



Usando Classificadores na Prática

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)

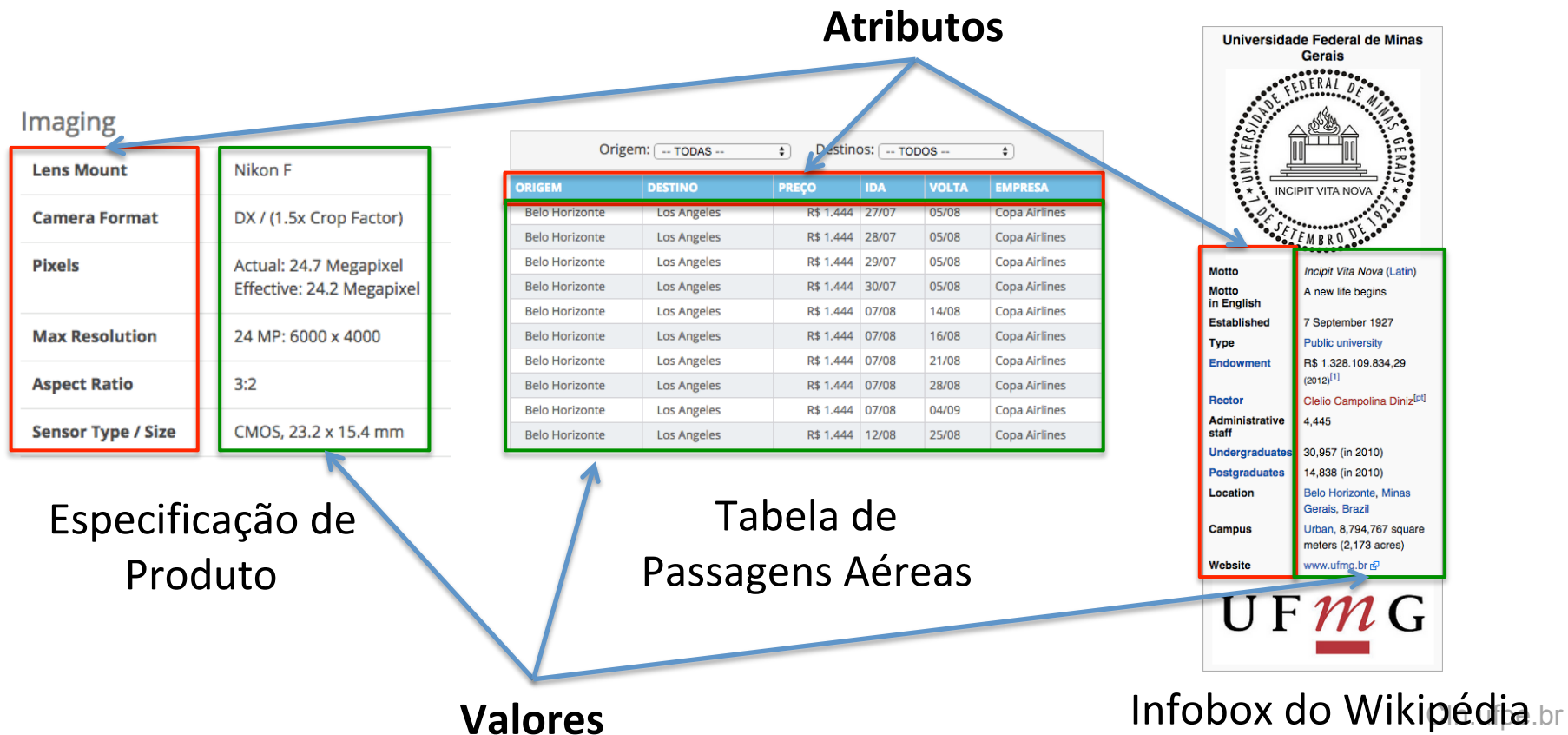


Projeto: Coleta e Busca de Entidades Estruturadas em um Domínio



Entidade Estruturada

- Def: objeto com atributos e valores associados
- Exemplos:





Benefícios: Melhoria na Busca por Entidades



citizen kane



Web

Images

Videos

Maps

News

More

Português

Sign in



gone with wind



Web

Images

Videos

Books

Shopping

More

Search tools

About 99,200,000 results (0.45 seconds)

Gone with the Wind (film) - Wikipedia, the free encyclopedia

[en.wikipedia.org/wiki/Gone_with_the_Wind_\(film\)](https://en.wikipedia.org/wiki/Gone_with_the_Wind_(film))

Gone with the Wind is a 1939 American epic historical romance film adapted from Margaret Mitchell's Pulitzer-winning 1936 novel. It was produced by David O. Vivien Leigh - Sidney Howard - Olivia de Havilland - Hattie McDaniel

Gone with the Wind - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Gone_with_the_Wind

Gone with the Wind is a novel written by Margaret Mitchell, first published in 1936. The story is set in Clayton County, Georgia, and Atlanta during the American ...
Gone with the Wind (film) - Margaret Mitchell - Scarlett - Rhett Butler's People

Gone with the Wind (1939) - IMDb

www.imdb.com/title/tt0031381/

★ ★ ★ ★ ★ Rating: 8.2/10 - 188,420 votes

Gone with the Wind - Margaret Mitchell's epic American classic! The Civil War- Gone with the Wind - Home video trailer with a sneak peak into the bonus ...

Full Cast & Crew - Awards - (1939) Poster - Trivia

Gone With the Wind(1939) - Rotten Tomatoes

www.rottentomatoes.com/m/gone_with_the_wind/

★ ★ ★ ★ ★ Rating: 95% - 79 votes

Critics Consensus: Filmed and presented on a scale not seen in modern productions, Gone with the Wind is, if not the definitive Hollywood film, then certainly ...

20 Things You Might Not Have Known About Gone with the ...

mentalfloss.com/.../20-things-you-might-not-have-known-about-gone-w...

Apr 11, 2014 - Gone with the Wind's original director was George Cukor, who had spent more than two years in planning and developing the film. Officially, he ...



