Usando Classificadores na Prática

Prof. Luciano Barbosa

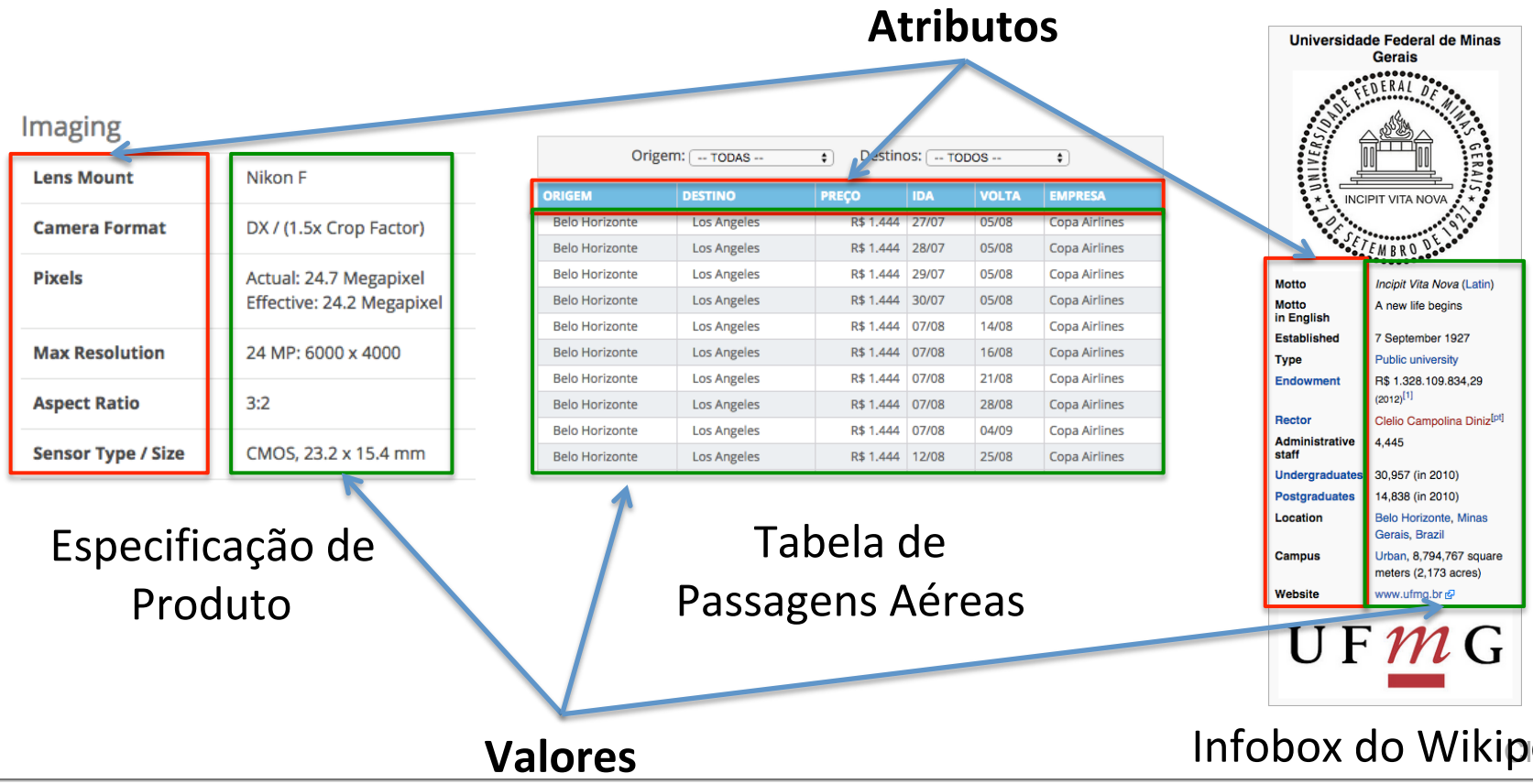
(Parte do material retirado dos slides dos livros adotados)

