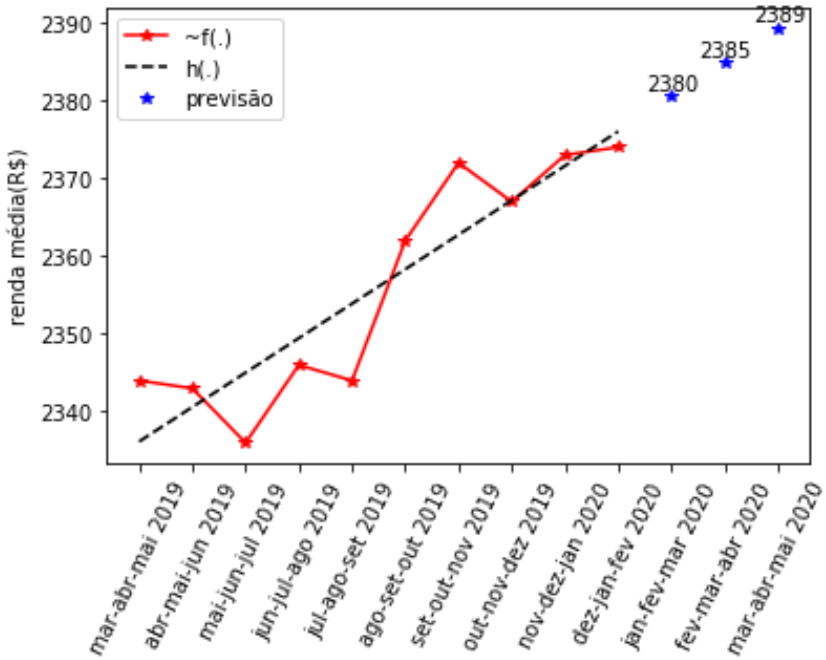
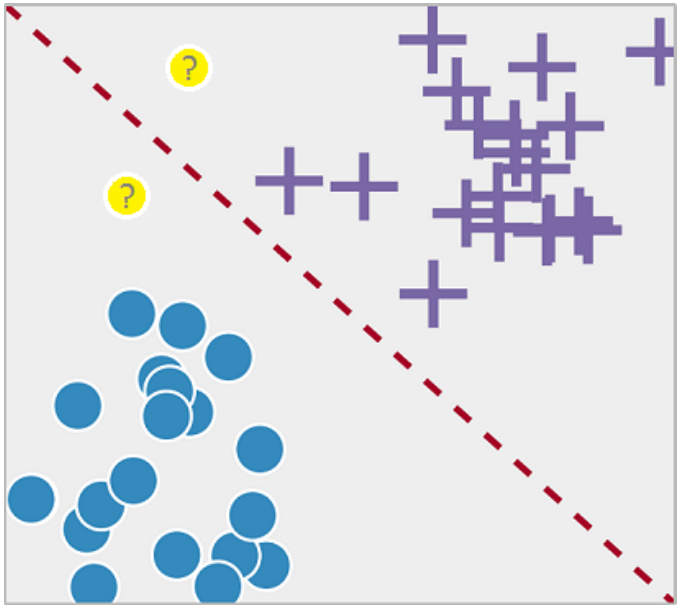
Universidade Federal da Paraíba
Coordenação do Curso de Ciência de Dados e Inteligência Artificial

