

Regressão Linear

Prática 08: Classificação

Prof^a Deborah Magalhães

Monitor: Davi Luis de Oliveira



UNIVERSIDADE
FEDERAL DO PIAUÍ

Olá!



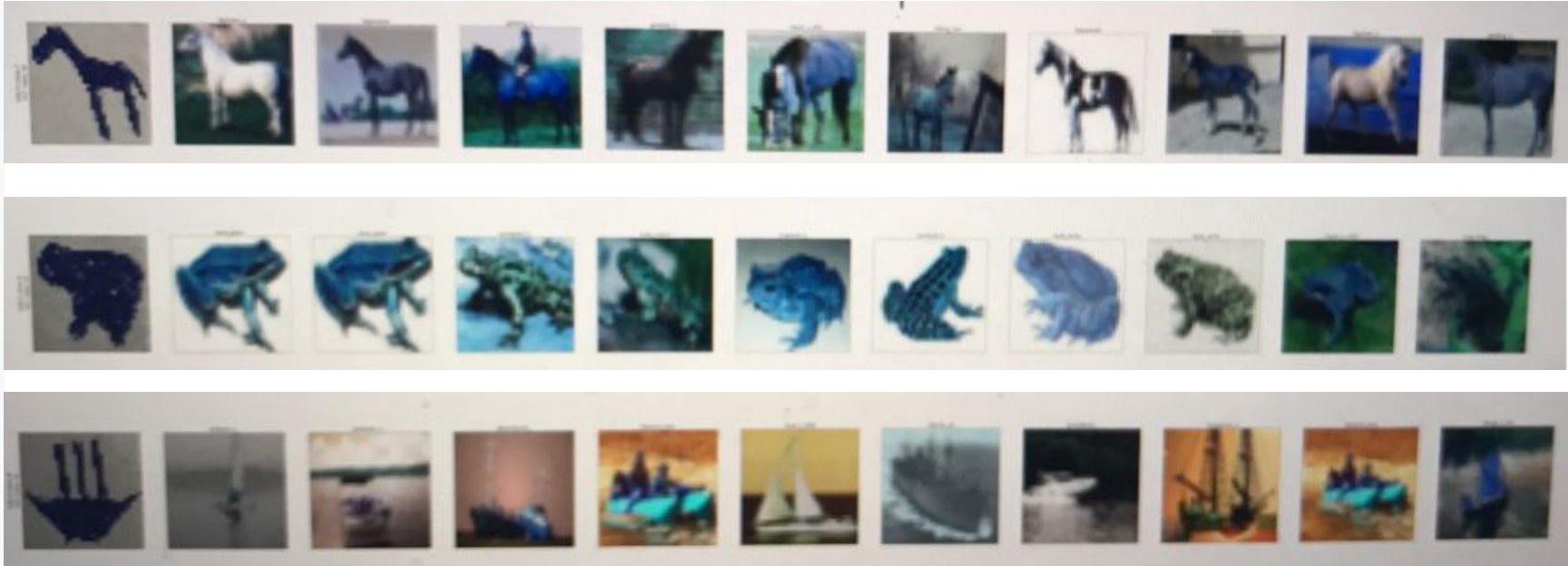
Curso: Bacharelado em Sistema de Informação

Disciplina: Sistemas Inteligentes

- ▷ **Aplicações de classificação**
- ▷ **Estudo de caso: Análise de Sentimento**