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

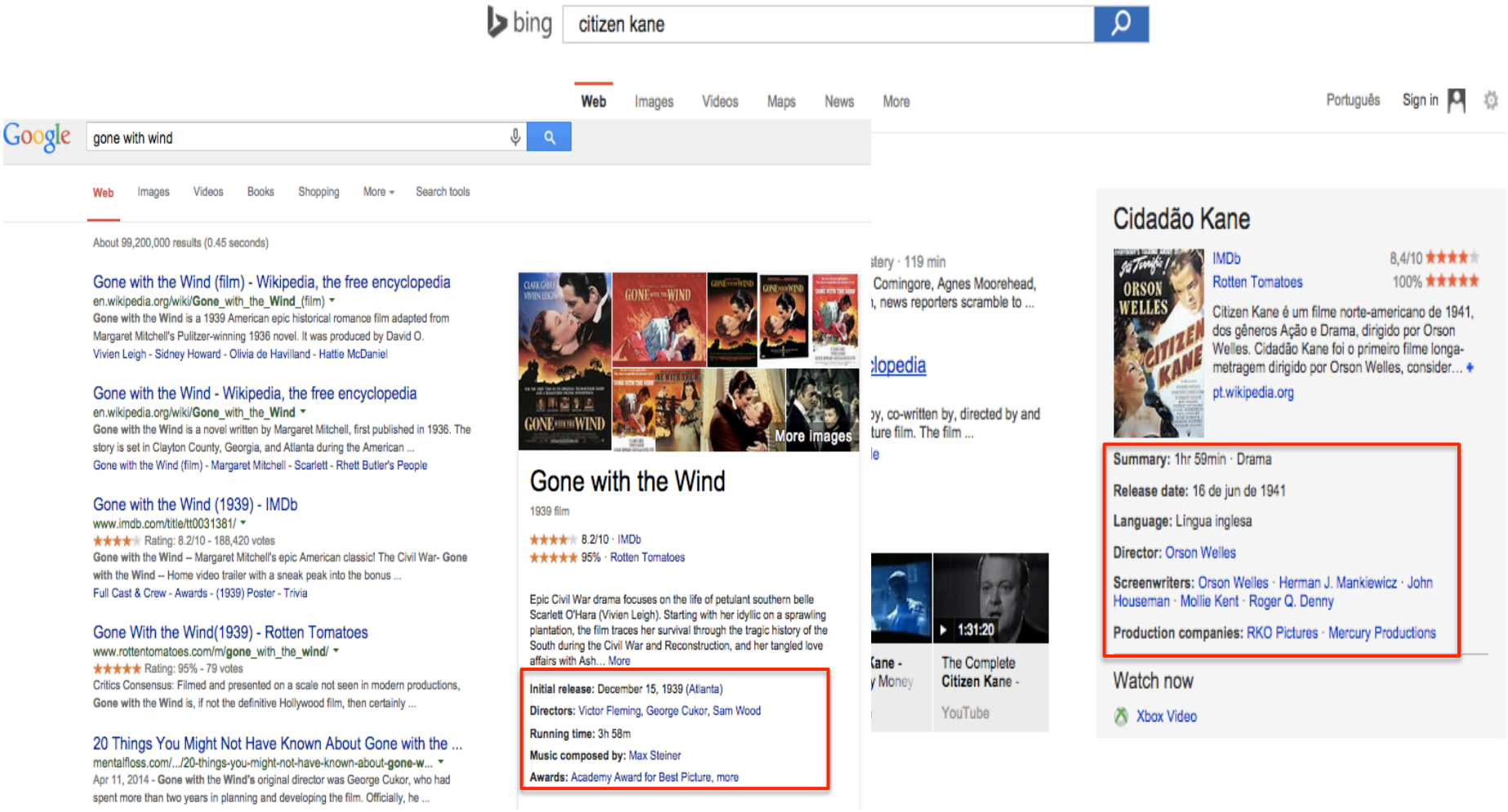
Projeto: Coleta e Busca de Entidades Estruturadas em um Domínio

Entidade Estruturada

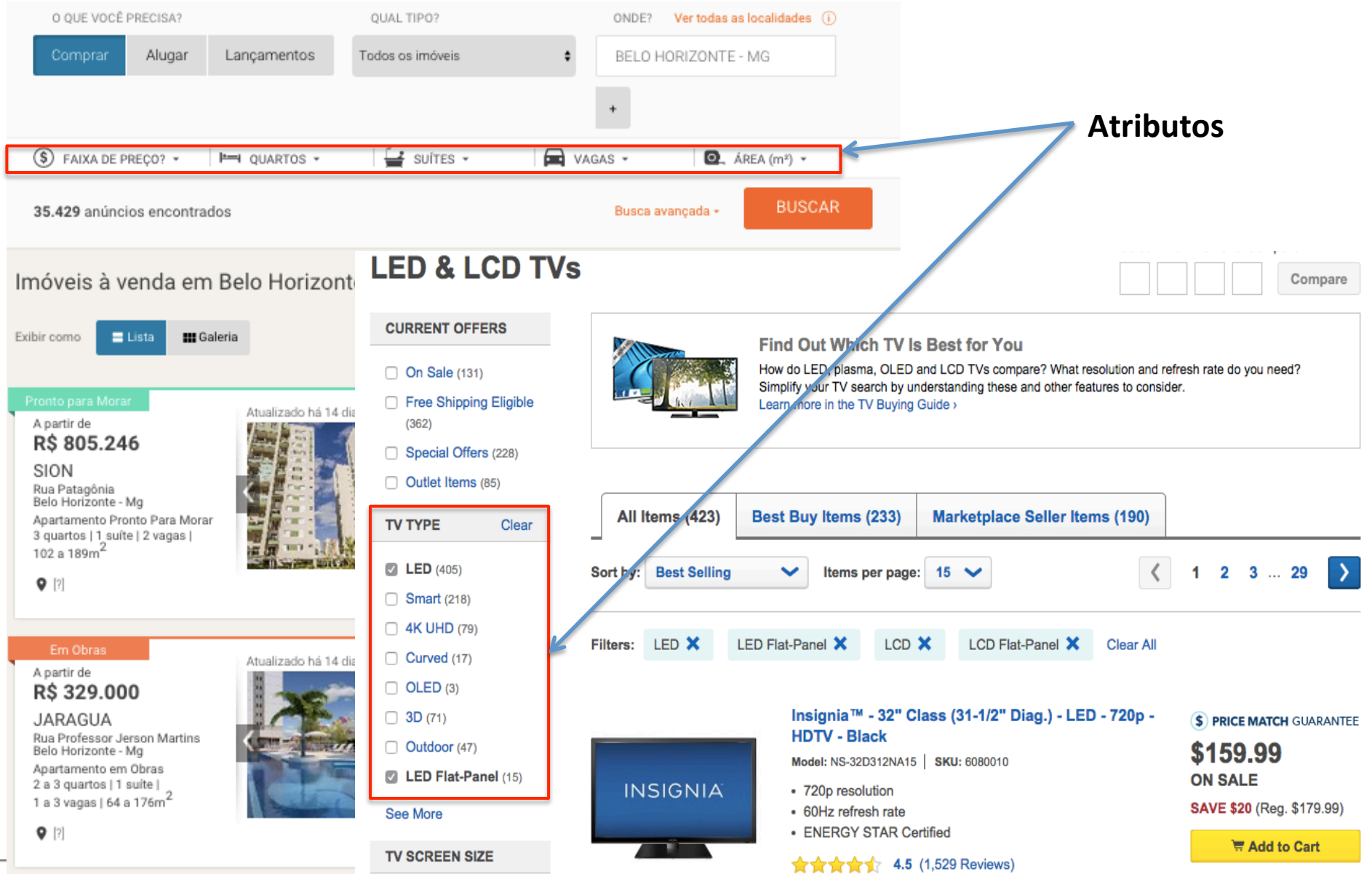
- Def: objeto com atributos e valores associados
- Exemplos:



Benefícios: Melhoria na Busca por Entidades



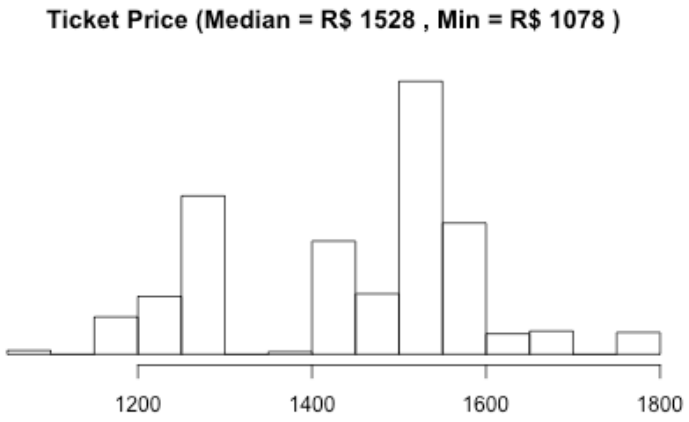
Benefícios: Busca Estruturada

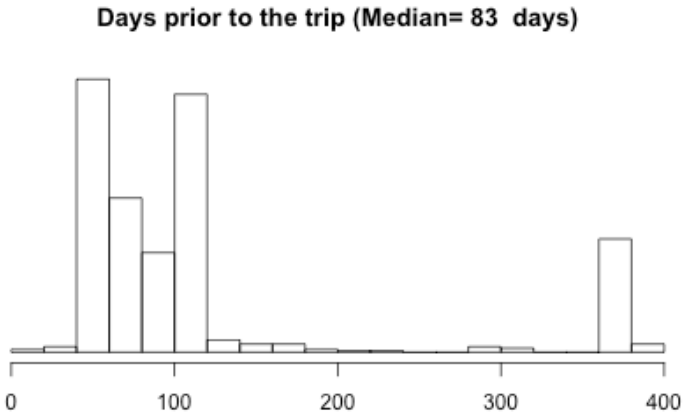


Benefícios: Análise Estatística

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





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

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: "" <takworlld@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	*spam*?
$\vec{x}_1 = ($	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

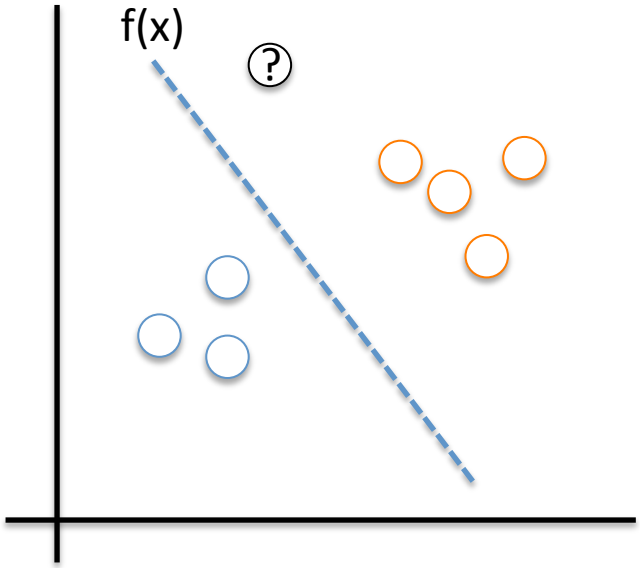
Modelo Supervisionado: Exemplo

- Filtragem de spam

	viagra	learning	the	dating	nigeria	*spam*?
$\vec{x}_1 = ($	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)

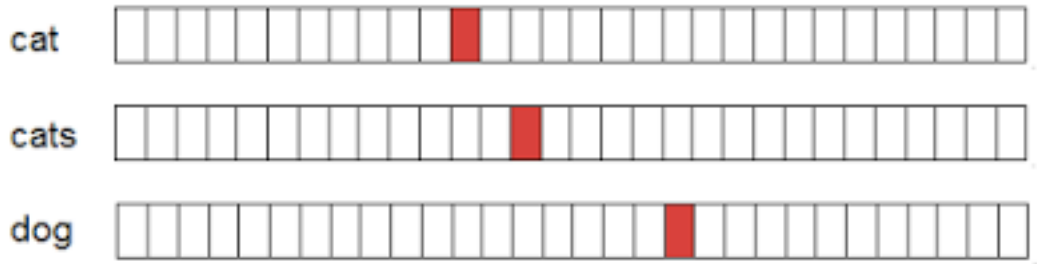
- Mais usado para texto: bag of words
  - Usa as tokens do documento

