linear é possível desenvolver um sistema computacional capaz de "diagnosticar" as seis doenças doenças de pele descritas anteriormente, de modo semelhante ao dermatologista!
- Para isso, precisamos formular o problema de diagnóstico médico como uma transformação linear $\mathbf{y} = \mathbf{W}\mathbf{x}$.

Vetor de Atributos

- Cada paciente vai ser representado por um vetor de dimensão $n = 34$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{32} \\ x_{33} \\ x_{34} \end{bmatrix} = \begin{bmatrix} \text{eritema} \\ \text{escala} \\ \text{bordas definidas} \\ \vdots \\ \text{infiltrado inflamatório mononuclear} \\ \text{infiltrado em banda} \\ \text{idade} \end{bmatrix} \quad (15)$$

- Para este problema, $x_j \in \{0, 1, 2, 3\}$, $j = 1, 2, \ldots, 33$.
- Somente a variável x_{34} assume valores maiores que 3.

Codificação da Saída

- Cada paciente possui um rótulo numérico de dimensão ($m = 6$) para a sua patologia, que é um identificador (ID) da patologia.

$$\text{Psoríase: } \mathbf{y} = \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}, \quad \text{Derm. Seborréica: } \mathbf{y} = \begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}, \quad \text{Líquen Plano:} \mathbf{y} = \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix} \tag{16}$$

$$\text{Pitiríase rósea: } \mathbf{y} = \begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix}, \quad \text{Derm. crônica: } \mathbf{y} = \begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix}, \quad \text{Pitiríase rubra pilar:} \mathbf{y} = \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix} \tag{17}$$

- Esta codificação assume que as classes são mutualmente exclusivas, pois os vetores-código são ortogonais entre si.
- No lugar de "0", pode-se usar "-1".

Formalização Matemática do Problema

- Note que no banco de dados teremos $N = 358$ vetores $\mathbf{x}_k \in \mathbb{R}^{34}$ e 358 vetores $\mathbf{y}_k \in \mathbb{R}^6$, $k = 1, \ldots, 358$, representando 358 pacientes e suas respectivas patologias.
- O índice k denota o k-ésimo paciente no banco de dados.
- Note que o objetivo é determinar uma matriz $\mathbf{W}$ que para um dado vetor de entrada (paciente) $\mathbf{x}_k$ forneça uma predição do vetor-código associado à patologia correspondente:

$$\mathbf{y}_k = \mathbf{W}\mathbf{x}_k, \quad \forall k = 1, \ldots, N = 358 \tag{18}$$

- Note que a matriz $\mathbf{W}$, de dimensões 6×34, atua como se fosse uma versão matemática do médico especialista.

Formalização Matemática do Problema

- Para facilitar, podemos organizar os $N = 358$ pacientes e os vetores-código de suas patologias nas colunas das matrizes $\mathbf{X}$ e $\mathbf{Y}$, dadas por:

$$\mathbf{X} = [\mathbf{x}_1 \mid \mathbf{x}_2 \mid \cdots \mid \mathbf{x}_{358}] \tag{19}$$

e

$$\mathbf{Y} = [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_{358}] \tag{20}$$

- Note que a matriz $\mathbf{X}$ tem dimensões 34×358 e a matriz $\mathbf{Y}$ tem dimensões 6×358.

Formalização Matemática do Problema

- A versão matricial da transformação $\mathbf{y}_k = \mathbf{W}\mathbf{x}_k$ pode ser obtida a partir da Eq. (20):

$$\begin{aligned} \mathbf{Y} &= [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_{358}], && (21)\\ &= [\mathbf{W}\mathbf{x}_1 \mid \mathbf{W}\mathbf{x}_2 \mid \cdots \mid \mathbf{W}\mathbf{x}_{358}], && (22)\\ &= \mathbf{W}[\mathbf{x}_1 \mid \mathbf{x}_2 \mid \cdots \mid \mathbf{x}_{358}] && (23) \end{aligned}$$

  que leva à seguinte expressão que envolve apenas matrizes: $\mathbf{Y}_{[6\times 358]} = \mathbf{W}_{[6\times 34]}\mathbf{X}_{[34\times 358]}$.
- As matrizes $\mathbf{X}$ e $\mathbf{Y}$ são montadas a partir do banco de dados de pacientes, enquanto a matriz $\mathbf{W}$, é desconhecida.
- Como a matriz $\mathbf{X}_{[34\times 358]}$ é retangular, não é possível obter sua inversa a fim de isolar a matriz $\mathbf{W}$ na expressão acima.