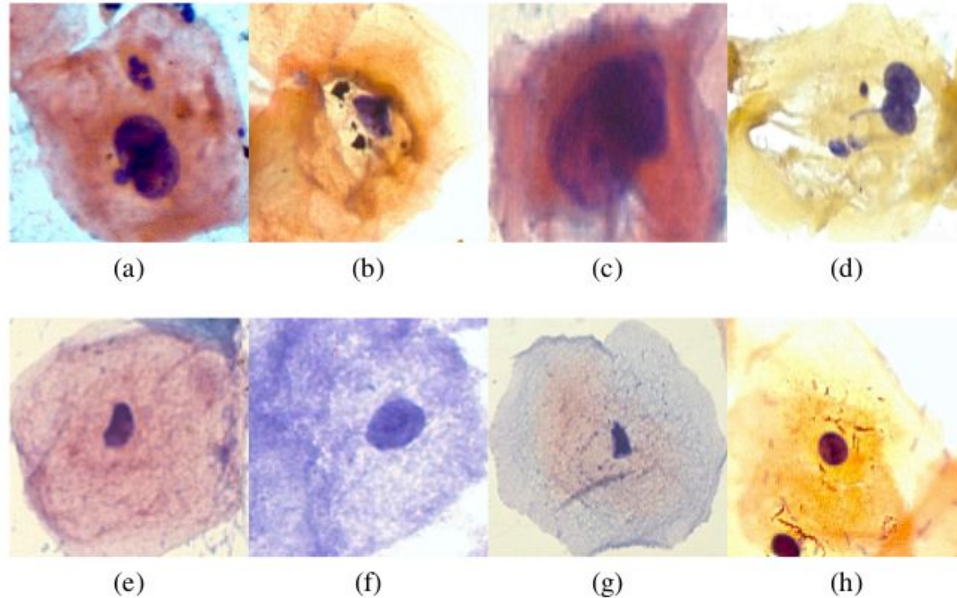
Você pode me encontrar em deborah.vm@gmail.com
(Dúvidas e sugestões serão bem-vindas =D)

Recuperação de imagens



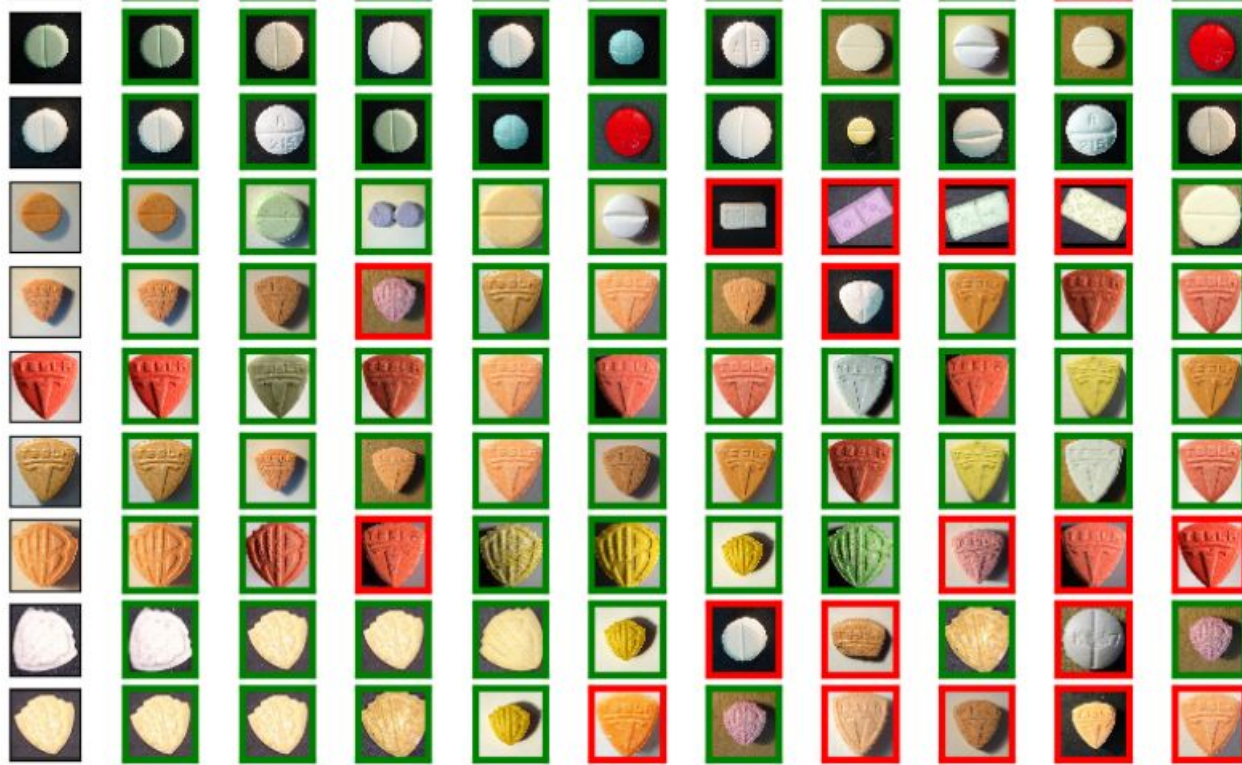
Recuperação de imagens por conteúdo a partir de desenhos feitos a mão