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

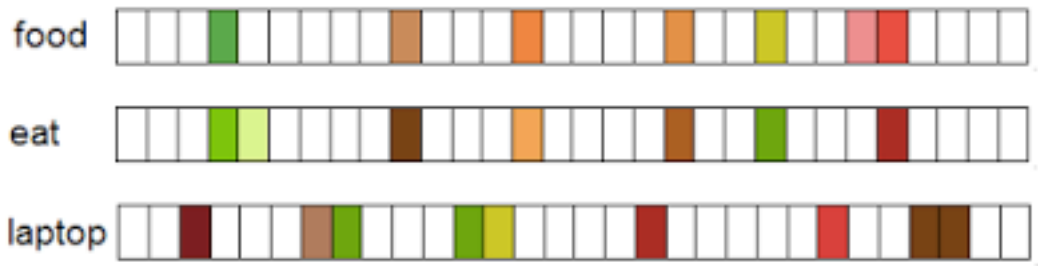
- Dimensionality: tamanho do vocabulário
- Problema: palavras similares têm representações diferentes



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

Features: Word Embeddings

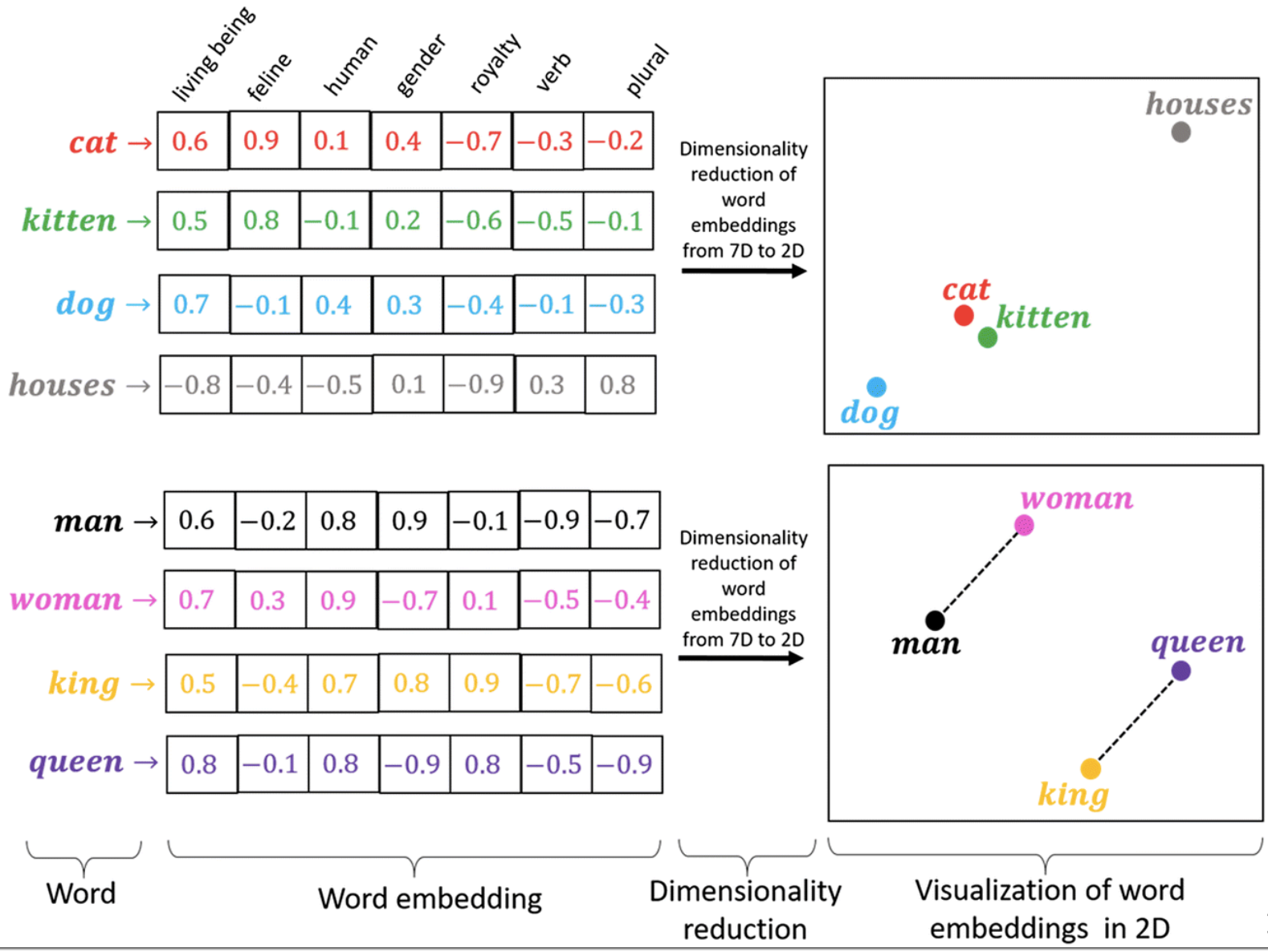
- Palavra representada por um vetor denso
- Mapeia semântica a um espaço geométrico (embedding space)
- Encapsula o contexto de uma palava para um vetor de pequena dimensionalidade (ex., 100, 200)
- Palavras similares estão próximas no espaço