Formalização Matemática do Problema

- A fim de isolar a matriz $\mathbf{W}$, vamos usar de um artifício baseado apenas nas dimensões das matrizes $\mathbf{Y}$, $\mathbf{W}$ e $\mathbf{X}$.
- Se a matriz $\mathbf{X}$ fosse quadrada, poderíamos obter sua inversa e isolar a matriz $\mathbf{W}$.
- A grande "sacada" do artifício está em manipular (ou atuar sobre) a matriz $\mathbf{X}$ a fim de obter uma matriz quadrada.
- Para isso, vamos multiplicar (pela direita) ambos os lados da equação acima pela matriz $\mathbf{X}^T$, que é a matriz transposta de $\mathbf{X}$, obtendo a seguinte expressão:

$$\mathbf{Y}_{[6\times 358]}\mathbf{X}^T_{[358\times 34]} = \mathbf{W}_{[6\times 34]}\mathbf{X}_{[34\times 358]}\mathbf{X}^T_{[358\times 34]} \qquad (24)$$

Formalização Matemática do Problema

- Com isso, percebemos que a matriz $\mathbf{X}\mathbf{X}^T$ é quadrada, de dimensão 34×34, podendo assim ser invertida.
- Multiplicando ambos os lados da equação (pela direita) por $(\mathbf{X}\mathbf{X}^T)^{-1}$, obtemos:

$$\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} = \mathbf{W}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \tag{25}$$

- De onde resulta a seguinte expressão para cálculo da matriz de transformação $\mathbf{W}$:

$$\boxed{\mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}}, \tag{26}$$

em que $\mathbf{W}$ é obtida a partir dos dados $\mathbf{X}$ e $\mathbf{Y}$.

Formalização Matemática do Problema

- A expressão mostrada na Eq. (26) é conhecida como estimador dos mínimos quadrados ordinários (MQO) da matriz de coeficientes $\mathbf{W}$.
- Esta expressão recebe este nome devido ao problema de otimização associado, que busca minimizar a norma do vetor de erros quadráticos relacionados ao modelo preditivo $\hat{\mathbf{Y}} = \mathbf{WX}$.
- Matematicamente, isto equivale a minimizar a seguinte função objetivo:

$$J_{MQO}(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{Y} - \mathbf{WX})^T (\mathbf{Y} - \mathbf{WX})\right\}, \qquad (27)$$

em que Tr denota o operador *traço* de uma matriz.

Regularização de Tikhonov

- Para evitar problemas causados pelo mal-condicionamento da matriz $\mathbf{X}$, pode-se usar a regularização de Tikhonov.
- A matriz de coeficientes $\mathbf{W}$ passa a ser estimada por

$$\boxed{\mathbf{W} = \mathbf{Y}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \lambda \mathbf{I}_n\right)^{-1}} \tag{28}$$

  em que $0 < \lambda \leq 1$ é chamada de constante de regularização e $\mathbf{I}_n$ é a matriz identidade de ordem n.
- A constante λ é, na verdade, um **hiperparâmetro**.
- Um hiperparâmetro é um parâmetro do modelo que deve ser pré-definido para que os parâmetros da função discriminante propriamente dita possam ser estimados.

Formalização Matemática do Problema

- De posse da matriz $\mathbf{W}$, podemos usá-la para construir um modelo preditor. Matematicamente, isto pode ser feito por meio da seguinte transformação matricial:

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}, \tag{29}$$

  em que o vetor $\mathbf{x} \in \mathbb{R}^{34}$ denota a versão "numérica" de um paciente qualquer, enquanto o vetor $\hat{\mathbf{y}} \in \mathbb{R}^{6}$ simboliza o vetor de predição da patologia do paciente $\mathbf{x}$.
- Cabe ao desenvolvedor do sistema, desenvolver uma interface amigável de modo a tornar a operação matemática acima transparente para o usuário.

Formalização Matemática do Problema

A expressão $\mathbf{y} = \mathbf{W}\mathbf{x}$ pode ser decomposta em m saídas individuais:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1j} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2j} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} & \cdots & w_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mj} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^T\mathbf{x} \\ \mathbf{w}_2^T\mathbf{x} \\ \vdots \\ \mathbf{w}_i^T\mathbf{x} \\ \vdots \\ \mathbf{w}_m^T\mathbf{x} \end{bmatrix} \tag{30}$$

Formalização Matemática do Problema

- Assim, cada componente do vetor $\mathbf{y}$ pode ser tomada individualmente e escrita de diferentes maneiras:

$$y_i = \mathbf{w}_i^T \mathbf{x} = \sum_{j=1}^{n} w_{ij} x_j \quad (31)$$

$$= w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{in}x_n \quad (32)$$

- Estas expressões definem a *função discriminante linear* da i-ésima classe, $i = 1, \ldots, m$.
- No presente estudo caso, a i-ésima classe corresponde à i-ésima patologia.