Auxílio ao diagnóstico médico



Imagens de células cervicais provenientes do exame de Papanicolau:
classificação binária anormal (a-d) e normal (e-h)

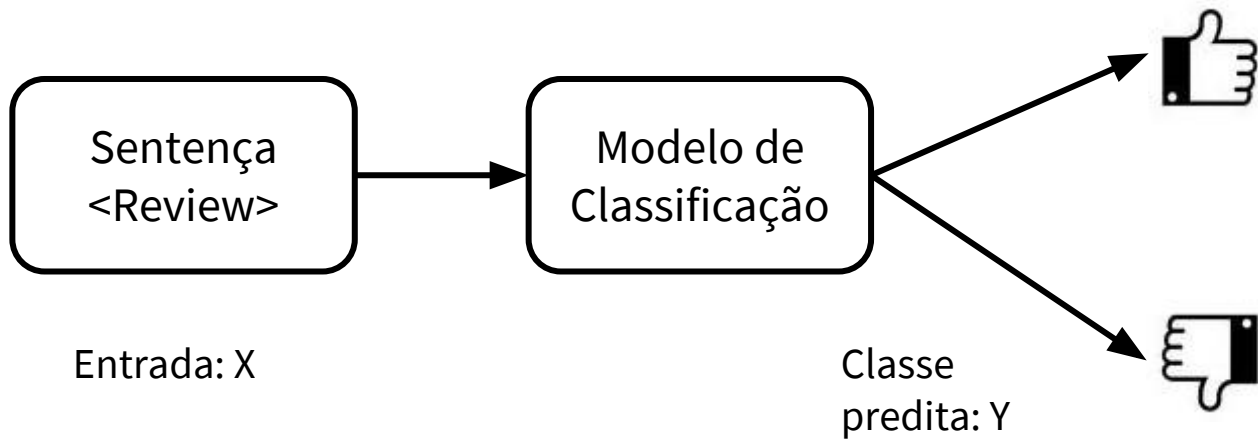
Detecção de pílulas ilícitas



Classificação binária(licita e ilícita)

Análise de sentimentos

6



Estudo de caso: Avaliação de um restaurante

Em geral, o que verificamos quando estamos procurando um restaurante?

”

*O que as pessoas falam
sobre a comida, preço,
ambiente, atendimento,
localização, prazo de
entrega*

”

Estudo de caso: Kazuco sushi

Kazuco Sushi

Japonesa • \$



ABRE ÀS 17:00

ENTREGA GRÁTIS

A Kazuco Sushi é a melhor opção de comida japonesa da região.

[Mais informações](#)



Estudo de caso: avaliações



Rogério

5,0 ★★★★★ 21/04/18

Otimo atendimento. entrega rapida.
comida gostosa!