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



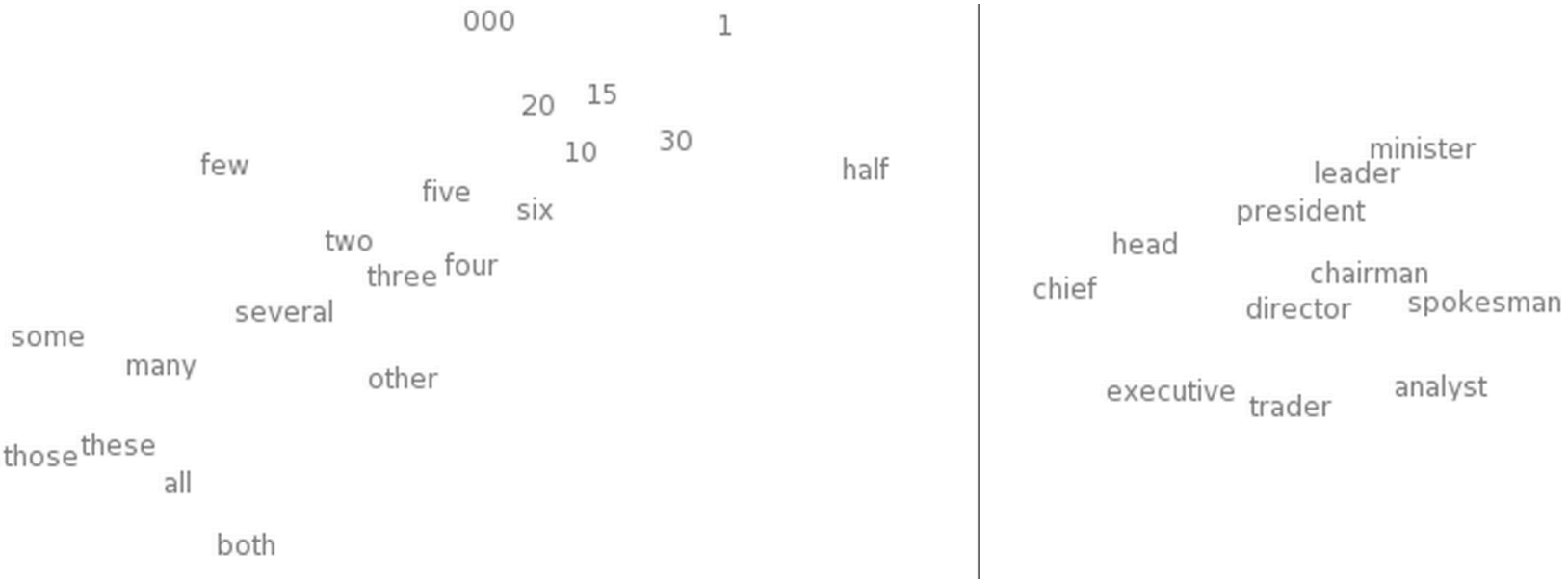
Examplo de Word Embeddings

- 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

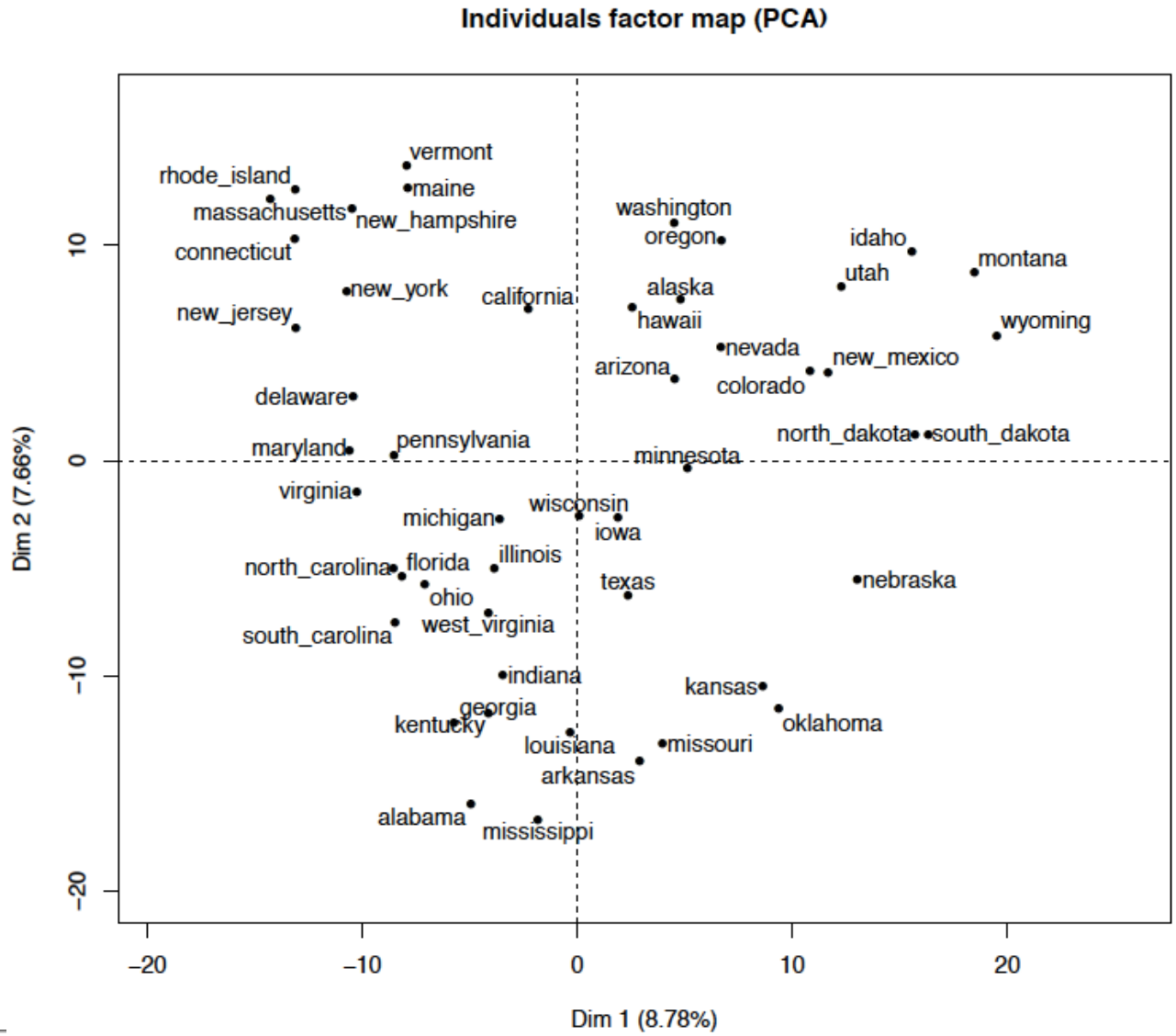
Img-Source: http://arxiv.org/pdf/1301.3781.pdf