Interpretação Geométrica da Função Discriminante

- A função $y_i = \mathbf{w}_i^T \mathbf{x}$, nesta forma envolvendo 2 vetores, é particularmente importante para interpretação do problema de classificação de padrões
- O vetor $\mathbf{w}_i$, de dimensão $n \times 1$, é chamado de vetor de coeficientes da função discriminante da i-ésima classe.
- O vetor $\mathbf{x}$, também de dimensão $n \times 1$, é chamado de vetor de atributos do objeto a ser classificado.
- O valor da saída y_i é dado pelo produto escalar de $\mathbf{w}_i$ com $\mathbf{x}$.

Interpretação Geométrica da Função Discriminante

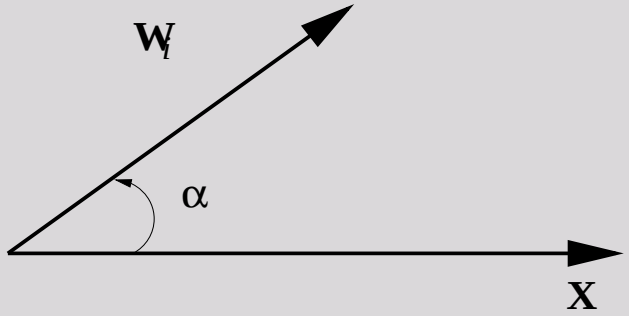
- Em classificação de padrões, o produto escalar é interpretado como uma medida de similaridade entre dois vetores quaisquer.
- Para ajudar na interpretação geométrica, usarei uma definição alternativa de produto escalar:

$$y_i = \mathbf{w}_i^T \mathbf{x} = \|\mathbf{w}_i\| \cdot \|\mathbf{x}\| \cdot \cos(\alpha) \tag{33}$$

em que o símbolo $\|\cdot\|$ define a norma euclidiana do vetor e α é o menor ângulo entre os vetores $\mathbf{w}_i$ e $\mathbf{x}$.

- A figura no próximo slide ilustra uma disposição hipotética dos vetores $\mathbf{w}_i$ e $\mathbf{x}$ para um ângulo $0^o < \alpha < 90^o$.

Interpretação Geométrica da Função Discriminante



- Quanto menor o ângulo α, mais próximos estarão os 2 vetores, e maior será o valor do produto escalar.
- O maior valor possível é atingido para $\alpha = 0$, quando os dois vetores estão alinhados sobre a mesma reta suporte.
- Assim, o produto escalar pode ser usado como medida de similaridade entre vetores.

Interpretação Geométrica da Função Discriminante

- O vetor $\mathbf{w}_i^T$ corresponde à i-ésima linha da matriz $\mathbf{W}$, $i = 1, \ldots, m$.
- Este vetor pode ser interpretado como um "protótipo" (ou modelo) dos pacientes da i-ésima classe.
- Assim, cada linha da matriz $\mathbf{W}$ contém um protótipo dos pacientes daquela classe.
- Logo, o maior valor de saída y_i indica qual protótipo é mais parecido com o vetor de entrada.
- Para a aplicação atual, o maior valor da saída vai indicar qual *protótipo* armazenado é aquele que mais se assemelha ao estado do paciente $\mathbf{x}$.

Interpretação Geométrica da Função Discriminante

- Cada vetor $\mathbf{w}_i^T$ pode ser estimado individualmente partir das matrizes $\mathbf{X}$ e $\mathbf{Y}$ por meio da seguinte expressão:

$$\boxed{\mathbf{w}_i^T = \mathbf{Y}_{[i,:]}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}}, \tag{34}$$

  em que $\mathbf{Y}_{[i,:]}$ simboliza a i-ésima linha da matriz $\mathbf{Y}$.

- A expressão anterior resulta em um vetor linha. Para obter o vetor coluna $\mathbf{w}_i$, podemos utilizar a seguinte expressão:

$$\boxed{\mathbf{w}_i = \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{Y}_{[i,:]}^T} \tag{35}$$

  que define o **estimador de mínimos quadrados** dos parâmetros da função discriminante da i-ésima classe.

Exemplo de Diagnóstico

- Um novo paciente do médico usuário do sistema computacional de auxílio ao diagnóstico foi codificado pelo seguinte vetor de atributos:

$$\mathbf{x}_{new} = [2\ 1\ 2\ 3\ 1\ 3\ 0\ 3\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 2\ 0\ 2\ 0\ 0\ 0\ 0\ 0\ 2\ 0\ 2\ 3\ 2\ 0\ 0\ 2\ 3\ 26]^T$$

- Ao multiplicarmos este vetor pela matriz $\mathbf{W}$ calculada na Equação (26), obtemos o seguinte vetor de saída $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}_{new} \tag{36}$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \\ \hat{y}_5 \\ \hat{y}_6 \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^T\mathbf{x}_{new} \\ \mathbf{w}_2^T\mathbf{x}_{new} \\ \mathbf{w}_3^T\mathbf{x}_{new} \\ \mathbf{w}_4^T\mathbf{x}_{new} \\ \mathbf{w}_5^T\mathbf{x}_{new} \\ \mathbf{w}_6^T\mathbf{x}_{new} \end{bmatrix} = \begin{bmatrix} -0.076297 \\ 0.113172 \\ 1.061544 \\ -0.123137 \\ -0.098041 \\ 0.015406 \end{bmatrix} \tag{37}$$

Exemplo de Diagnóstico (cont.)