Camille

5,0 ★★★★★ 22/03/18

Sushi maravilhoso! super recomendo



Marta

5,0 ★★★★★ 22/04/18

Ótimo

Estudo de caso: avaliações



André

3,0 ★★★★★ 27/04/18

A comida era boa mas o tempo de entrega não foi bacana.



Ana

5,0 ★★★★★ 12/04/18

o tempo de entrega esta de 80/95 minutos, acho muito tempo, até porque não é tão longe da minha casa, mas tirando isso tudo massa...gostei



Sidarta

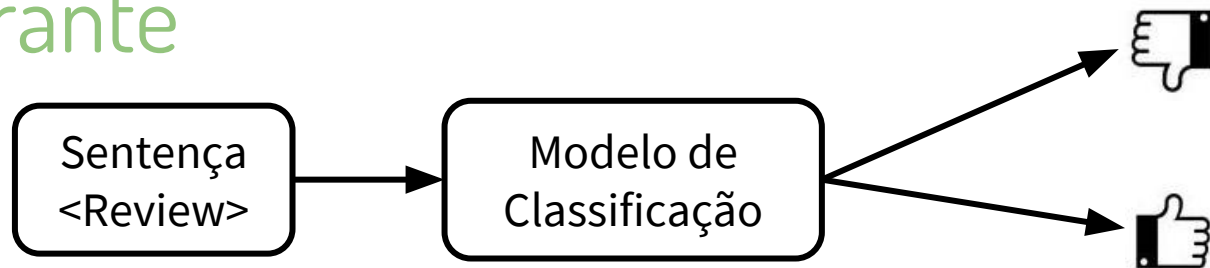
4,0 ★★★★★ 18/04/18

Sushi muito bom. Recomendo. Só não dei 5 estrelas porque não veio o wasabi que foi solicitado.

*O que podemos inferir das
avaliações?*

”

Estudo de caso: sistema de avaliação de restaurante



1. Todas as avaliações do restaurante

Camille: sushi maravilhoso.
Super recomendo.

Sidarta: Sushi muito bom.
Recomendo. Só não dei 5
estrelas porque não veio o
wasabi solicitado

2. Selecionar as sentenças com "sushi"

"Sushi maravilhoso"

"Sushi muito bom"

3. Classificador de sentimentos

Classificador **linear**
baseado em **limiar**

Estudo de caso: sistema de avaliação de restaurante

Palavras Positivas	Palavras Negativas
Bom, maravilhoso, gostosa, rápida	Não, demorado



Setenças

Entrada: X



Classificador baseado em limiar

Contar o número de palavras positivas & Contar o número de palavras negativas

IF (positivo > negativo) \hat{y} = 👍

ELSE \hat{y} = 👎

Estudo de caso: classificador baseado em limiar

Sentença 01: **Ótimo** atendimento. entrega **rapida**. comida **gostosa**.

Sentença 02: A comida era **boa** mas o tempo de entrega **não foi bacana**.

- As palavras possuem diferentes graus de sentimento.
 - Ótimo > boa
- Palavras sozinhas não são boas o suficiente para transmitir sentimentos

Estudo de caso: sistema de avaliação de restaurante

- ▷ O classificador linear utilizará os dados de treinamento para aprender o peso de cada palavra

Palavra	Peso
bom	+1.5
ótimo	+2.5
gostosa	+2.0
ruim	-1.5
Não foi bacana	-2.0
rapido	+3.0
demorado	-3.0
Tempo, restaurante, comida, atendimento, entrega	0.0

Estudo de caso: sistema de avaliação de restaurante

- ▷ Setença 01: **Ótimo** atendimento. entrega **rapida**. comida **gostosa**.

$$\hat{y} = + 2.5 + 3.0 + 2.0 = 7.5$$



Palavra	Peso
bom	+1.5
ótimo	+2.5
gostosa	+2.0
ruim	-1.5
Não foi bacana	-2.0
rapido	+3.0
demorado	-3.0
Tempo, restaurante, comida, atendimento, entrega	0.0