Word Embeddings em 2D

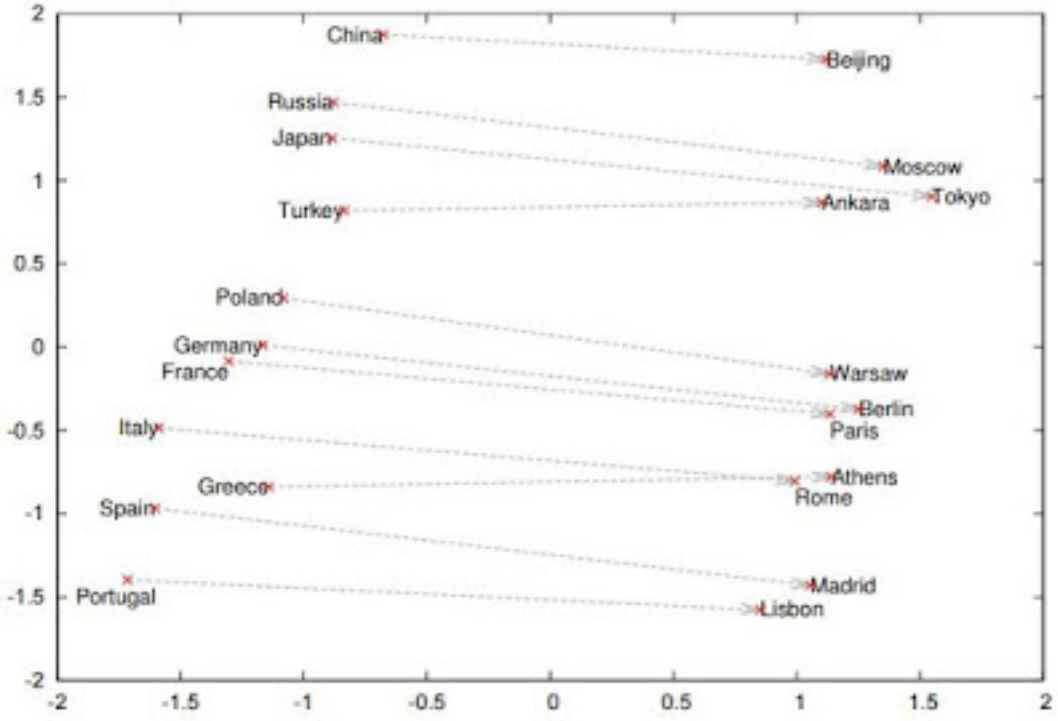


Img-Source: http://metaoptimize.s3.amazonaws.com/cw-embeddings-ACL2010/embeddings-mostcommon.EMBEDDING_SIZE=50.png

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

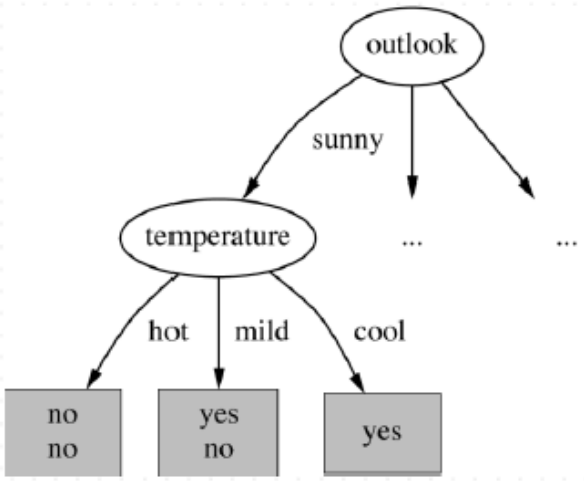
Feature Selection: Information Gain

- Quão bem uma feature divide as classes

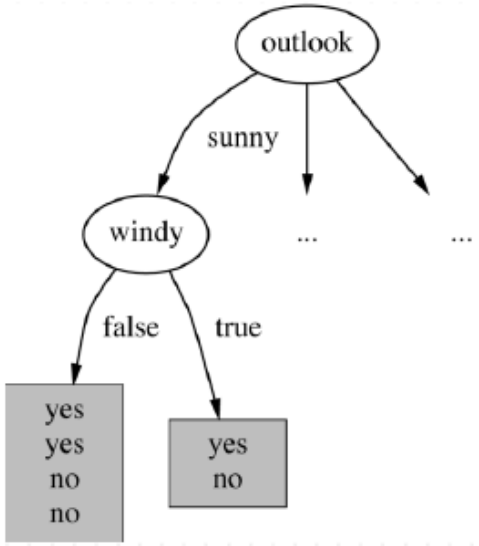
Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	High	True	***No***