Gone with the Wind

1939 film

★ ★ ★ ★ ★ 8.2/10 - IMDb

★ ★ ★ ★ ★ 95% - Rotten Tomatoes

Epic Civil War drama focuses on the life of petulant southern belle Scarlett O'Hara (Vivien Leigh). Starting with her idyllic on a sprawling plantation, the film traces her survival through the tragic history of the South during the Civil War and Reconstruction, and her tangled love affairs with Ash... More

Initial release: December 15, 1939 (Atlanta)

Directors: Victor Fleming, George Cukor, Sam Wood

Running time: 3h 58m

Music composed by: Max Steiner

Awards: Academy Award for Best Picture, more

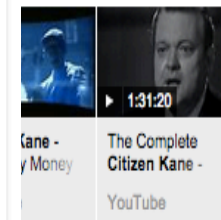
story - 119 min

Comingore, Agnes Moorehead, 1, news reporters scramble to ...

[lopedia](#)

y, co-written by, directed by and ture film. The film ...

le



Cidadão Kane



IMDb

8,4/10 ★ ★ ★ ★ ★

Rotten Tomatoes

100% ★ ★ ★ ★ ★

Citizen Kane é um filme norte-americano de 1941, dos gêneros Ação e Drama, dirigido por Orson Welles. Cidadão Kane foi o primeiro filme longametragem dirigido por Orson Welles, consider... +

pt.wikipedia.org

Summary: 1hr 59min · Drama

Release date: 16 de jun de 1941

Language: Língua inglesa

Director: Orson Welles

Screenwriters: Orson Welles · Herman J. Mankiewicz · John Houseman · Mollie Kent · Roger Q. Denny

Production companies: RKO Pictures · Mercury Productions

Watch now

Xbox Video



Benefícios: Busca Estruturada

O QUE VOCÊ PRECISA?

Comprar

Alugar

Lançamentos

Todos os imóveis

QUAL TIPO?

ONDE?

Ver todas as localidades

BELO HORIZONTE - MG

+

FAIXA DE PREÇO?

QUARTOS

SUÍTES

VAGAS

ÁREA (m²)

35.429 anúncios encontrados

Busca avançada

BUSCAR

Imóveis à venda em Belo Horizonte

Exibir como

Lista

Galeria

Pronto para Morar

A partir de **R\$ 805.246**

SION

Rua Patagônia
Belo Horizonte - Mg

Apartamento Pronto Para Morar
3 quartos | 1 suíte | 2 vagas |
102 a 189m²

[?]

Atualizado há 14 dias

Em Obras

A partir de **R\$ 329.000**

JARAGUA

Rua Professor Jerson Martins
Belo Horizonte - Mg

Apartamento em Obras
2 a 3 quartos | 1 suíte |
1 a 3 vagas | 64 a 176m²

[?]

Atualizado há 14 dias

LED & LCD TVs

CURRENT OFFERS

☐ On Sale (131)

☐ Free Shipping Eligible (362)

☐ Special Offers (228)

☐ Outlet Items (85)

TV TYPE

Clear

☒ LED (405)

☐ Smart (218)

☐ 4K UHD (79)

☐ Curved (17)

☐ OLED (3)

☐ 3D (71)

☐ Outdoor (47)

☒ LED Flat-Panel (15)

See More

TV SCREEN SIZE

Find Out Which TV Is Best for You

How do LED, plasma, OLED and LCD TVs compare? What resolution and refresh rate do you need? Simplify your TV search by understanding these and other features to consider.
[Learn more in the TV Buying Guide](#)

All Items (423)

Best Buy Items (233)

Marketplace Seller Items (190)

Sort by: Best Selling

Items per page: 15

< 1 2 3 ... 29 >

Filters: LED X LED Flat-Panel X LCD X LCD Flat-Panel X Clear All

Insignia™ - 32" Class (31-1/2" Diag.) - LED - 720p - HDTV - Black
Model: NS-32D312NA15 | SKU: 6080010

- 720p resolution
- 60Hz refresh rate
- ENERGY STAR Certified

★★★★★ 4.5 (1,529 Reviews)

PRICE MATCH GUARANTEE

\$159.99

ON SALE

SAVE \$20 (Reg. \$179.99)

Add to Cart

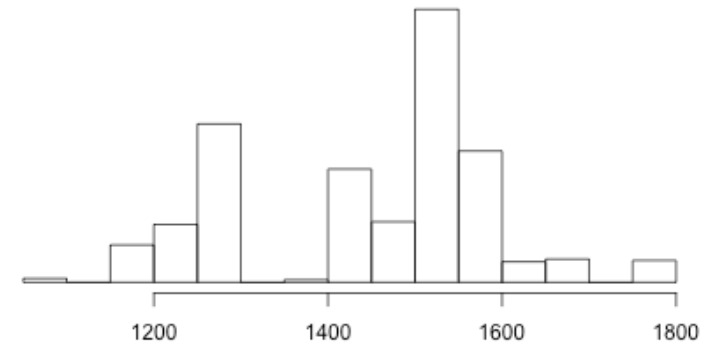


Benefícios: Análise Estatística

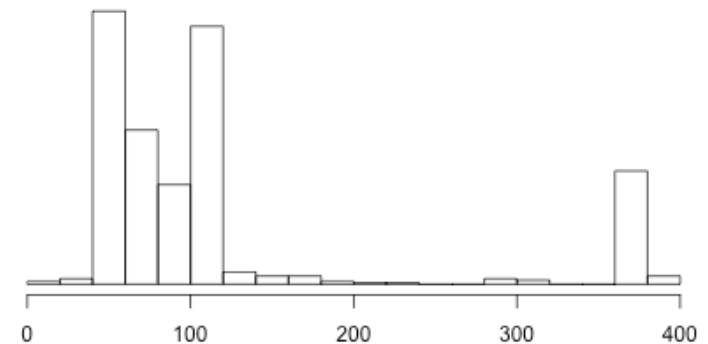
Origem: -- TODAS --		Destinos: -- TODOS --			
ORIGEM	DESTINO	PREÇO	IDA	VOLTA	EMPRESA
Belo Horizonte	Los Angeles	R\$ 1.444	27/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	28/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	29/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	30/07	05/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	14/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	16/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	21/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	28/08	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	07/08	04/09	Copa Airlines
Belo Horizonte	Los Angeles	R\$ 1.444	12/08	25/08	Copa Airlines



Ticket Price (Median = R\$ 1528 , Min = R\$ 1078)



Days prior to the trip (Median= 83 days)





Benefícios: Mercado de Dados



RESTAURANTS

43 restaurant specific attributes for restaurants of every type in the US, UK, France, Germany, and Australia.