Estudo de caso: sistema de avaliação de restaurante

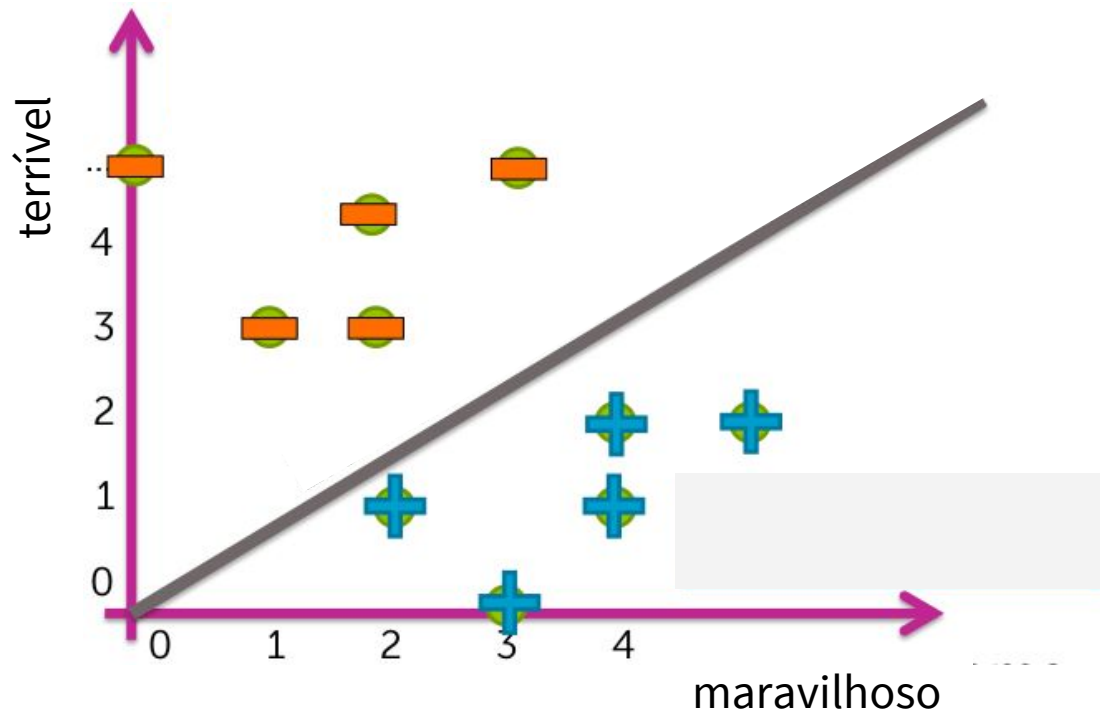
- ▷ Setença 02: A comida era **boa** mas o tempo de entrega **não foi bacana**.