Modelo de Aprendizagem Linear I

Prof. Gilberto Farias

Regressão x Classificação

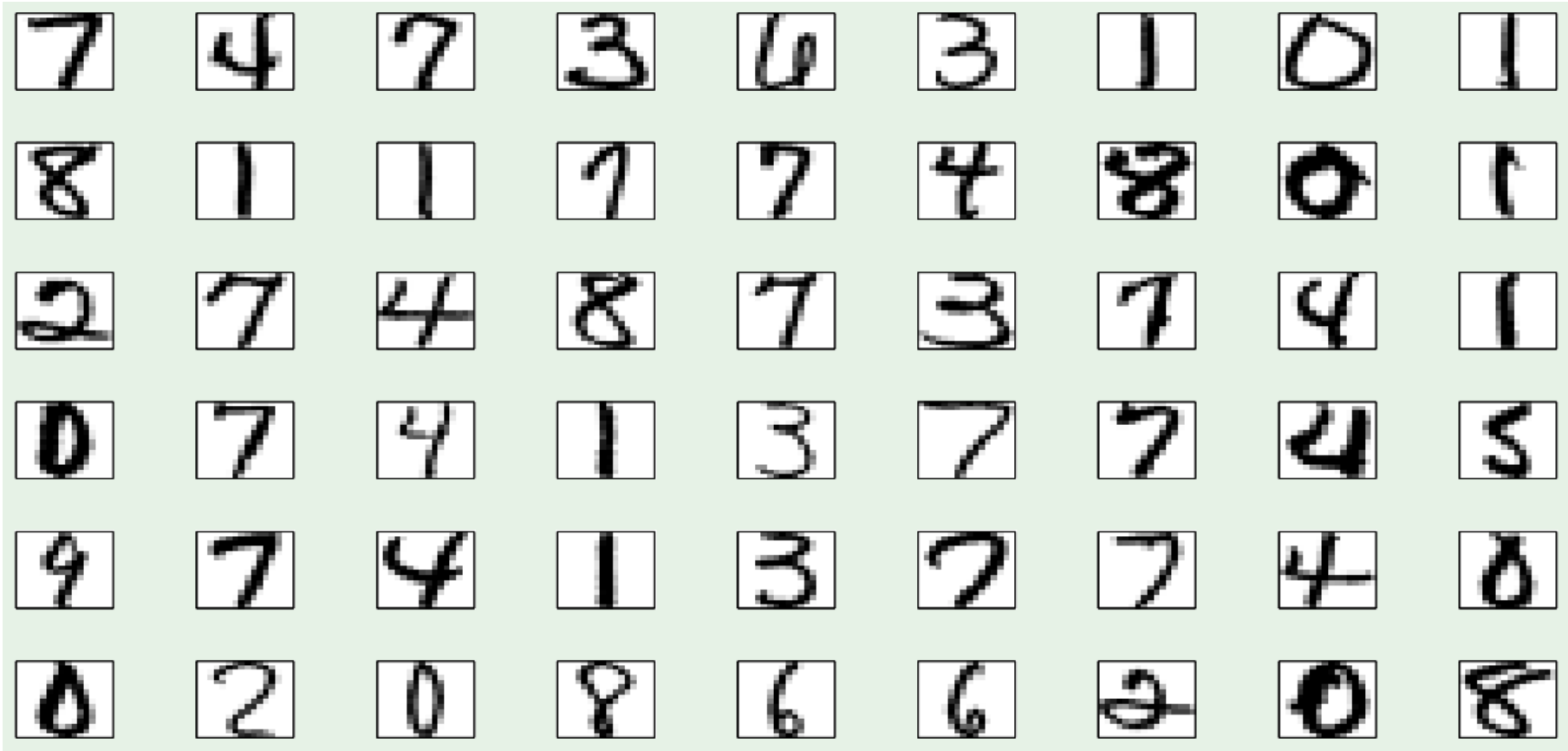




Roteiro

- Representando dados reais
  - Revisando o PLA
  - Prática SimpleSpam
- Classificação Linear
  - Uso da Regressão Linear
  - Prática Classificação com Regressão Linear
- Como aproximar $h(x)$ de $f(x)$?
- Métricas de aprendizado

Representando dados reais

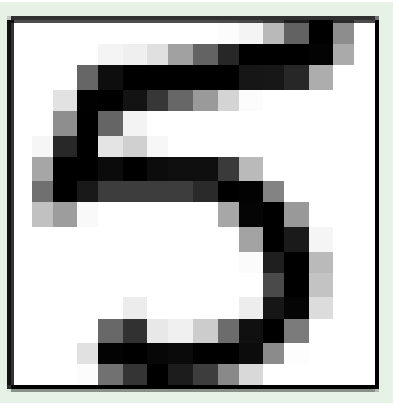


Notas de aula Yaser Abu-Mostafa

Representação da entrada

Entrada bruta $\boldsymbol{x} = (x_0, x_1, \cdots, x_{256})$

Modelo linear $\boldsymbol{w} = (w_0, w_1, \cdots, w_{256})$



Características: extrair informação útil

intensidade e simetria $\boldsymbol{x} = (x_0, x_1, x_2)$

Modelo linear (w_0, w_1, w_2)