[LEARN MORE](#)



DOCTORS

Database of over 1 million physician, dentist, and healthcare provider listings with the key data you need to make informed decisions.

[LEARN MORE](#)



HOTELS

Database of 140,000 hotel listings with over 35 attributes covering everything you need to know about a hotel.

[LEARN MORE](#)



Grande Interesse da Indústria

 Microsoft Azure

 Google fusion tables
beta

 factual

 DataMarket

infochimps




Sugestões de Tópico

- Produtos
 - Carros
 - Câmeras fotográficas
- Emprego
- Música



Chamada



Standing Queries

- Classificadores de texto escritos à mão
- Monitoramento de informação
- Executar uma consulta periodicamente para novas notícias em um tópico
- Usado para monitorar posts em mídia social
- Ex.: Google alerts
 - <https://www.google.com.br/alerts>



Exemplo de Classificação de Texto: Filtragem de Spam

From: "" <takworld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

=====

Click Below to order:

<http://www.wholesaledaily.com/sales/nmd.htm>

=====



Aprendizado Supervisionado

- Objetivo: inferir uma função a partir de exemplos dados para predizer classes de novos exemplos
- Duas fases:
 - Treinamento: aprende a função a partir de exemplos
 - Execução: usa a função para predizer a classe de um exemplo dado



Modelo Supervisionado

- Conjunto de treinamento: instâncias e rótulos
- Instância representada por seu vetor de características: x_i
- Aprender função $f(x)=y$ que melhor prediz o valor de y dado x
- Para y categórico -> classificação
- Para y numérico -> regressão

Conjunto de Treinamento

	viagra	learning	the	dating	nigeria	<i>spam?</i>
$\vec{x}_1 = ($	1	0	1	0	0)	$y_1 = 1$
$\vec{x}_2 = ($	0	1	1	0	0)	$y_2 = -1$
$\vec{x}_3 = ($	0	0	0	0	1)	$y_3 = 1$

↑
Categórico



Modelo Supervisionado: Exemplo

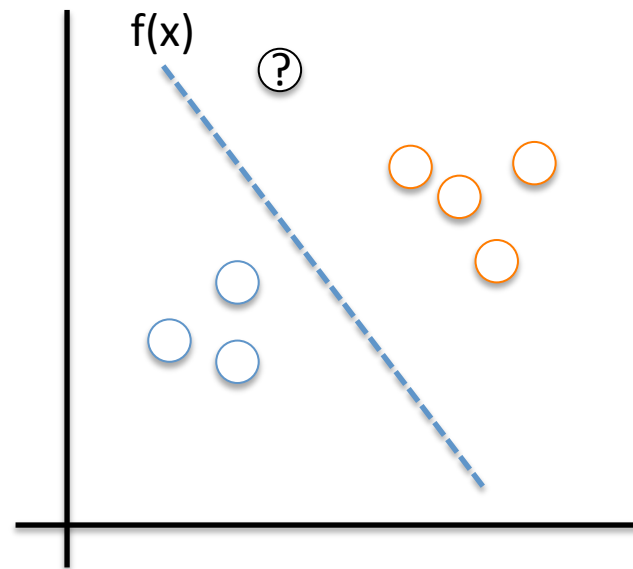
- Filtragem de spam

	viagra	learning	the	dating	nigeria	<i>spam?</i>
$\vec{x}_1 = ($	1	0	1	0	0)	$y_1 = 1$
$\vec{x}_2 = ($	0	1	1	0	0)	$y_2 = -1$
$\vec{x}_3 = ($	0	0	0	0	1)	$y_3 = 1$

- Features:
 - Palavras: viagra, learning, the, dating nigeria
 - Presença ou ausência
- Classe y : spam (+1) ou não spam (-1)



Resultado do Modelo Supervisionado





Features

- Grande importância no resultado da classificação
- Importante: alta correlação com o saída da classificação
 - Ex1: previsão de chuva: temperatura, humidade
 - Ex2: análise de sentimentos: palavras com polaridade (negativa/positiva)
- Classificadores podem usar qualquer tipo de feature
 - Palavras, pontuação, capitalização etc



Features: Bag of Words (BofW)

- Mais usado para texto: bag of words
 - Usa as tokens do documento

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

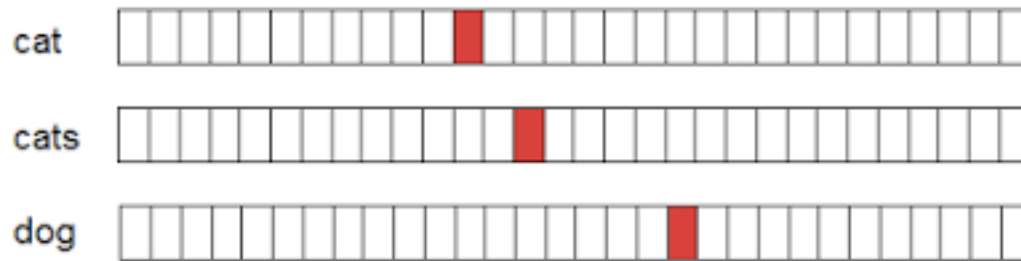


great	2
love	2
recommend	1
laugh	1
happy	1
...	...



One-hot Encoding

- Dimensionality: tamanho do vocabulário
- Problema: palavras similares têm representações diferentes

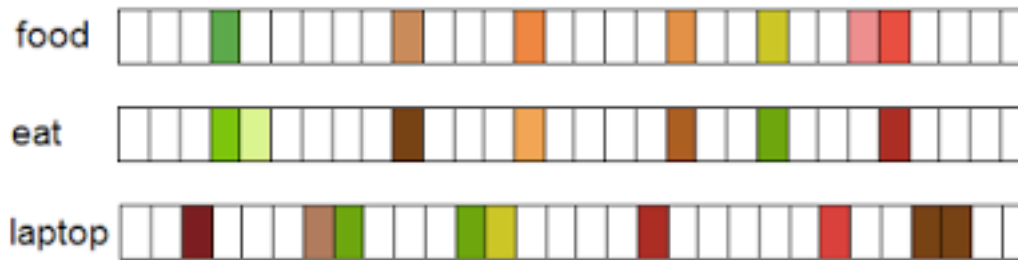


Img-Source: <http://veredshwartz.blogspot.com.br/2016/01/representing-words.html>