$$\hat{y} = + 1.5 - 2.0 = - 0.5$$

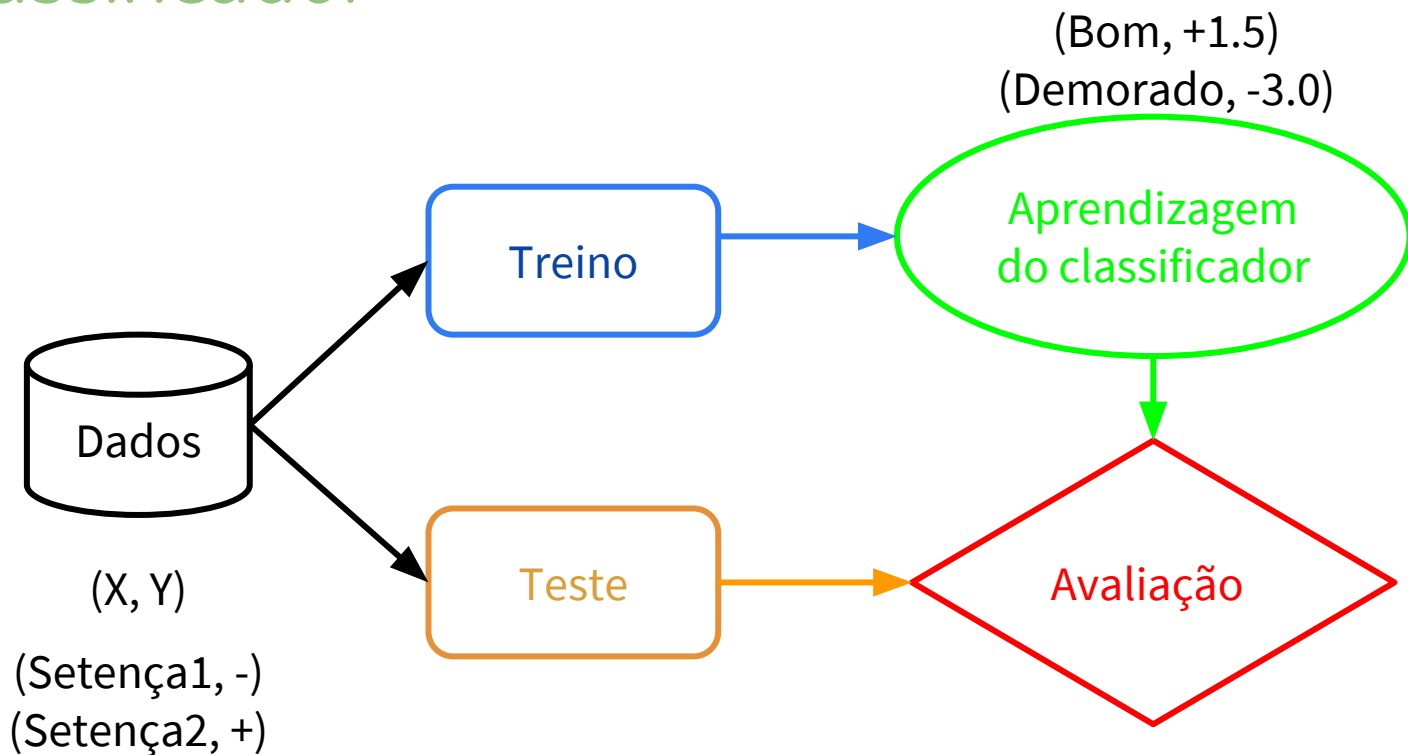


Palavra	Peso
bom	+1.5
ótimo	+2.5
gostosa	+2.0
ruim	-1.5
Não foi bacana	-2.0
rapido	+3.0
demorado	-3.0
Tempo, restaurante, comida, atendimento, entrega	0.0

Estudo de caso: fronteira de decisão



Estudo de caso: treinando e avaliando o classificador



Estudo de caso: avaliação do classificador

$$\text{Erro} = \frac{\text{número de erros}}{\text{tamanho teste}}$$

$$\text{Acurácia} = \frac{\text{número de acertos}}{\text{tamanho teste}}$$

$$\text{Erro} = 1 - \text{Acurácia}$$

y	\hat{y}	
+	+	correto
+	-	erro
-	+	erro
-	-	correto

Estudo de caso: avaliação do classificador

- ▶ Classificação binária (normal/anormal e positivo/negativo):
 - ▶ A acurácia do preditor deve ser melhor que 50%
- ▶ Para n classes, a acurácia deve ser maior que $1/n$
 - ▶ Se você tem 4 classes, a acurácia deve ser maior que 25%





Desbalanceamento: em 2010, 90% dos emails eram considerados spam

Estudo de caso: tipos de erros

		Predito (\hat{y})	
		+	-
Rótulo (y)	+	Verdadeiro Positivo	Falso Negativo
	-	Falso Positivo	Verdadeiro Negativo

