



Modelos Preditivos Séries Temporais

Capítulo 1. Introdução à Modelagem Preditiva

Prof. Fernando Mourão e Prof. Túlio Vieira



Aula 1.1. Introdução à modelagem preditiva

- Definir formalmente a modelagem preditiva.
- Apresentar os principais termos relacionados à área.
- Discutir os cinco passos principais do processo de criação de modelos preditivos.
- Discutir os principais desafios relacionados à modelagem preditiva em cenários reais.

O que é modelagem preditiva?

- Uso de dados, algoritmos e métodos oriundos da Estatística, Aprendizado de Máquinas e Mineração de Dados para se determinar as chances de resultados futuros, ou desconhecidos, com base em dados passados. O objetivo é ir além de saber o que aconteceu ao fornecer uma melhor estimativa do que acontecerá no futuro.

■ Por que modelagem preditiva?

- A análise criteriosa de dados tanto internos como externos a uma organização torna-se cada vez mais necessária, dada a escassez de tempo e a cobrança por agilidade e flexibilidade imposta pelo mercado na **tomada de decisão**.

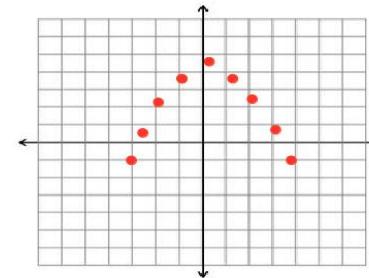
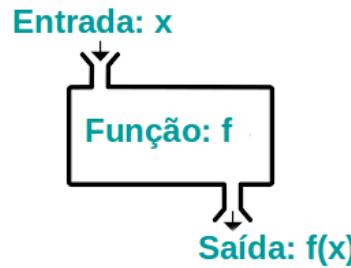




Conhecimento: Processo de uso da informação para a tomada de decisão.

Terminologia

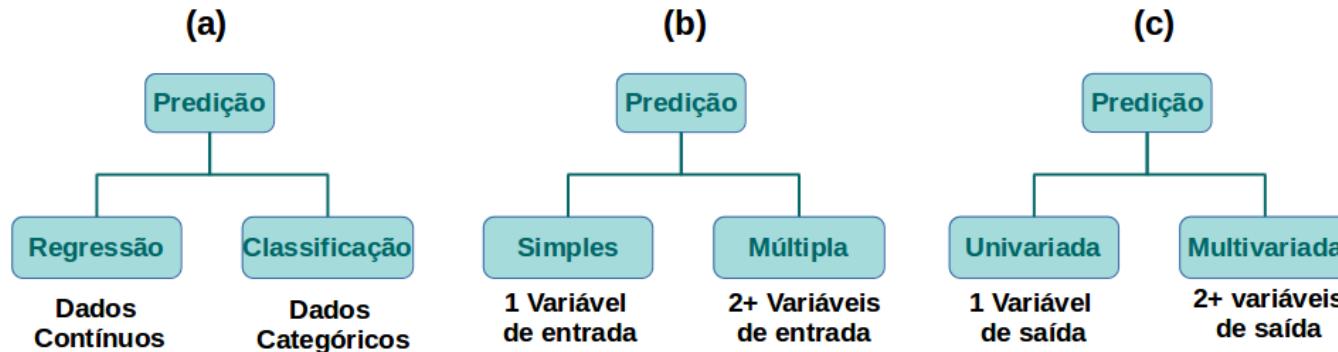
- Considerando-se o escopo da Modelagem Preditiva, sem perda de generalidade, podemos considerar um modelo simplesmente como uma função matemática.