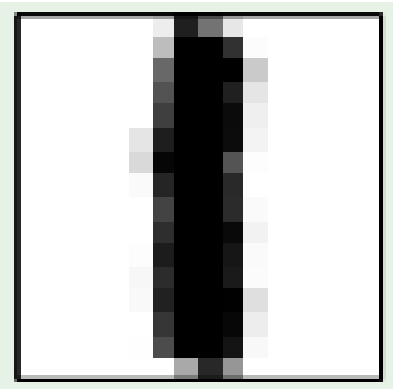
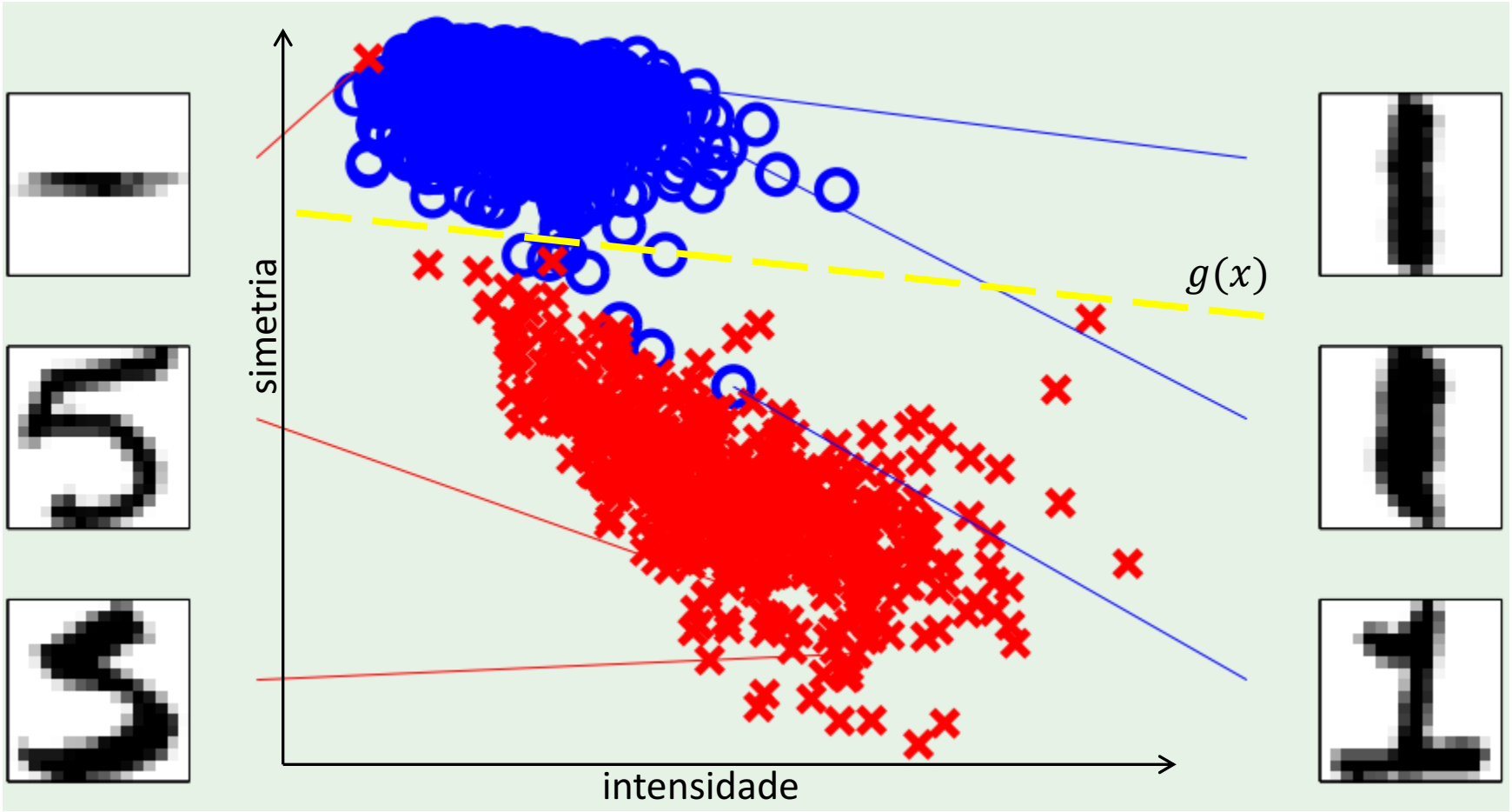