- Analizando as componentes do vetor $\mathbf{y}$, percebemos que a maior delas é a terceira componente (i.e. $\hat{y}_3 = 1.061544$).
- Assim, a regra de decisão é formalmente dada por

$$j^* = \text{índice da classe de } \mathbf{x}_{new} = \arg\max_{\forall j}\{\hat{y}_j\}, \qquad (38)$$

em que a função max retorna o maior valor entre todas as saídas y_j e a função arg retorna o índice (i.e. a posição) da maior saída dentro do vetor.

- Assim, o sistema computacional está sugerindo que o paciente pelo vetor de atributos clínicos e histopatológicos, $\mathbf{x}_{new}$, apresenta características da patologia **Líquen Plano**.

- É comum aplicar algumas funções ao vetor de saídas $\mathbf{y}_{new}$ para deixar mais em evidência a saída de maior valor.
- Uma das opções mais comuns é aplicar uma normalização das saídas $\hat{y}_i$ pelo uso da função *softmax*:

$$\hat{y}_i^* = \frac{e^{\hat{y}_i}}{\sum_{r=1}^{m} e^{\hat{y}_r}}, \quad i = 1, \ldots, m. \tag{39}$$

- A aplicação da função *softmax* faz com que todas as saídas sejam positivas e que a soma delas seja igual a 1:

$$\hat{y}_i^* > 0 \quad \text{e} \quad \sum_{i=1}^{m} \hat{y}_i^* = 1. \tag{40}$$

- Contudo, não se deve interpretar as saídas normalizadas $\hat{y}_i^*$ como probabilidades por causa disso.

- Aplicando a função softmax ao vetor de saídas preditas da Eq. (36), temos

$$\hat{\mathbf{y}} = \begin{bmatrix} -0.076297 \\ 0.113172 \\ \mathbf{1.061544} \\ -0.123137 \\ -0.098041 \\ 0.015406 \end{bmatrix} \quad \Rightarrow \quad \hat{\mathbf{y}}^* = \begin{bmatrix} 0.11965 \\ 0.14462 \\ \mathbf{0.37332} \\ 0.11418 \\ 0.11708 \\ 0.13115 \end{bmatrix} \tag{41}$$

- Uma altenativa interessante consiste em usar uma função de quantização:

$$\hat{y}_i^* = \begin{cases} 1, & \text{se } \hat{y}_i \geq \beta \\ 0, & \text{se } \hat{y}_i < \beta \end{cases}, \quad i = 1, \ldots, m, \tag{42}$$

tal que $0 < \beta < 1$ é um limiar de quantização, normalmente constante e igual a β =0,5.

- Se a codificação da saída foi feita usando ± 1, a expressão acima é alterada para

$$\hat{y}_i^* = \begin{cases} +1, & \text{se } \hat{y}_i \geq \beta \\ -1, & \text{se } \hat{y}_i < \beta \end{cases}, \quad i = 1, \ldots, m, \tag{43}$$

com o limiar definido como $\beta = 0$.

- Aplicando a função de quantização da Eq. (42), temos

$$\hat{\mathbf{y}} = \begin{bmatrix} -0.076297 \\ 0.113172 \\ \mathbf{1.061544} \\ -0.123137 \\ -0.098041 \\ 0.015406 \end{bmatrix} \quad \Rightarrow \quad \hat{\mathbf{y}}^* = \begin{bmatrix} 0 \\ 0 \\ \mathbf{1} \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{44}$$

- Se, no momento da quantização, ocorrer mais de uma saída igual a 1, ou todas as saídas iguais a zero, pode-se optar por não classificar o vetor $\mathbf{x}_{new}$.
- Esta estratégia é conhecida como *opção de rejeição*.

Implementação no Matlab/Octave

- A Eq. (26) pode ser implementada no Matlab/Octave por meio da seguinte linha de comando:

```
» W = Y*X'*inv(X*X');
```

- Contudo, este procedimento não é recomendado devido ao seu elevado custo computacional (exige muito uso de memória) e alta susceptibilidade a erros numéricos quando a matriz $\mathbf{X}\mathbf{X}^T$ está próxima da singularidade.
- O resultado será confiável apenas se a matriz for de posto completo, i.e., posto$(\mathbf{X}) = \min(n, N)$.
- Além disso, esta expressão não escalona bem para dados de alta dimensão.