Features: Word Embeddings

- Palavra representada por um vetor denso
- Mapeia semântica a um espaço geométrico (embedding space)
- Encapsula o contexto de uma palavra para um vetor de pequena dimensionalidade (ex., 100, 200)
- Palavras similares estão próximas no espaço



Img-Source: <http://veredshwartz.blogspot.com.br/2016/01/representing-words.html>

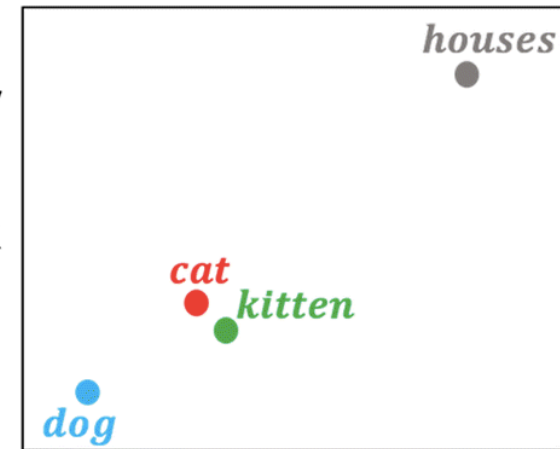
- Construído usando técnicas de redução de dimensionalidade
 - Redes neurais (word2vec)
 - Fatorização de matrizes (Latent Semantic Indexing)



Exemplos de Word Embeddings

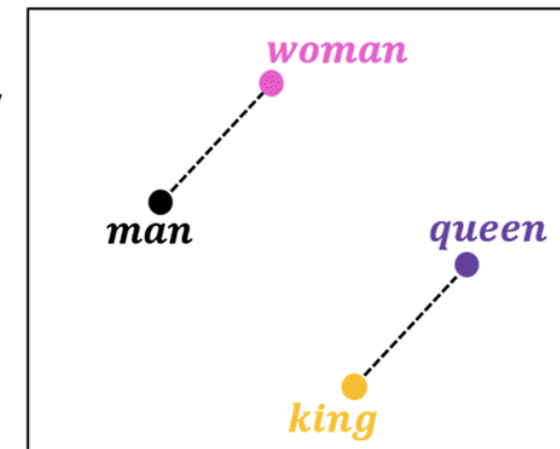
	living being	feline	human	gender	royalty	verb	plural
<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Dimensionality
reduction of
word
embeddings
from 7D to 2D



<i>man</i> →	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i> →	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i> →	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i> →	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Dimensionality
reduction of
word
embeddings
from 7D to 2D