Feature Selection: Information Gain

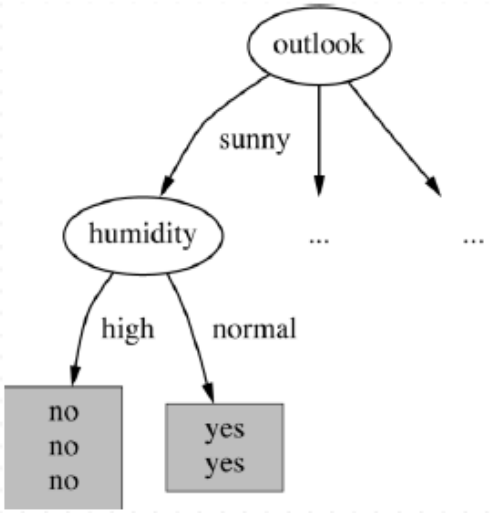
- Quão bem uma feature divide as classes



Temperature = 0.571



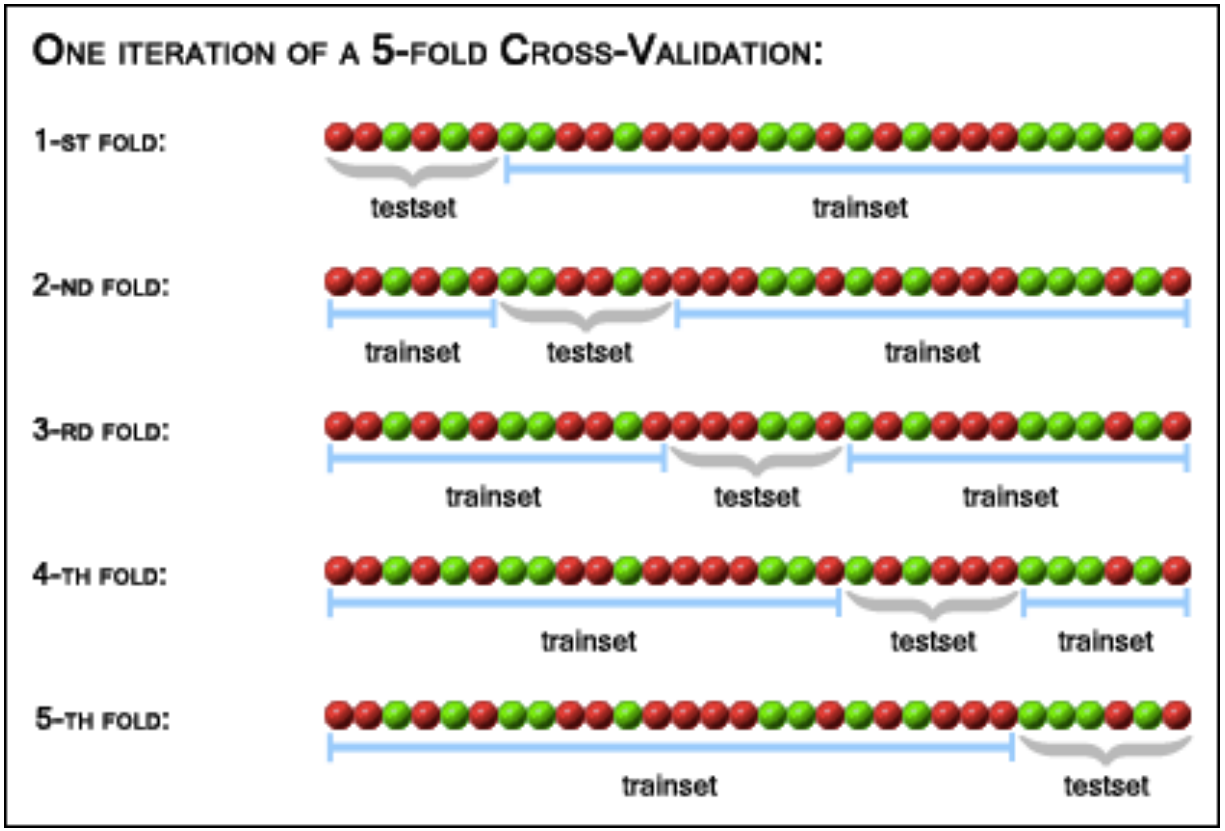
Windy = 0.020



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)