Implementação no Matlab/Octave (cont.-1)

- Para evitar problemas na inversão da matriz $\mathbf{X}\mathbf{X}^T$, costuma-se usar a regularização de Tikhonov.
- Para isso, faz-se necessário definir uma constante $0 < \lambda \leq 1$, chamada de constante de regularização.
- Assim, tem-se que usar as seguintes linhas de comando:

```
» lam=0.01;

» W = Y*X'*inv(X*X'+lam*eye(size(X*X')));
```

- Esta expressão favore soluções de norma mínima, porém ainda consome muita memória e não escalona bem para dados de alta dimensão.

Implementação no Matlab/Octave (cont.-2)

- Para evitar problemas com matrizes de posto incompleto, recomenda-se o uso do comando `pinv`:

```
» W = Y*pinv(X);
```

- Este procedimento escalona bem para dados de alta dimensão, pois faz uso eficiente de memória.
- Esta abordagem utiliza uma versão aproximada da *Decomposição em Valores Singulares* (SVD) para calcular a inversa de $\mathbf{X}\mathbf{X}^T$ e tratar matrizes de posto incompleto, ou seja

$$\text{posto}(\mathbf{X}) < \min(n, N). \tag{45}$$

Implementação no Matlab/Octave (cont.-1)

- Uma alternativa bem interessante para implementar a Eq. (26) no Matlab/Octave faz uso do operador *barra invertida* "/":

```
» W = Y/X;
```

- Esta alternativa não envolve a inversão explícita de $\mathbf{X}\mathbf{X}^T$, pois usa o método de eliminação de Gauss.
- Este método escalona muito bem para dados de alta dimensão e possui o menor custo de processamento (ou seja, é mais rápido) que os métodos anteriores.
- Versão com regularização:

```
» W = (Y*X')/(X*X'+lam*eye(size(X*X')));
```

Método HOLDOUT de Avaliação do Classificador

- Na implementação da Eq. (26) usamos todos os N pares entrada-saída $(\mathbf{x}_k, \mathbf{y}_k)$, $k = 1, \ldots, N$, disponíveis.
- Porém, avaliar o desempenho do classificador com os mesmos dados usados no seu projeto não é uma boa prática.
- A prática correta requer a separação da matriz $\mathbf{X}$ e, por extensão, da matriz $\mathbf{Y}$, em duas partes.
- A primeira parte $(\mathbf{X}_{trn}, \mathbf{Y}_{trn})$ contendo N_{trn} exemplos de treino.
- A segunda parte $(\mathbf{X}_{tst}, \mathbf{Y}_{tst})$ contendo N_{tst} exemplos de teste.

$$\mathbf{X} = [\mathbf{X}_{trn} \mid \mathbf{X}_{tst}] \quad (46)$$

$$\mathbf{Y} = [\mathbf{Y}_{trn} \mid \mathbf{Y}_{tst}] \quad (47)$$

tal que $N = N_{trn} + N_{tst}$.

Método HOLDOUT de Avaliação do Classificador

- Este procedimento chama-se *Hold-out*, que literalmente significa segurar uma parte dos dados fora, para teste do classificador.
- Para isso, a escolha dos exemplos que comporão as matrizes de treino/teste deve ser randomizada. Porém, tal randomização pode conduzir a um desempenho bom (ou ruim) por mero fruto do acaso, se for realizado uma única vez.
- Assim, a randomização das matrizes de treino-teste deve ser repetida por N_r rodadas independentes. Para isso, fazemos

$$\mathbf{X}(r) = [\mathbf{X}_{trn}(r) \mid \mathbf{X}_{tst}(r)] \quad (48)$$

$$\mathbf{Y}(r) = [\mathbf{Y}_{trn}(r) \mid \mathbf{Y}_{tst}(r)] \quad (49)$$

em que $\mathbf{X}(r)$ e $\mathbf{Y}(r)$ contém os mesmos vetores-coluna que as matrizes $\mathbf{X}$ e $\mathbf{Y}$, porém em posições diferentes a cada rodada r.

Método HOLDOUT de Avaliação do Classificador

- Observar que as posições aleatorizadas das colunas da matriz de rótulos $\mathbf{Y}(r)$ devem ser as mesmas da matriz $\mathbf{X}(r)$ a fim de não perder a correspondência paciente-diagnóstico.
- O procedimento de selecionar aleatoriamente os exemplos de treino-teste por várias rodadas independentes é chamado de *método de Monte Carlo*.
- Para cada rodada, devemos re-estimar a matriz $\mathbf{W}$ e recalcular os índices numéricos que caracterizam o desempenho do classificador.
- Ao final das N_r rodadas, devemos fornecer estatísticas descritivas dos índices de desempenho, tais como média, desvio-padrão, mediana, valor mínimo e valor máximo.