Word

Word embedding

Dimensionality
reduction

Visualization of word
embeddings in 2D



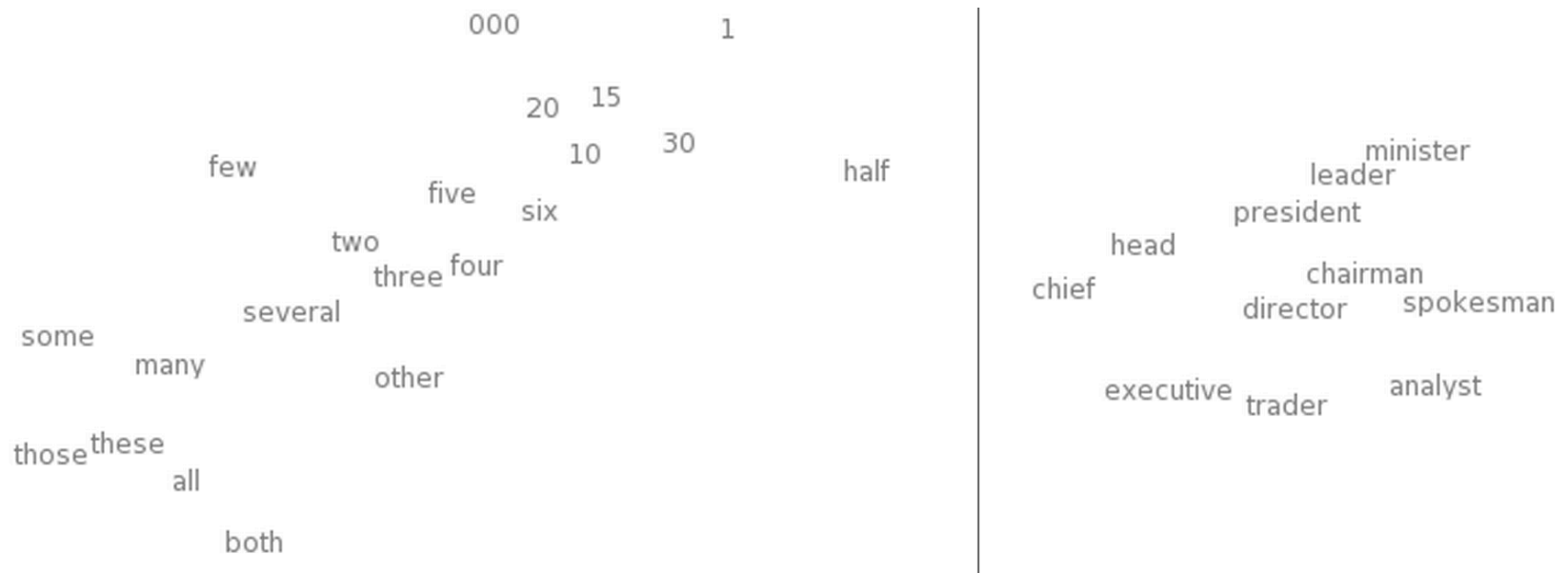
Exemplo de Word Embeddings

- Palavras mais próximas

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES



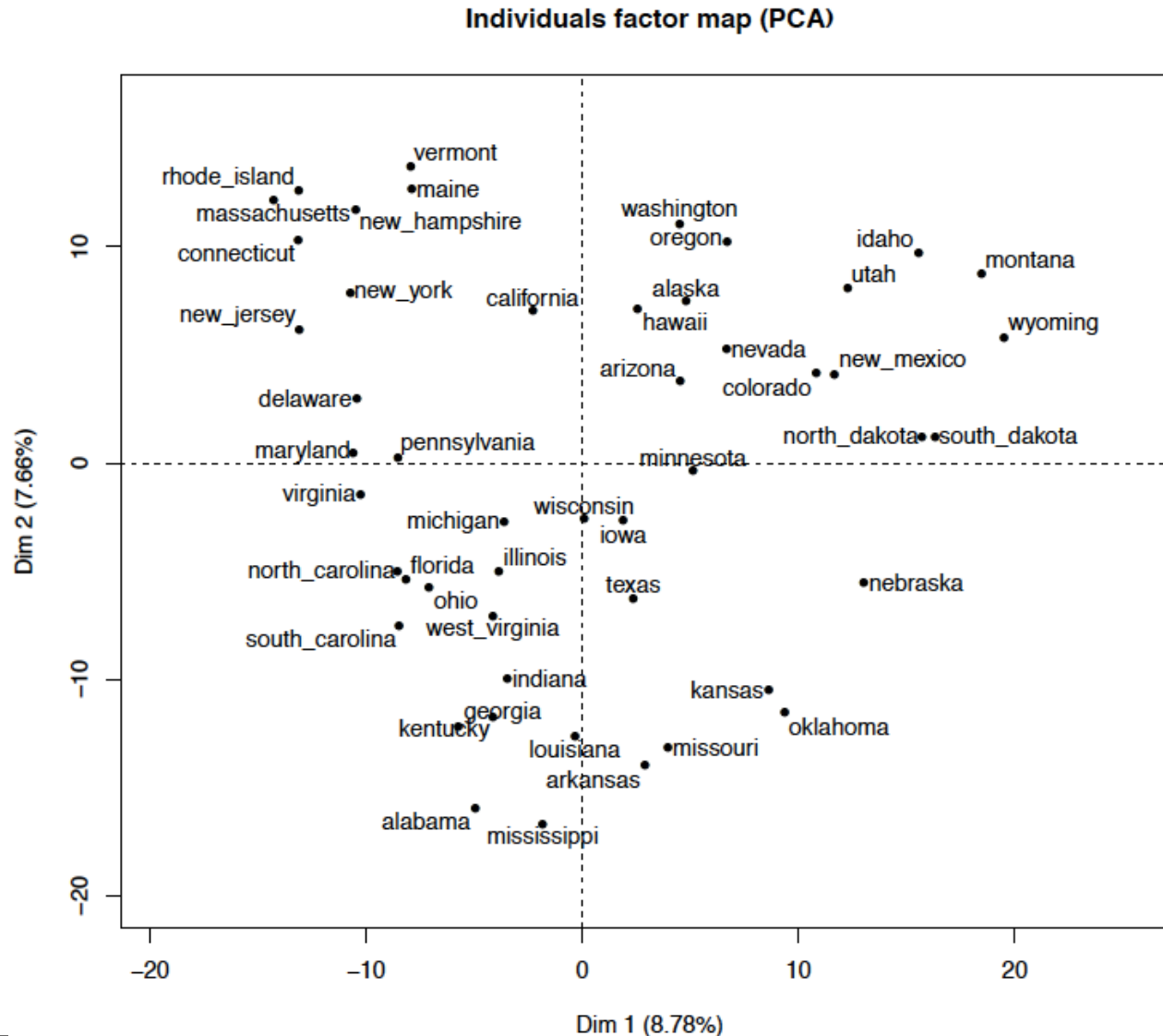
Word Embeddings em 2D



Img-Source: http://metaoptimize.s3.amazonaws.com/cw-embeddings-ACL2010/embeddings-mostcommon.EMBEDDING_SIZE=50.png

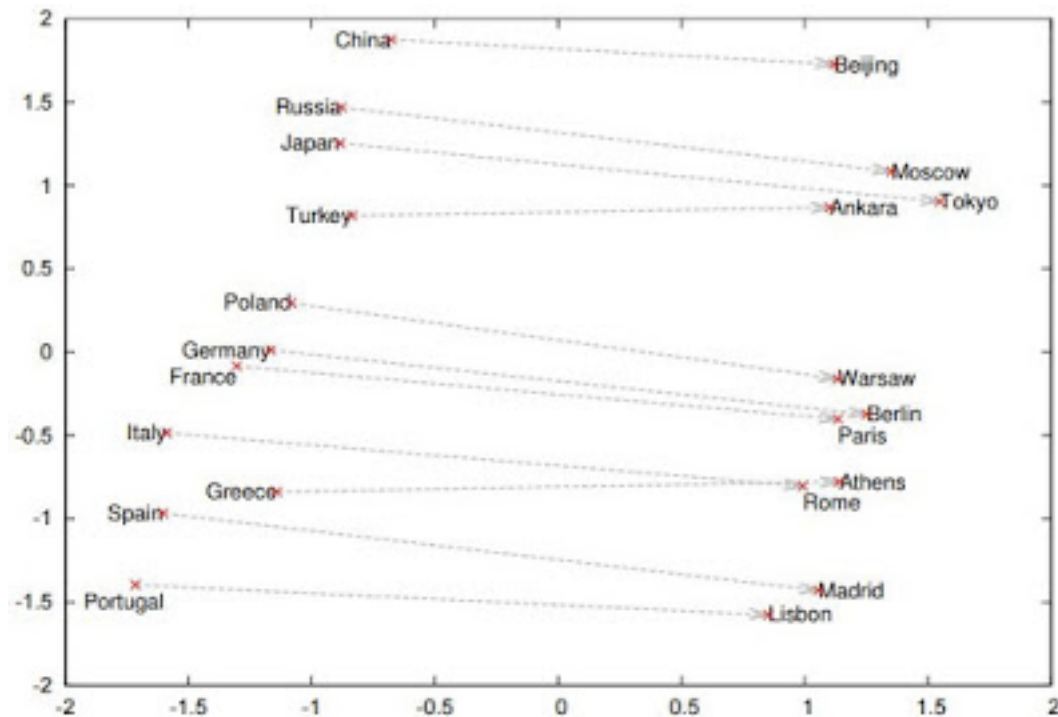


Word Embeddings em 2D





Relações em Word Embeddings



Img-Source: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>



Modelos Existentes

- Palavras
 - Word2Vec: <https://code.google.com/p/word2vec/>
 - GloVe: <http://nlp.stanford.edu/projects/glove/>
 - FastText: <https://fasttext.cc/>
- Sentenças/documentos
 - Doc2Vec: <https://radimrehurek.com/gensim/models/doc2vec.html>
 - BERT: <https://github.com/google-research/bert>



Feature Selection: Por quê?