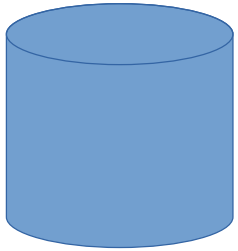
Ilustração da classificação

$\boldsymbol{x} = (x_0, x_1, x_2)$ x_1: intensidade x_2: simetria



Adaptado das notas de aula de Yaser Abu-Mostafa

Classificar *email* como span



Análise dos dados

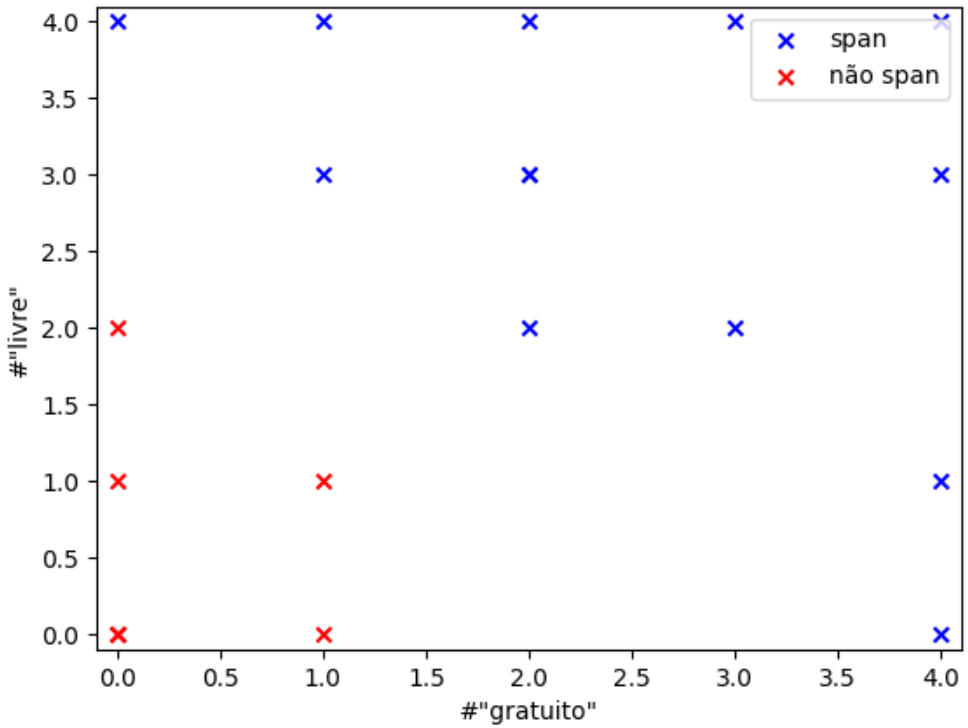
Sistemas/UFPB	PROGEP/CQVSST/DQVS REITERA A OF...	25 de mai.
Lucidio Cabral	Fwd: Convite: Desenvolvimento de Curso - -...	22 de mai.
Lourdes Maria Rodri.	Convocação para a 1ª Reunião Extraord...	22 de mai.
rhaian jose farias .	RE: Prova 1 - Rhaian Barros - ESTRUTU...	22 de mai.
Yuri .. Christian 6	Disponibilização dos cursos do Coursera g...	22 de mai.
Ruy, Ruy, Alisson 4	[professoresppgi_ufpb:3205] Cadastro de o...	18 de mai.
OR Spectrum (ORSP)	Are you willing to review ORSP-D-19-0...	17 de mai.
Thiago Gouveia	Fwd: [seminarios-grafos] Próximo seminári...	13 de mai.
Danielle Rousy Dias.	LC-EaD-Plano de Aplicação de Provas-2...	5 de mai.

Banco de mensagens de *emails*

#"gratuito"	#"livre"	span
2	3	+1
4	3	+1
0	0	-1
4	0	+1
0	4	+1
0	0	-1
4	1	+1
2	2	+1
0	0	-1
1	0	-1
3	2	+1
0	1	-1
1	4	+1
2	4	+1
1	3	+1
1	1	-1
0	2	-1
2	3	+1
4	4	+1
3	4	+1

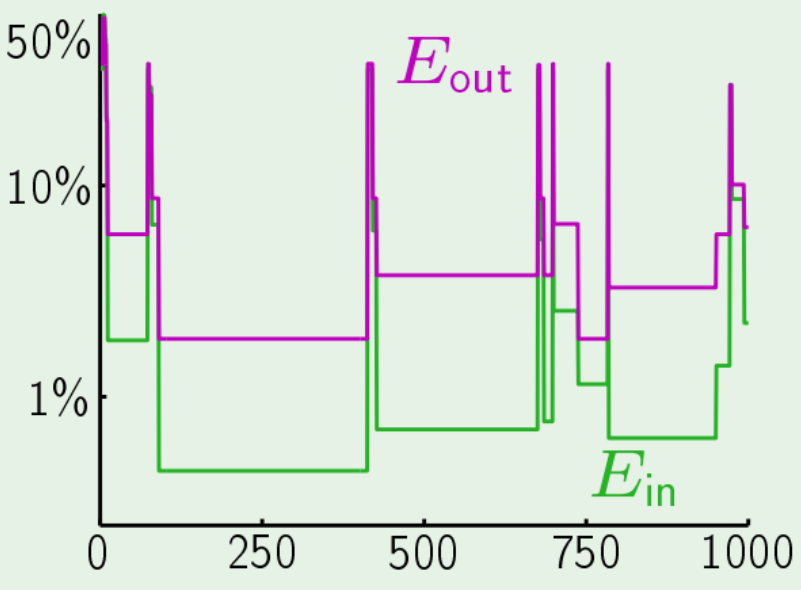
Exemplos de Treinamento

X		Y
#”gratuito”	#”livre”	span
2	3	+1
4	3	+1
0	0	-1
4	0	+1
0	4	+1
0	0	-1
4	1	+1
2	2	+1
0	0	-1
1	0	-1
3	2	+1
0	1	-1
1	4	+1
2	4	+1
1	3	+1
1	1	-1
0	2	-1
2	3	+1
4	4	+1
3	4	+1