Exemplos de Índices de Desempenho

- Taxa de acerto global (P_{acerto}):

$$P_{acerto} = \frac{N_{acertos}}{N_{tst}}, \tag{50}$$

em que $N_{acertos}$ é o número de exemplos de teste corretamente classificados.

- Taxa de acerto da i-ésima classe ($P_{acerto}(\omega_i)$):

$$P_{acerto}(\omega_i) = \frac{N_{acertos}(\omega_i)}{N_{tst}(\omega_i)}, \tag{51}$$

em que $N_{tst}(\omega_i)$ é o número de exemplos de teste com rótulos iguais a ω_i, e $N_{acertos}(\omega_i)$ é o número de exemplos de teste da i-ésima classe corretamente classificados.

Pseudocódigo: Classificador Linear de Mínimos Quadrados

1. **Inicialização**

   1.1 - Carregar dados e montar matrizes $\mathbf{X}$ e $\mathbf{Y}$.
   1.2 - Determinar n, m, N, N_{trn}, N_{tst} e N_r.

2. **Treinamento/Teste**
   FOR $r = 1$ **TO** N_r

   2.1 - Definir dados de treino: $\mathbf{X}_{trn}(r)$ e $\mathbf{Y}_{trn}(r)$.
   2.2 - Definir dados de teste: $\mathbf{X}_{tst}(r)$ e $\mathbf{Y}_{tst}(r)$.
   2.3 - Estimar $\mathbf{W}(r) = \mathbf{Y}_{trn}(r) \cdot \text{pinv}(\mathbf{X}_{trn}(r))$.
   2.4 - Classificar exemplos de teste: $\mathbf{Y}_{pred} = \mathbf{W}\mathbf{X}_{tst}(r)$.
   2.5 - Calcular taxa de acerto na rodada r:
   $P_{acerto}(r) = F[\mathbf{Y}_{tst(r)}, \mathbf{Y}_{pred}(r)]$.

   ENDFOR

3. **Avaliação de Desempenho**

   3.1 - Calcular estatísticas de desempenho global.
   3.2 - Calcular estatísticas de desempenho por classe.

Regra de Aprendizado do Perceptron Simples

- Perceba que a determinação da matriz $\mathbf{W}$ via Eq. (26) exige o armazenamento em memória de todos os vetores de atributos $\mathbf{x}_i$ e seus respectivos vetores-rótulos $\mathbf{y}_i$, $i = 1, 2, \ldots, 358$.
- Uma alternativa mais econômica no uso de memória consiste em utilizar a regra de aprendizado do Perceptron, que na sua forma matricial é dada por

$$\boxed{\mathbf{W}(n+1) = \mathbf{W}(n) + \eta \mathbf{e}(n)\mathbf{x}^T(n)} \qquad (52)$$

em que $0 < \eta \ll 1$ é chamada de passo de aprendizagem e n denota o instante ou iteração atual.

Regra de Aprendizado do Perceptron Simples

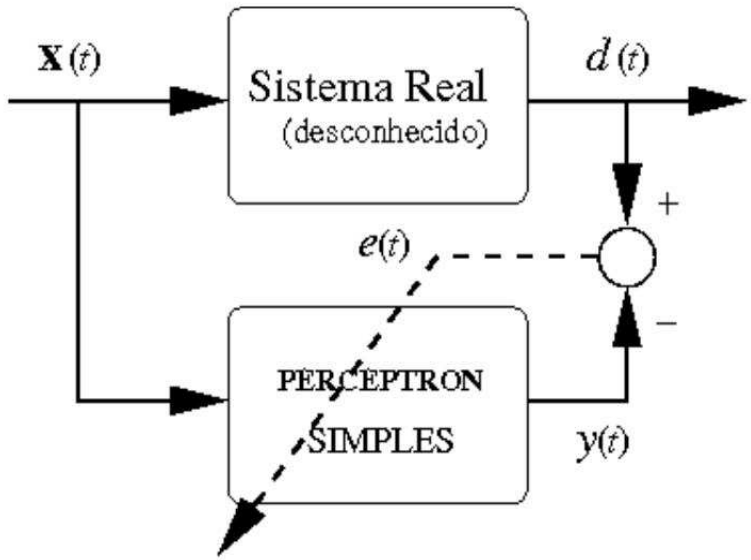
- O vetor de erros na iteração n é dado por

$$\begin{aligned} \mathbf{e}(n) &= \mathbf{y}(n) - \mathbf{y}(n), & (53) \\ &= \mathbf{y}(n) - \text{sinal}(\mathbf{W}(n)\mathbf{x}(n)). & (54) \end{aligned}$$