- Coleções de texto têm um número grande de features
 - 10.000 a 1.000.000 palavras únicas (ou mais)
- Alguns classificadores não conseguem trabalhar com muitas features
- Reduz tempo de treinamento
 - Para alguns métodos tempo de treinamento é quadrático (ou pior) no número de features
- Torna o tempo de classificação mais rápido
- Pode melhorar generalização (evitar overfitting)



Feature Selection: Estratégias

- Filter
 - Independente do algoritmo de classificação
 - Baseado em medidas de teoria da informação, dependência estatística etc
- Wrapper
 - Usa um algoritmo de classificação
 - Calcula a acurácia do classificador criado com o conjunto selecionado de features



Feature Selection: Frequência

- Frequência
 - Método mais simples
 - Usa os termos mais comuns
 - Na prática, cerca de 90% tão bom qto os melhores métodos
- Exemplo de features para spam
 - Palavras: viagra, cialis
 - Frase: impress ... girl
 - From: inicia com números
 - Subject: todo maiúsculo
 - SpamAssassin
 - http://spamassassin.apache.org/old/tests_3_3_x.html



Feature Selection: Information Gain

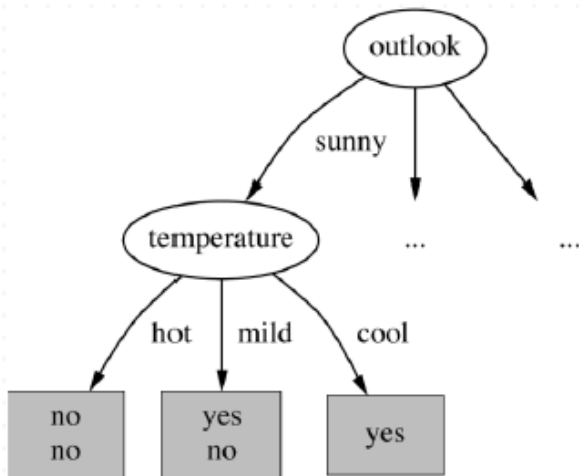
- Quão bem uma feature divide as classes

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	<i>No</i>
Sunny	Hot	High	True	<i>No</i>
Overcast	Hot	High	False	<i>Yes</i>
Rainy	Mild	High	False	<i>Yes</i>
Rainy	Cool	Normal	False	<i>Yes</i>
Rainy	Cool	Normal	True	<i>No</i>
Overcast	Cool	Normal	True	<i>Yes</i>
Sunny	Mild	High	False	<i>No</i>
Sunny	Cool	Normal	False	<i>Yes</i>
Rainy	Mild	Normal	False	<i>Yes</i>
Sunny	Mild	Normal	True	<i>Yes</i>
Overcast	Mild	High	True	<i>Yes</i>
Overcast	Hot	Normal	False	<i>Yes</i>
Rainy	Mild	High	True	<i>No</i>

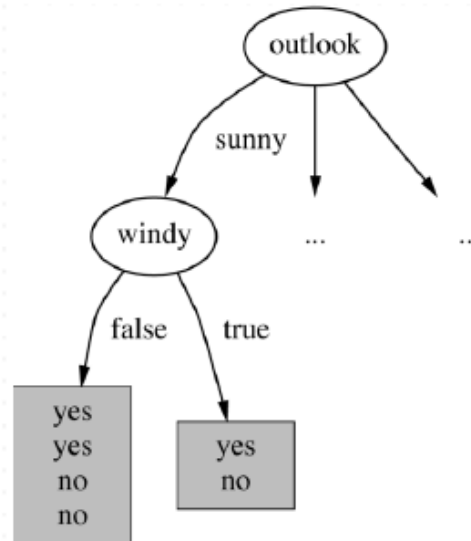


Feature Selection: Information Gain

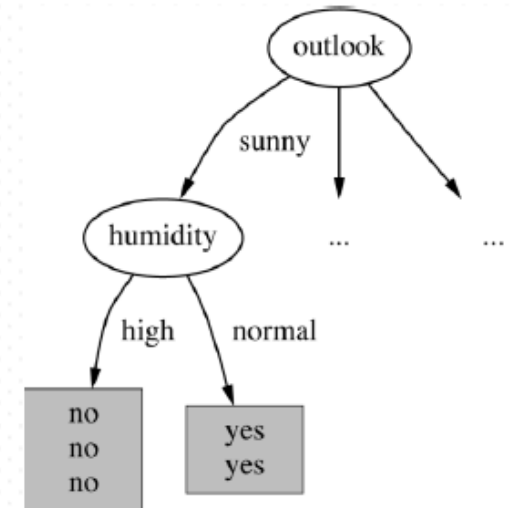
- Quão bem uma feature divide as classes