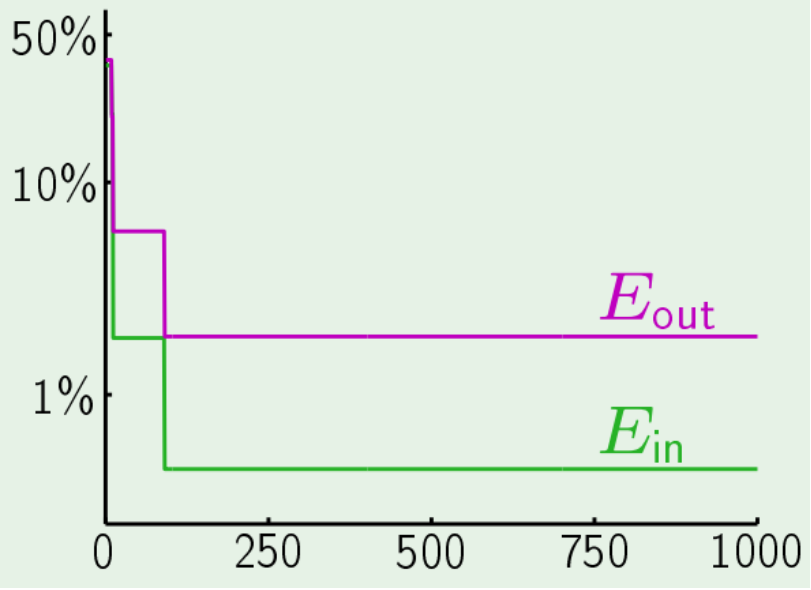


Analisando o PLA para dados linearmente não separáveis

PLA:



Pocket:



Regressão linear para Classificação

- Regressão linear aprende uma função de valor real $y = f(x) \in \mathbb{R}$
- Valores binário também são valores reais! $\pm 1 \in \mathbb{R}$
- Use a regressão linear para pegar um $\boldsymbol{w}$ onde $w^{\mathrm{T}} x_n \approx y_n = \pm 1$
- Use $\mathrm{sign}(w^{\mathrm{T}} x_n)$ para concordar com $y_n = \pm 1$
- Constrói uma boa ponderação inicial para classificação.

$$h(x) = sign\left(\sum_{i=0}^{d} w_i x_i\right)$$

Modelo *perceptron*

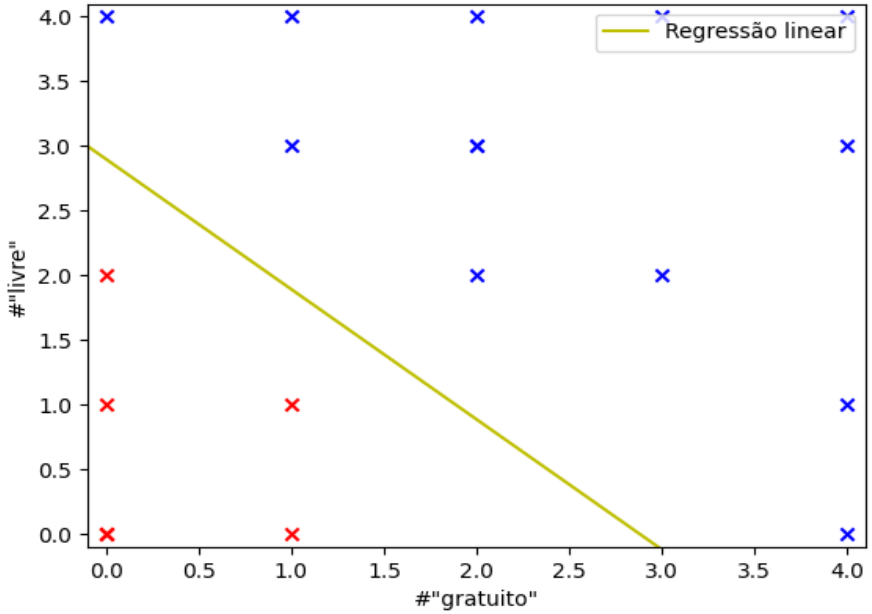
Função hipótese g()

X		Y
#"gratuito"	#"livre"	span
2	3	+1
4	3	+1
0	0	-1
4	0	+1
0	4	+1
0	0	-1
4	1	+1

Exemplos de treinamento

Regressão linear

Treino de *w*



X		Y
#"gratuito"	#"livre"	span
0	2	??
2	3	??
4	4	??
3	4	??

Emails para classificar

$g(x) = sign(w^T x)$

X		Y
#"gratuito"	#"livre"	span
0	2	-1
2	3	+1
4	4	+1
3	4	+1

Como Aproximar $h(x)$ de $f(x)$?

O que significa "$h \approx f$"??

- No aprendizado não se espera replicar a função alvo f perfeitamente;
- A medição de erro quantifica a aproximação da função hipótese h;

Medição de erro: $\boldsymbol{E(h, f)}$

Medida de erro pontual: $e\big(h(x), f(x)\big)$

Exemplos:

Erro quadrático: $\boldsymbol{e\big(h(x), f(x)\big) = (h(x) - f(x))^2}$

Erro binário: $\boldsymbol{e\big(h(x), f(x)\big) = [\![h(x) \neq f(x)]\!]}$

Medindo o erro da função hipótese

- O erro $E(h, f)$ é a média dos erros individuais $e\big(h(x), f(x)\big)$

- Erro dentro da amostra:

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} e\big(h(x_n), f(x_n)\big)$$

- Erro fora da amostra:

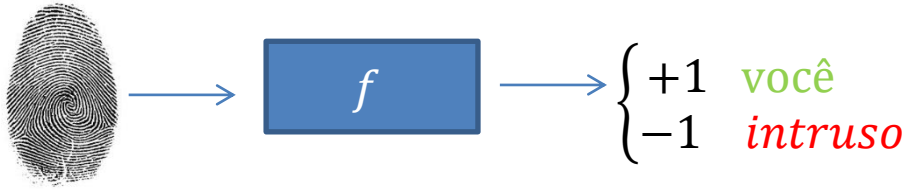
$$E_{out}(h) = E_x\big[e\big(h(x), f(x)\big)\big]$$

Escolhendo uma medição de erro

Tipos de erros:

1. falso positivo
2. falso negativo

Verificação de digital



Como penalizar cada tipo de erro?

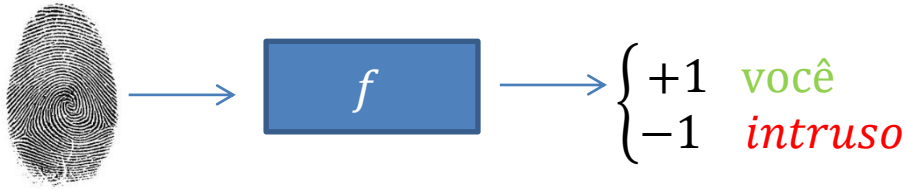
		f	
		+1	-1
h	+1		
	-1		

Escolhendo uma medição de erro

Tipos de erros:

1. falso positivo
2. falso negativo

Verificação de digital



Como penalizar cada tipo de erro?

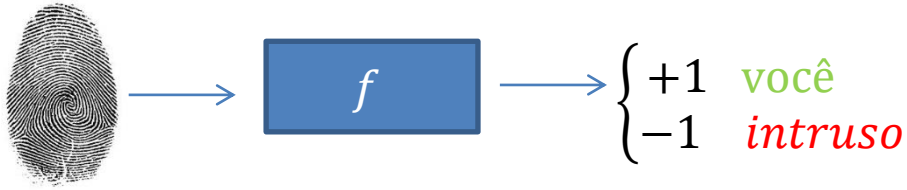
		f	
		+1	-1
h	+1	sem erro	
	-1		sem erro

Escolhendo uma medição de erro

Tipos de erros:

1. falso positivo
2. falso negativo

Verificação de digital



Como penalizar cada tipo de erro?

		f	
		+1	-1
h	+1	sem erro	falso aceite
	-1	falso rejeito	sem erro

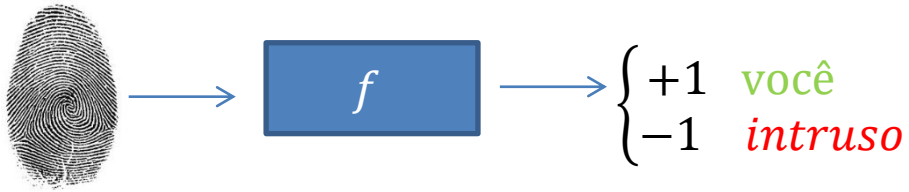
Depende da aplicação!

Medição de erro – para supermercados

Supermercado verifica a digital para dar descontos

Falso negativo é custoso;

Falso positivo é aceitável.



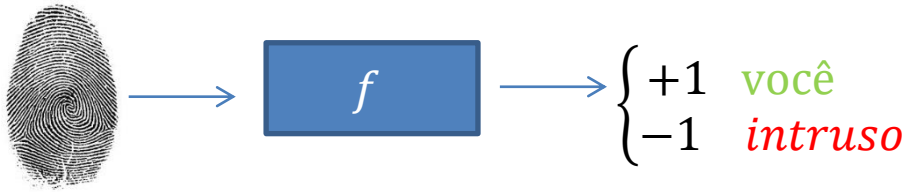
		f	
		+1	-1
h	+1	0	1
	-1	10	0

Medição de erro – para bancos

Banco verifica a digital para acessar conta

Falso positivo é um desastre;

Falso negativo é aceitável.



		f	
		+1	-1
h	+1	0	1000
	-1	1	0

Métricas de qualidade do Aprendizado de Máquina

Matriz de confusão

h

f	-1	+1
-1	#Verdadeiros Negativos (VN)	#Falsos Positivos (FP)
+1	#Falsos Negativos(FN)	#Verdadeiros Positivo (VP)

Exemplo:

	classificado não spam	classificado spam
não spam	1000 (VN)	150 (FP)
spam	50 (FN)	100 (VP)

Acurácia

Quão frequente o classificador está correto?

$$acuracia = \frac{VP + VN}{VP + VN + FP + FN}$$

Exemplo

	classificado não spam	classificado spam
não spam	1000 (VN)	150 (FP)
spam	50 (FN)	100 (VP)

$$acuracia = \frac{100 + 1000}{100 + 1000 + 50 + 150} = 85\%$$

Paradoxo da Acurácia

	classificado não spam	classificado spam
não spam	1150 (VN)	0 (FP)
spam	150 (FN)	0 (VP)