Erros têm impactos distintos em aplicações diferentes

Estudo de caso: matriz de confusão

			
60		50	10
40		5	35

$$\text{Acurácia} = \frac{50 + 35}{100}$$

$$\text{Acurácia} = 0.85 \text{ ou } 85\%$$

Estudo de caso: matriz de confusão

	Saudável	Crise de garganta	Gripado
Saudável	60	8	2
Crise de Garganta	4	12	4
Gripado	0	2	8

$$\text{Acurácia} = \frac{60 + 12 + 8}{100}$$

$$\text{Acurácia} = 0.80 \text{ ou } 80\%$$

Prática:

”

Passo 1: Importar graphlab e carregar os dados

```
import graphlab  
produtos = graphlab.SFrame("amazon_baby.gl/")
```

Passo 1: saída

In [4]: produtos

Out[4]:

name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3.0
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0

Passo 2: contar as palavras do review

```
produtos["conta_palavra"] = graphlab.text_analytics.count_words(produtos["review"])
```

Passo 2: saída

In [6]: produtos

Out[6]:

name	review	rating	conta_palavra
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3.0	{'and': 5, '6': 1, 'stink': 1, 'because' ...
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0	{'and': 3, 'love': 1, 'it': 2, 'highly': 1, ...
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0	{'and': 2, 'quilt': 1, 'it': 1, 'comfortable': ...
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0	{'ingenious': 1, 'and': 3, 'love': 2, ...
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0	{'and': 2, 'parents!!': 1, 'all': 2, 'puppet.': ...
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0	{'and': 2, 'this': 2, 'her': 1, 'help': 2, ...
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0	{'shop': 1, 'noble': 1, 'is': 1, 'it': 1, 'as': ...

Passo 3: definir uma avaliação positiva e uma avaliação negativa

```
produtos = produtos[produtos["rating"] != 3]  
produtos["sentimento"] = produtos["rating"] >=4
```

Passo 3: saída


In [16]: `products.head()`

Out[16]:

name	review	rating	word_count	sentiment
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0	{ 'and': 3, 'love': 1, 'it': 2, 'highly': 1, ...	1
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0	{ 'and': 2, 'quilt': 1, 'it': 1, 'comfortable': ...	1
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0	{ 'ingenious': 1, 'and': 3, 'love': 2, ...	1
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0	{ 'and': 2, 'parents!!': 1, 'all': 2, 'puppet.': ...	1
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0	{ 'and': 2, 'cute': 1, 'help': 2, 'doll': 1, ...	1
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0	{ 'shop': 1, 'be': 1, 'is': 1, 'it': 1, 'as': ...	1
Baby Tracker® - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5.0	{ 'feeding.': 1, 'and': 2, 'all': 1, 'right': 1, ...	1


Passo 3: saída (histograma dos dados iniciais)

```
produtos["rating"].show(view="Categorical")
```

Value	Count	Percent	
5	107,054	58.33%	
4	33,205	18.092%	
3	16,779	9.142%	
1	15,183	8.273%	
2	11,310	6.162%	

Passo 3: saída (desconsiderando as avaliações 3.0)

```
produtos["rating"].show(view="Categorical")
```

Value	Count	Percent	
5	107,054	64.2%	
4	33,205	19.913%	
1	15,183	9.105%	
2	11,310	6.783%	

Passo 3: saída (desconsiderando as avaliações 3.0)

```
produtos["sentimento"].show(view="Categorical")
```

Value	Count	Percent	
1	140,259	84.112%	
0	26,493	15.888%	

[illegible]