Temperature = 0.571



Windy = 0.020

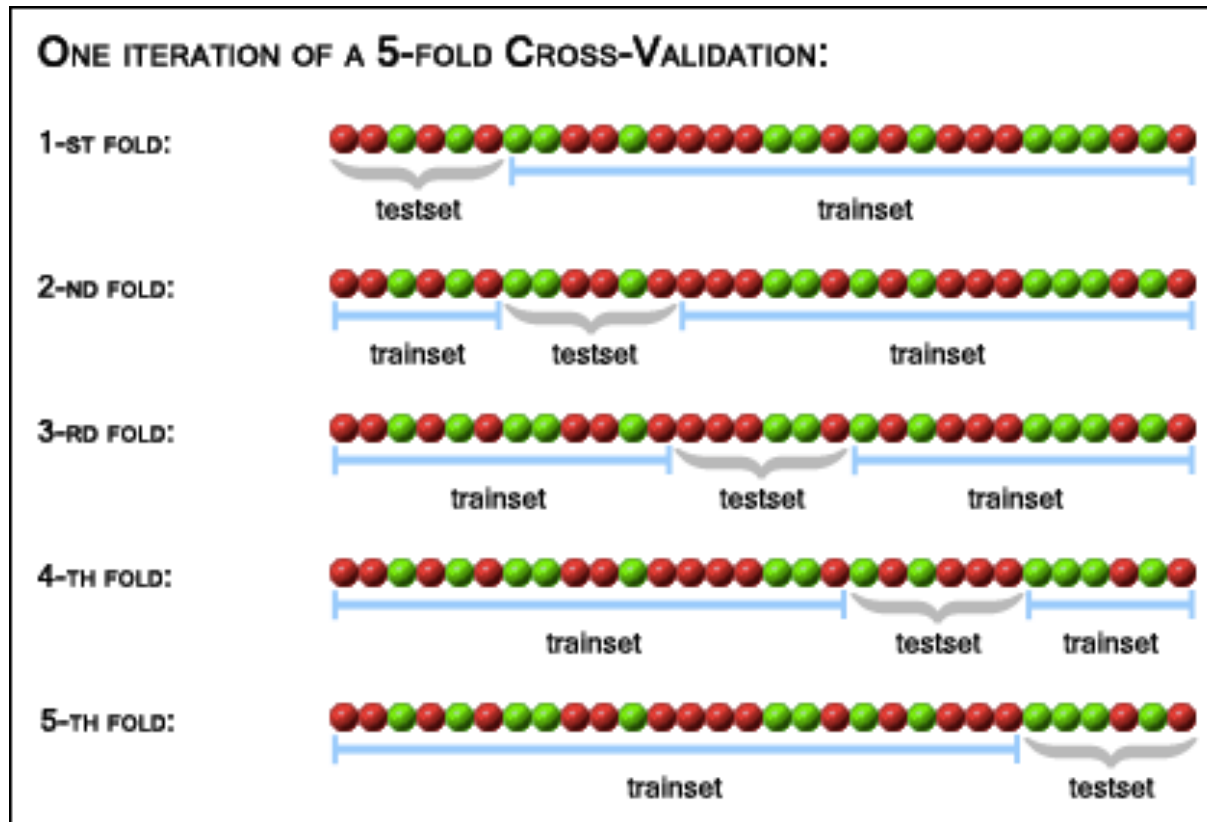


Humidity = 0.971



Avaliação do Modelo

- Conjunto de teste (holdout)
- Validação cruzada





Medidas de Avaliação

- Precision
- Recall
- F1 (F-measure)
- Accuracy

	in the class	not in the class
predicted to be in the class	true positives (TP)	false positives (FP)
predicted to not be in the class	false negatives (FN)	true negatives (TN)

$$\text{precision: } P = TP / (TP + FP)$$

$$\text{recall: } R = TP / (TP + FN)$$

$$F_1 = \frac{1}{\frac{1}{2} \frac{1}{P} + \frac{1}{2} \frac{1}{R}} = \frac{2PR}{P + R}$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FN + FP)$$

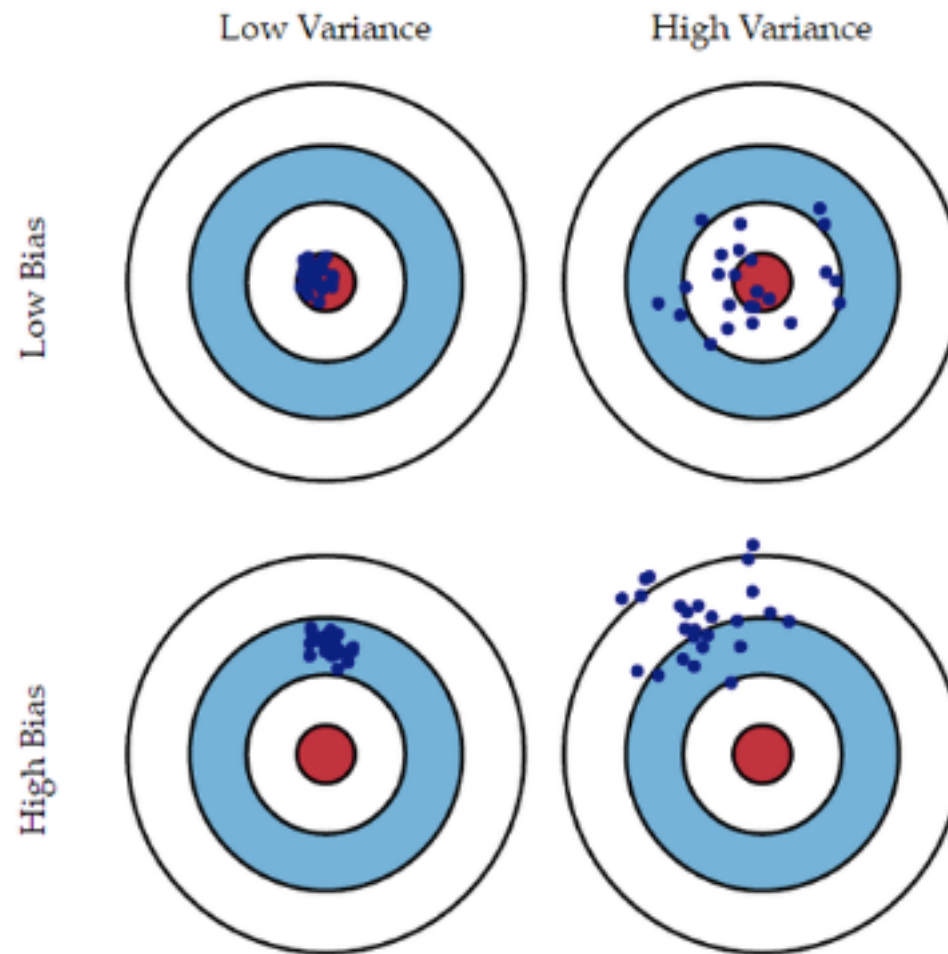


Diagnóstico de Modelos

- Como melhorar o classificador?
 - Features, dados etc
- Exemplo:
 - Alvo: 5% de erro
 - Erro no treinamento: 15% (**viés: $15-5=10$**)
 - Erro no teste: 16% (**variância: $16-15=1$**)
 - Precisa melhorar o desempenho no conjunto de treinamento
- Viés:
 - Desempenho no conjunto de treinamento
 - Depende do alvo
- Variância: diferença de desempenho entre treinamento e teste