Classificador burro

$$acuracia = \frac{0 + 1150}{0 + 1150 + 0 + 150} = 88{,}4\%$$

Precisão

Daqueles que classifiquei como corretos, quantos efetivamente eram?

$$precisao^{+} = \frac{VP}{VP + FP} \qquad precisao^{-} = \frac{VN}{VN + FN}$$

Exemplo

	classificado não spam	classificado spam
não spam	1000 (VN)	150 (FP)
spam	50 (FN)	100 (VP)

$$precisao^{+} = \frac{100}{100 + 150} = 40{,}0\% \qquad precisao^{-} = \frac{1000}{1000 + 50} = 95\%$$

Precisão
Acurácia

Recall

Quando realmente é da classe X, o quão frequente você classifica como X?

$$recall^{+} = \frac{VP}{VP + FN} \qquad recall^{-} = \frac{VN}{VN + FP}$$

Exemplo

	classificado não spam	classificado spam
não spam	1000 (VN)	150 (FP)
spam	50 (FN)	100 (VP)

$$recall^{+} = \frac{100}{100 + 50} = 66\% \qquad recall^{-} = \frac{1000}{1000 + 150} = 87\%$$

F1 score

$$f1^{+} = \frac{2 \cdot precisao^{+} \cdot recall^{+}}{precisao^{+} + recall^{+}} \qquad f1^{-} = \frac{2 \cdot precisao^{-} \cdot recall^{-}}{precisao^{-} + recall^{-}}$$

$f1 =$ média ponderada $f1^{+}$ e $f1^{-}$

Exemplo

$$\begin{cases} precisao^{+} = 0{,}4 \\ recall^{+} = 0{,}66 \end{cases} \qquad \begin{cases} precisao^{-} = 0{,}95 \\ recall^{-} = 0{,}87 \end{cases}$$

$$f1^{+} = \frac{2 \cdot 0{,}4 \cdot 0{,}66}{0{,}4 + 0{,}66} = 0{,}49 \qquad f1^{-} = \frac{2 \cdot 0{,}95 \cdot 0{,}87}{0{,}95 + 0{,}87} = 0{,}91$$

$$f1 = \frac{150 \cdot 0{,}49 + 1150 \cdot 0{,}91}{150 + 1150} = 86\%$$

Valor financeiro esperado de um classificador

Valor esperado do Classificador Binário (-1,+1)

Tipos de acertos:

- Acerto 1 (a_1): classifica objeto como +1 e de fato ele é +1
- Acerto 2 (a_2): classifica objeto como -1 e de fato ele é -1

Tipos de erros:

- Erro 1 (e_1): classifica objeto como -1 e de fato ele é -1
- Erro 2 (e_1): classifica objeto como -1 e de fato ele é +1

$$v_e = p_{a_1} \cdot v_{a_1} - p_{e_1} \cdot v_{e_1} + p_{a_2} \cdot v_{a_2} - p_{e_2} \cdot v_{e_2}$$

p_{a_1} = ??

p_{e_1} = ??

p_{a_2} = ??

p_{e_2} = ??

Valor esperado do Classificador Binário (-1,+1)

Tipos de acertos:

- Acerto 1 (a_1): classifica objeto como +1 e de fato ele é +1
- Acerto 2 (a_2): classifica objeto como -1 e de fato ele é -1

Tipos de erros:

- Erro 1 (e_1): classifica objeto como -1 e de fato ele é -1
- Erro 2 (e_1): classifica objeto como -1 e de fato ele é +1

$$v_e = p_{a_1} \cdot v_{a_1} - p_{e_1} \cdot v_{e_1} + p_{a_2} \cdot v_{a_2} - p_{e_2} \cdot v_{e_2}$$

$$p_{a_1} = recall^+ \qquad p_{e_1} = 1 - recall^+$$

$$p_{a_2} = recall^- \qquad p_{e_2} = 1 - recall^-$$

Classificador de Crediário

Tipos de acertos:

- Acerto 1 (a_1): libera crediário ao cliente bom pagador
- Acerto 2 (a_2): nega crediário ao cliente caloteiro

Tipos de erros:

- Erro 1 (e_1): libera crediário ao cliente caloteiro
- Erro 2 (e_1): nega crediário ao cliente bom pagador

$$v_e = p_{a_1} \cdot v_{a_1} - p_{e_1} \cdot v_{e_1} + p_{a_2} \cdot v_{a_2} - p_{e_2} \cdot v_{e_2}$$

v_{a_1} ??

v_{a_2} ??

v_{e_1} ??

v_{e_2} ??