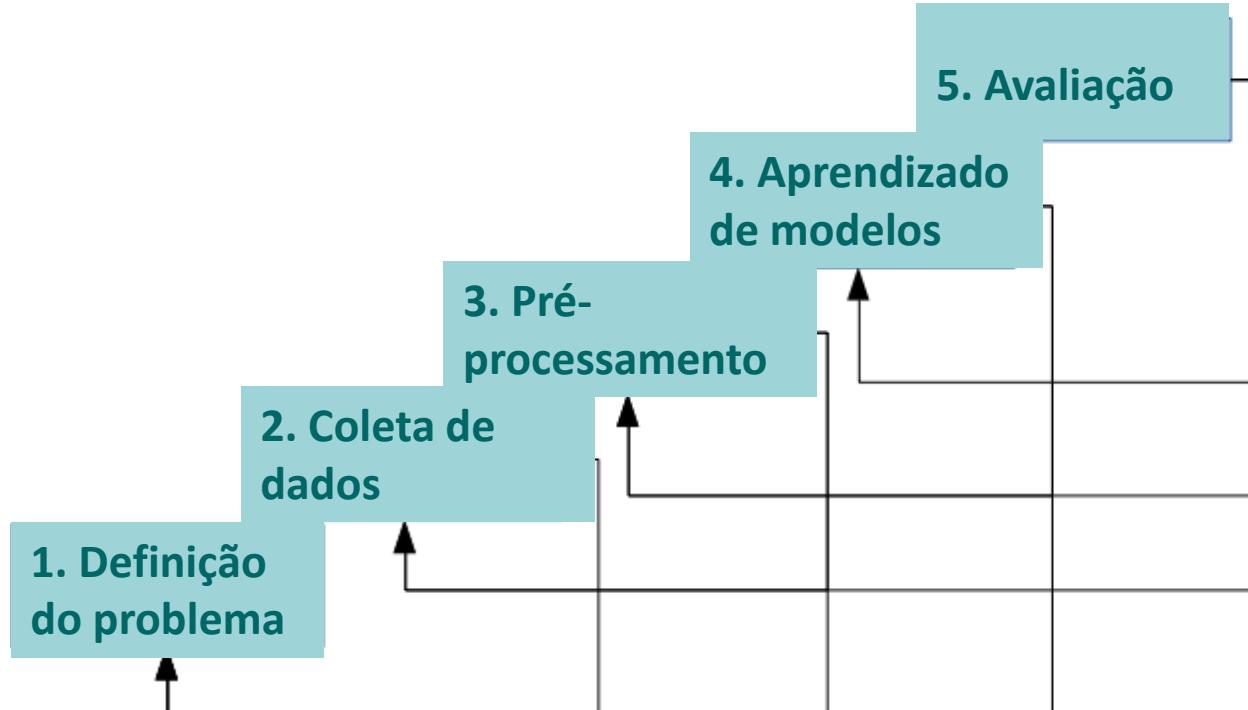
- O processo de aprendizado consiste em se definir os valores dos parâmetros para a função que melhor se ajusta aos dados de entrada.

Terminologia

- Classificação dos modelos de predição de acordo com (a) tipos de dados; (b) número de variáveis independentes; (c) número de variáveis dependentes.



Como criar modelos preditivos?



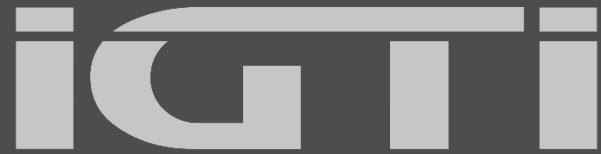
Desafios relacionados

- Identificar as premissas corretas para cada problema.
- Limitação dos dados.
- Qualidade dos dados.
- Complexidade de modelo.
- Volume de dados & desempenho.

Conclusão

- Definição formal da modelagem preditiva.
- Os principais termos relacionados à área.
- Cinco passos principais do processo de criação de modelos preditivos.
- Os principais desafios relacionados à modelagem preditiva em cenários reais.

- Amostragem em modelagem preditiva.
- Conhecer os principais termos e conceitos estatísticos relacionados à amostragem.
- Elaboração de planejamentos amostrais.
- Técnicas de amostragem.
- Qualidade dos dados.



Modelos Preditivos Séries Temporais

Capítulo 2. Coleta de Dados

Prof. Fernando Mourão e Prof. Túlio Vieira



Aula 2.1. Técnicas de amostragem

- Entender o papel da amostragem em modelagem preditiva.
- Conhecer os principais termos e conceitos estatísticos relacionados à amostragem.
- Introduzir planejamentos amostrais.
- Diferenciar as principais técnicas de amostragem existentes.

População

Predição

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Amostra

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$

Amostragem

- **Quando amostrar?**

- População infinita ou arbitrariamente grande.
- Estudo de comportamentos dinâmicos.
- Restrições de tempo e custo computacional.
- Predições sobre dados desconhecidos.
- Identificação de comportamentos verdadeiramente significativos.

- **Quando não amostrar?**

- População é pequena e o custo de se amostrar for similar ao de se utilizar toda a população.
- Tamanho da amostra necessária para conduzir o estudo for muito grande.
- Necessidade de precisão completa sobre os resultados.
- População muito heterogênea e o erro introduzido pelo uso da amostra se tornar grande.

1. **População-alvo:** subconjunto de objetos pertinentes para o estudo;
2. **Unidade amostral:** indivíduos ou grupos de indivíduos;
3. **Tamanho da população:** número total de unidades amostrais na população-alvo;
4. **Tamanho da amostra:** número total de unidades amostrais na amostra;
5. **Técnica de amostragem:** método de seleção das amostras.