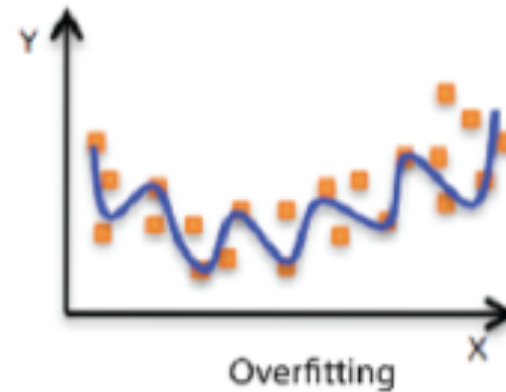
Viés e Variância





Overfitting

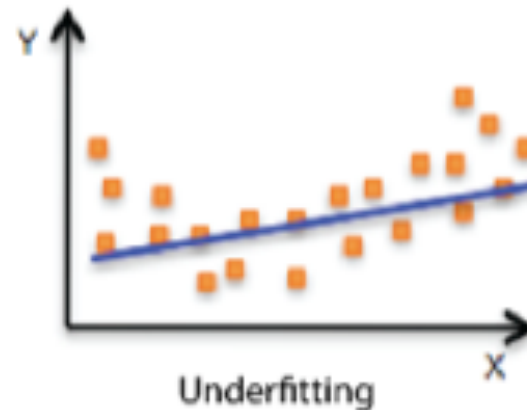
- Bom desempenho no conjunto de treinamento
- Problema em generalizar
- Baixo viés e alta variância
- Exemplo
 - Erro no treinamento: 1%
 - Erro no teste: 11%





Underfitting

- Modelo não modela bem o conjunto de treinamento
- Alto viés e baixa variância
- Exemplo
 - Erro no treinamento: 15%
 - Erro no teste: 16%





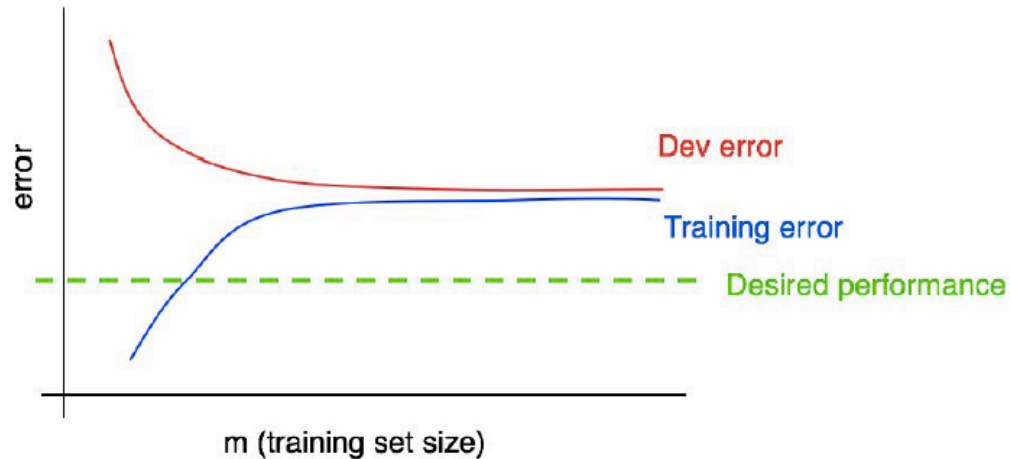
Outros Cenários

- Underfitting e overfitting
 - Exemplo
 - Erro no treinamento: 15%
 - Erro no teste: 30%
- Ideal
 - Exemplo
 - Erro no treinamento: 0.5%
 - Erro no teste: 1%



Lidando com Viés

- Alto viés (underfitting)

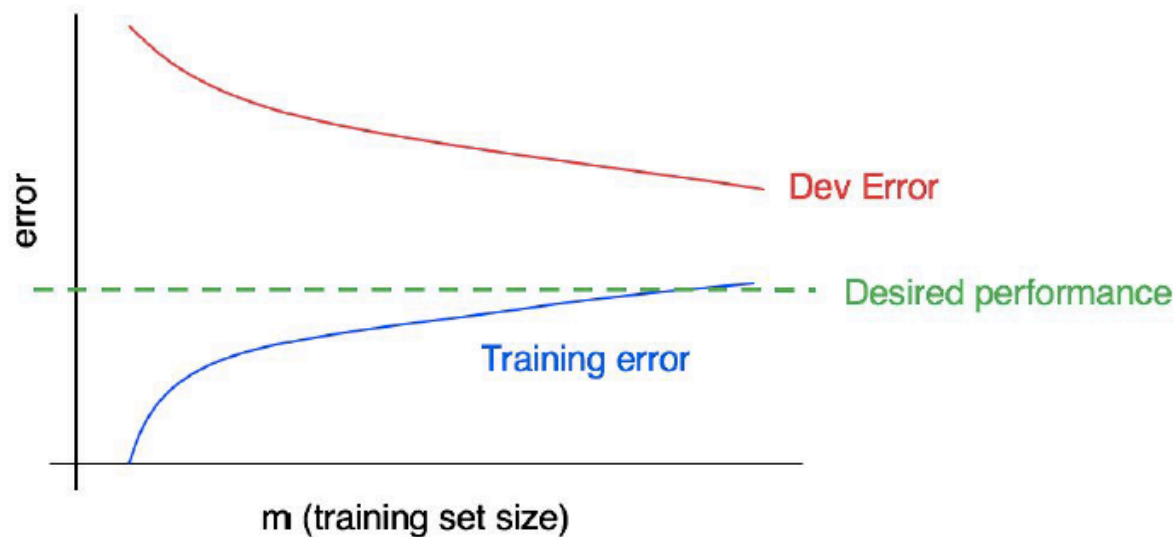


- Aumentar a complexidade do modelo
- Mais features
- Não ajuda adicionar mais dados ao treinamento



Lidando com Variância

- Alta variância (overfitting)



- Adicionar dados ao conjunto de treinamento
- Feature selection



Procedimento para Criação de Classificadores

- Definir features
- Obter dados rotulados
 - Pares (x_i, y_i) , onde x é um vetor de features e y é o rótulo
- Separar em 3 grupos
 - Treinamento (grande, ex.: 70%)
 - Validação (menor, ex.: 10%)
 - Teste (pequeno, ex.: 20%)



Procedimento para Criação de Classificadores

- Usar software para treinar o classificador -> conjunto de treinamento
- Usar validação para escolher melhores parâmetros do classificador (model selection)
- Avaliar no conjunto de teste
 - Accuracy
 - Precision
 - Recall
 - F-measure
- Fazer diagnóstico do modelo
- Building a machine learning application:
<http://docs.aws.amazon.com/machine-learning/latest/dg/building-machine-learning.html>



Ferramentas

- Python: Scikit-learn
 - <http://scikit-learn.org/stable/>
- Java: Weka
 - <http://www.cs.waikato.ac.nz/ml/weka/>