Passo 4: saída do classificador logístico

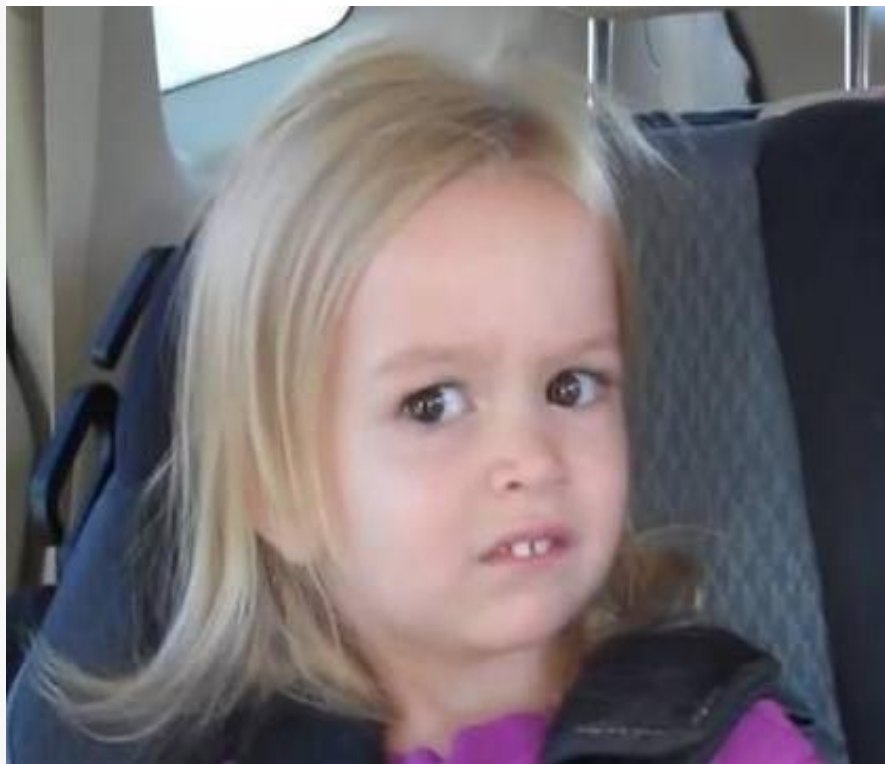
Iteration	Passes	Step size	Elapsed Time	Training-accuracy	Validation-accuracy
1	5	0.000002	1.888753	0.841481	0.839989
2	9	3.000000	3.737210	0.947425	0.894877
3	10	3.000000	4.475131	0.923768	0.866232
4	11	3.000000	5.244382	0.971779	0.912743
5	12	3.000000	6.005588	0.975511	0.908900
6	13	3.000000	6.839366	0.899991	0.825967
7	15	1.000000	8.057046	0.984548	0.921451
8	16	1.000000	8.806937	0.985118	0.921871
9	17	1.000000	9.546010	0.987066	0.919709
10	18	1.000000	10.326059	0.988715	0.916256

Passo 4: saída do classificador SVM

Iteration	Passes	Step size	Elapsed Time	Training-accuracy	Validation-accuracy
1	6	0.000001	2.213088	0.841481	0.839989
2	9	5.000000	3.654153	0.952603	0.900552
3	10	5.000000	4.406180	0.950970	0.889353
4	11	5.000000	5.147151	0.850391	0.785761
5	13	1.000000	6.265389	0.974042	0.918418
6	14	1.000000	7.090160	0.975459	0.918298
7	15	1.000000	7.913898	0.977602	0.918839
8	16	1.000000	8.774206	0.981536	0.920490
9	17	1.000000	9.575727	0.987141	0.918388
10	18	1.000000	10.410586	0.989037	0.915566

Atividade prática individual (valendo 2.5 pts)

- Apresentar a matriz de confusão para cada um dos classificadores
- Calcular a acurácia com base na matriz de confusão e verificar se dá próximo a acurácia de validação na 10 iteração para cada classificador
- Qual dos classificadores possuem mais FP e qual possui mais FN?



**Dúvidas? Sugestões?
Inquietações?
Aconselhamentos?**

- ▶ Desabafe em:
deborah.vm@gmail.com