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

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

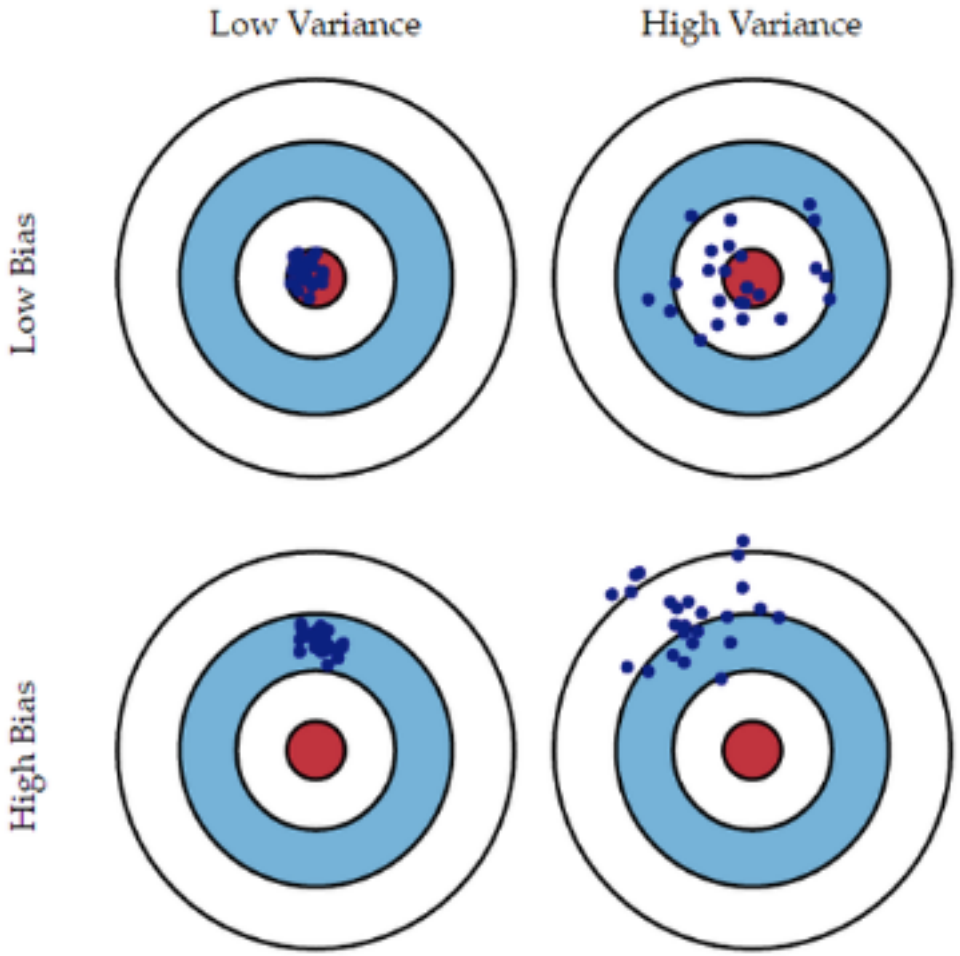
$$F_1 = \frac{1}{\frac{1}{2}\frac{1}{P} + \frac{1}{2}\frac{1}{R}} = \frac{2PR}{P+R}$$

Accuracy = (TP+TN)/(TP+TN + FN + FP)

Diagnóstico de Modelos

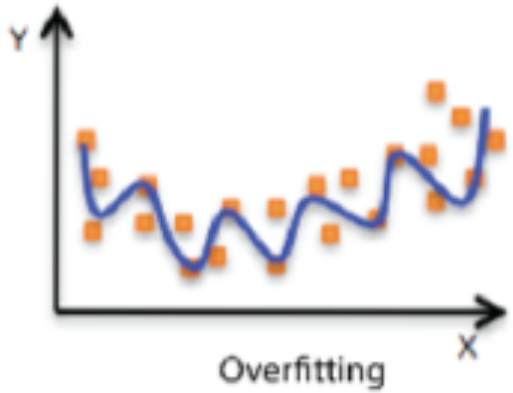
- Como melhorar o classificador?
  - Features, dados etc
- Exemplo:
  - Alvo: 5% de erro
  - Erro no treinamento: 15% (**viés: 15-5= 10**)
  - Erro no teste: 16% (**variância: 16-15 = 1**)
  - Precisa melhor o desempenho no conjunto de treinamento
- Viés:
  - Desempenho no conjunto de treinamento
  -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

Viés e Variância



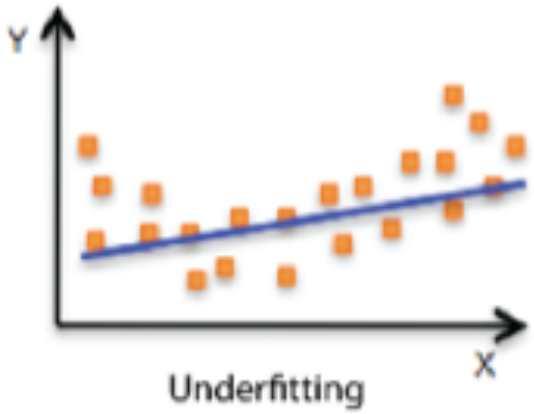
Overfitting

- 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%

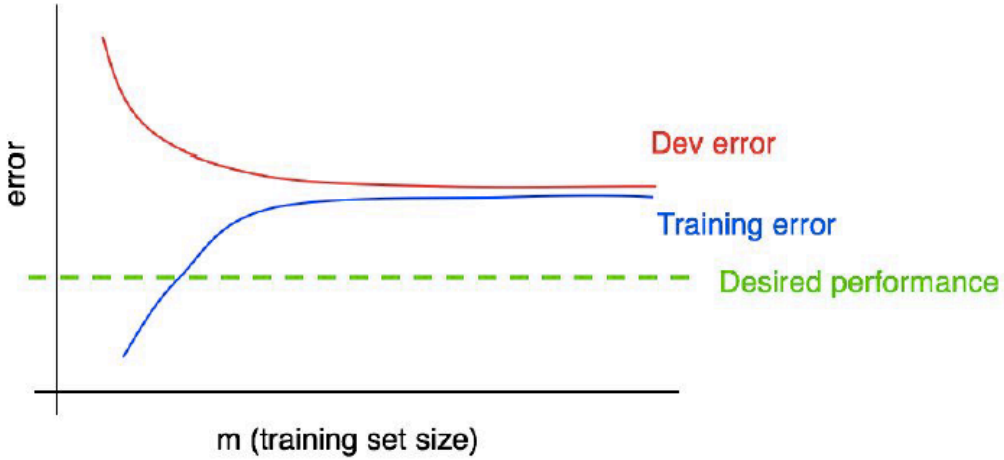


Outros Cenários

- 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/