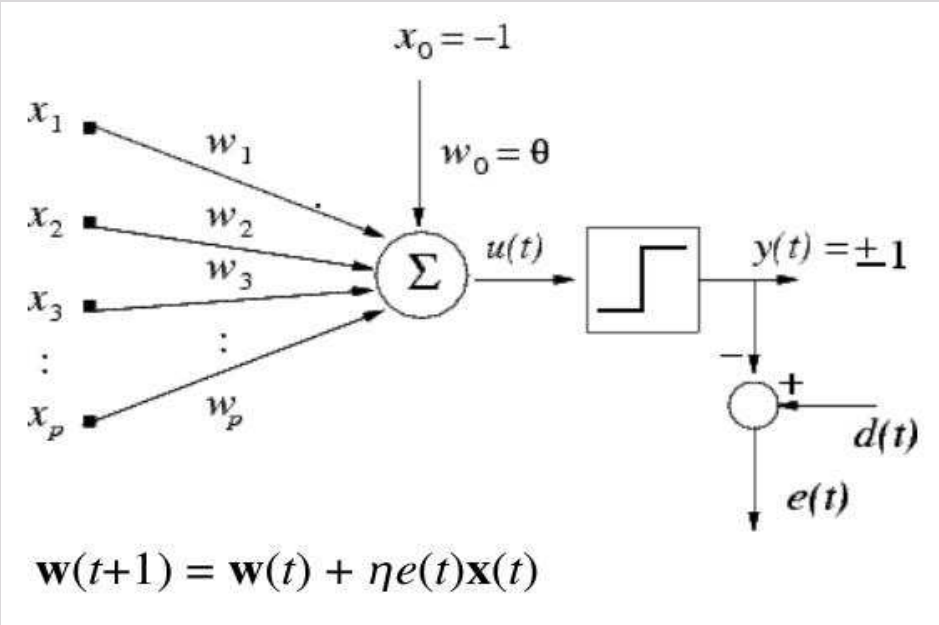
em que $\mathbf{y}(n) = \text{sinal}(\mathbf{W}(n)\mathbf{x}(n))$ denota a saída da rede naquele instante.

Regra de Aprendizado do Perceptron Simples

O processo de aprendizagem, ou seja, de modificação dos parâmetros do neurônio M-P é guiado pelo erro (e) e pelo vetor de entrada ($\mathbf{x}$)!



Neurônio de McCulloch & Pitts + Regra de Aprendizado



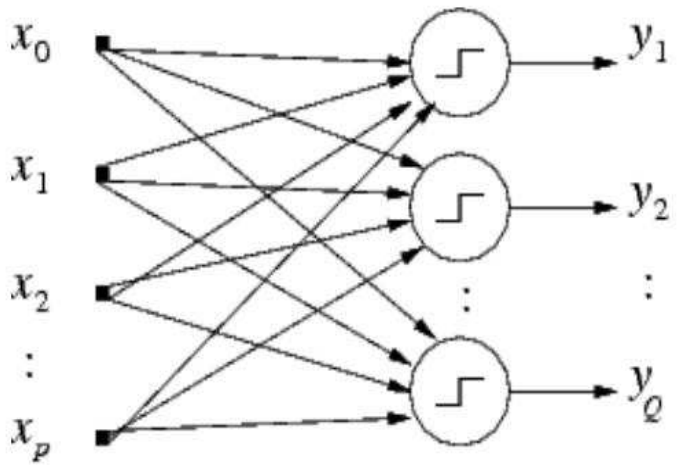
Algoritmo Perceptron Simples (1 neurônio)

1. Início (t=0)
   - 1.1 – Definir valor de η entre 0 e 1.
   - 1.2 – Iniciar $\mathbf{w}(0)$ com valores nulos ou aleatórios.
2. Funcionamento
   - 2.1 – Selecionar vetor de entrada $\mathbf{x}(t)$.
   - 2.2 – Calcular ativação $u(t)$.
   - 2.3 – Calcular saída $y(t)$.
3. Treinamento
   - 3.1 – Calcular erro: $e(t) = d(t) - y(t)$
   - 3.2 – Ajustar pesos via regra de aprendizagem.
   - 3.3 – Verificar critério de parada.
     - 3.3.1 – Se atendido, finalizar treinamento.
     - 3.3.2 – Caso contrário, fazer t=t+1 e ir para Passo 2.