Classificador de Crediário

Tipos de acertos:

- Acerto 1 (a_1): libera crediário ao cliente bom pagador
- Acerto 2 (a_2): nega crediário ao cliente caloteiro

Tipos de erros:

- Erro 1 (e_1): libera crediário ao cliente caloteiro
- Erro 2 (e_1): nega crediário ao cliente bom pagador

$$v_e = p_{a_1} \cdot v_{a_1} - p_{e_1} \cdot v_{e_1} + p_{a_2} \cdot v_{a_2} - p_{e_2} \cdot v_{e_2}$$

v_{a_1} : valor mediano do empréstimo

v_{a_2} : R$0,0 (deixou de levar calote)

v_{e_1} : valor mediano do empréstimo

v_{e_2} : R$0,0 (deixou de ganhar na venda)

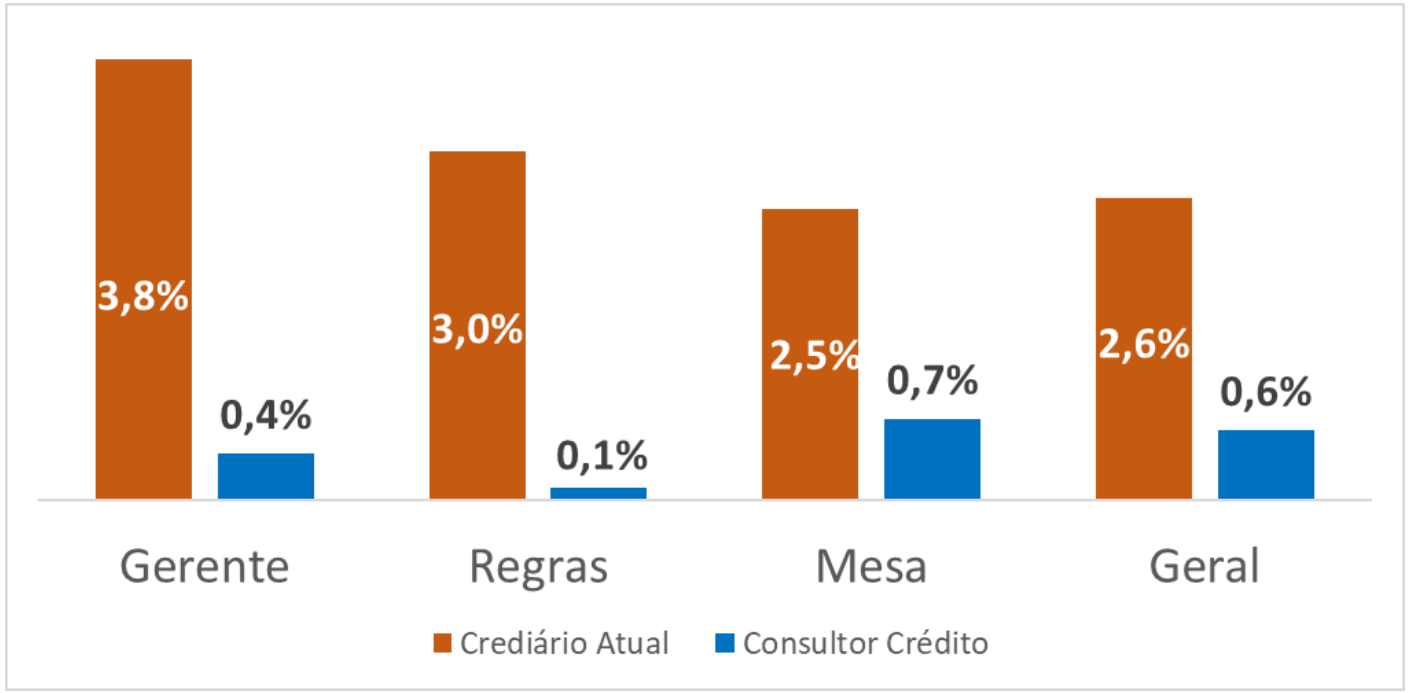
Estudo de caso: Crediário

Empresa de Varejo (base de crediário 2018-2021)

Estatística	Valor
nº orçamentos	179.262
nº aprovados	147.096
nº negados	32.166
Média de aprovados anual	**36.774**

Valor Orçamento	
média	R$ 1.188,70
desvio padrão	R$ 827,20
valor mínimo	R$ 16,00
1 quartil	R$ 500,38
2 quartil (mediana)	**R$ 1003,70**
3 quartil	R$ 1.609,95
valor máximo	R$ 24.142,90

Erro percentual na liberação de crédito



161.336
orçamentos testados

-2%
de redução do erro
pelo Consultor de Crédito

99,3%
Significância estatística do teste
(0,6 p.p.)

Valor esperado por uso do Classificador de Crédito

$$V_e = p_{acerto} \cdot mediana_{valor} - p_{erro} \cdot mediana_{valor}$$

Sistema	Valor Esperado (V_e)
Crediário Atual	R$947,90
Consultor de Crediário	R990,90

Ganho de R$43,00

Valor Esperado Anual do classificador

+4,5%
aumento na lucratividade

R$1.581.282,00
ganho esperado anual

R$ 43,00
ganho esperado por cada orçamento

36.774
orçamentos aprovados por ano