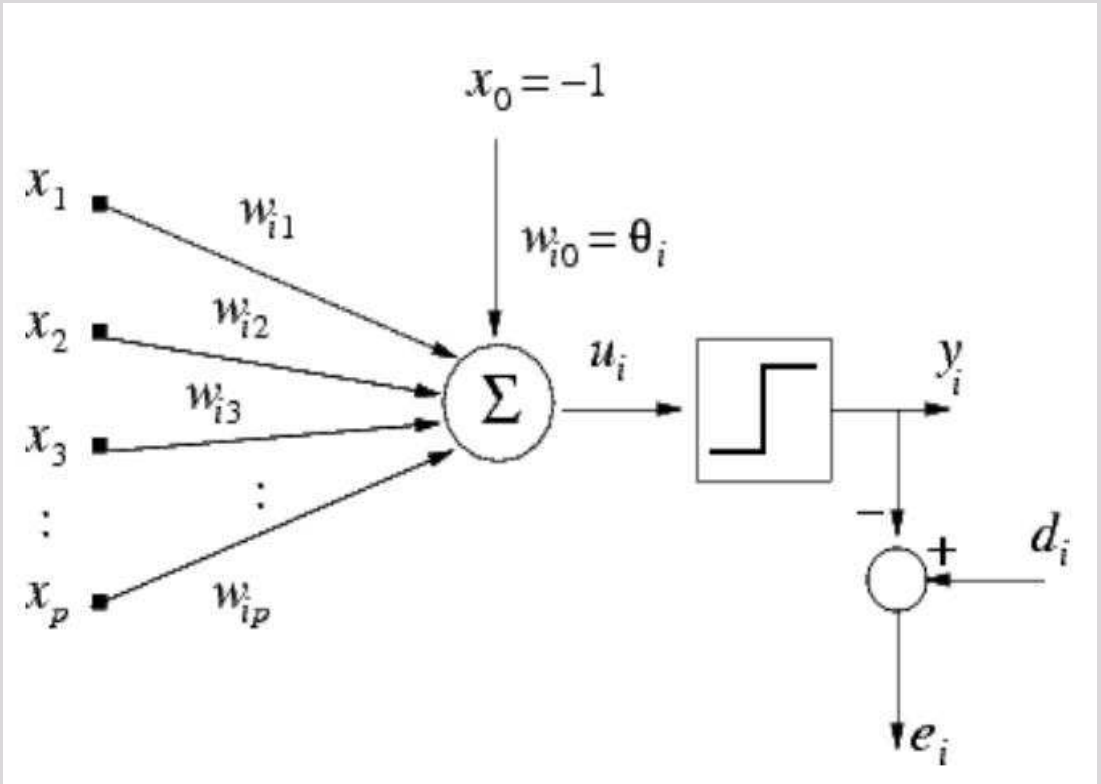
Arquitetura da Rede Perceptron Simples (Q neurônios)

Um único neurônio M-P categoriza apenas duas classes de dados.

Em problemas com múltiplas classes, deve-se utilizar vários neurônios em paralelo.



Representação do i-ésimo neurônio da rede PS.



Funcionamento do i-ésimo neurônio da rede PS.

O funcionamento de cada neurônio individualmente é o mesmo.

Assim, a ativação do i-ésimo neurônio da rede PS é dada por:

$$u_i = \mathbf{w}_i^T \mathbf{x} = w_{i1} x_1 + w_{i2} x_2 + ... + w_{ip} x_p$$

A saída do i-ésimo neurônio é dada por:

$$y_i = \text{sinal}(u_i) = \text{sinal}(\mathbf{w}_i^T \mathbf{x})$$

O erro do i-ésimo neurônio é dado por: $e_i = d_i - y_i$

onde d_i é a saída desejada do i-ésimo neurônio.

$i = 1, ..., Q$ ($Q \geq 1$ é o número de neurônios de saída).

Treinamento do i-ésimo neurônio da rede PS.

Como cada neurônio tem seu próprio vetor de pesos $\mathbf{w}_i$, $i = 1, 2, ..., Q$, então teremos agora Q regras de aprendizagem!

Ou seja, uma regra de aprendizagem para cada vetor $\mathbf{w}_i$.

Assim, a regra de aprendizagem do i-ésimo neurônio é dada por:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \eta\, e_i(t)\, \mathbf{x}(t)$$

Em que $0 < \eta \ll 1$ e $i=1, 2, ..., Q$.

Algoritmo Perceptron Simples (Q neurônios)

1. Início (t=0)
   - 1.1 – Definir valor de η entre 0 e 1.
   - 1.2 – Iniciar $\mathbf{w}_i(0)$ com valores aleatórios.
2. Funcionamento
   - 2.1 – Selecionar o vetor de entrada $\mathbf{x}(t)$.
   - 2.2 – Calcular as Q ativações $u_i(t)$.
   - 2.3 – Calcular as Q saídas $y_i(t)$.
3. Treinamento
   - 3.1 – Calcular os Q erros: $e_i(t) = d_i(t) - y_i(t)$
   - 3.2 – Ajustar os Q vetores de pesos $\mathbf{w}_i(t)$.
   - 3.3 – Verificar critério de parada.
     - 3.3.1 – Se atendido, finalizar treinamento.
     - 3.3.2 – Caso contrário, fazer t=t+1 e ir para Passo 2.