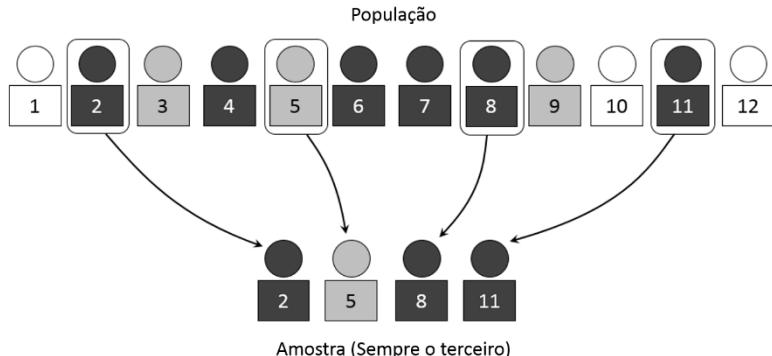
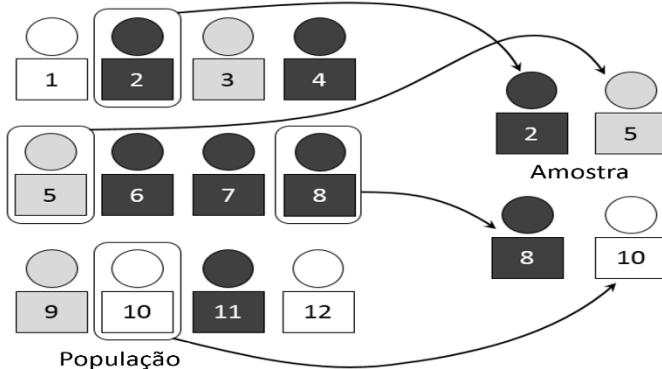
Amostragem probabilística

- **Aleatória simples**

- Elementos da população são sorteados seguindo uma distribuição uniforme.

- **Aleatória sistemática**

- Amostra obtida a partir da progressão aritmética sobre ponto de partida selecionado aleatoriamente.



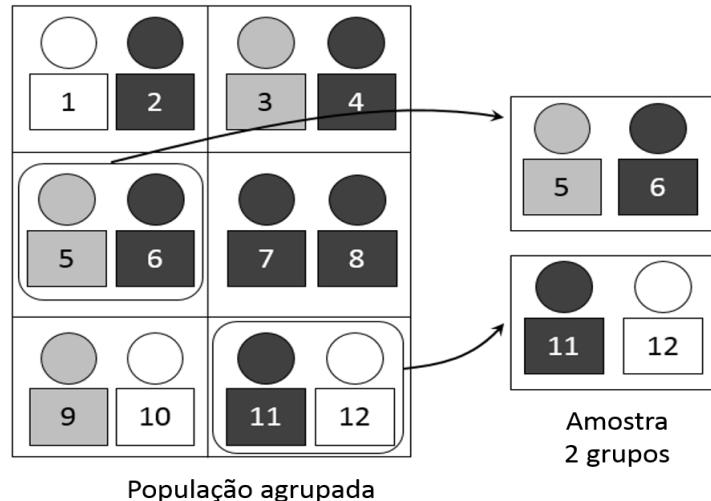
Amostragem probabilística

- **Aleatoriedade estratificada**

- População é dividida em estratos de acordo com alguma característica relevante e selecionam-se indivíduos.

- **Por conglomerado**

- População é organizada em grupos de acordo com alguma característica relevante e selecionam-se grupos.



- **Por conveniência**

- Elementos são incluídos na amostra sem a probabilidade previamente especificadas ou conhecidas de eles serem selecionados.

- **Intencional ou por julgamento**

- O pesquisador avalia quais pessoas detêm maior conhecimento do tema a ser estudado e escolhe os elementos que julga serem os mais representativos da população.

- **Snowball**

- A cada um que se enquadra, o entrevistador pede que este lhe indique onde é possível encontrar outro para entrevistar, até chegar ao número de entrevistas desejadas.

- **Por quotas**

- Para cada entrevistador é atribuída uma cota de entrevistas e este escolherá pessoas que estejam dentro do perfil da pesquisa.

Conclusão

- O papel da amostragem em modelagem preditiva.
- Os principais termos e conceitos estatísticos relacionados à amostragem.
- Planejamentos amostrais.
- As principais técnicas de amostragem existentes.

■ Próxima aula

- Qualidade de dados.
- Erro amostral vs. não amostral.
- Principais erros em amostragem.



Aula 2.2. Qualidade da amostra

- ❑ Discutir sobre as características que definem qualidade de amostras.
- ❑ Diferenciar erro amostral e não amostral.
- ❑ Revisar os principais tipos de erros durante o processo de amostragem.

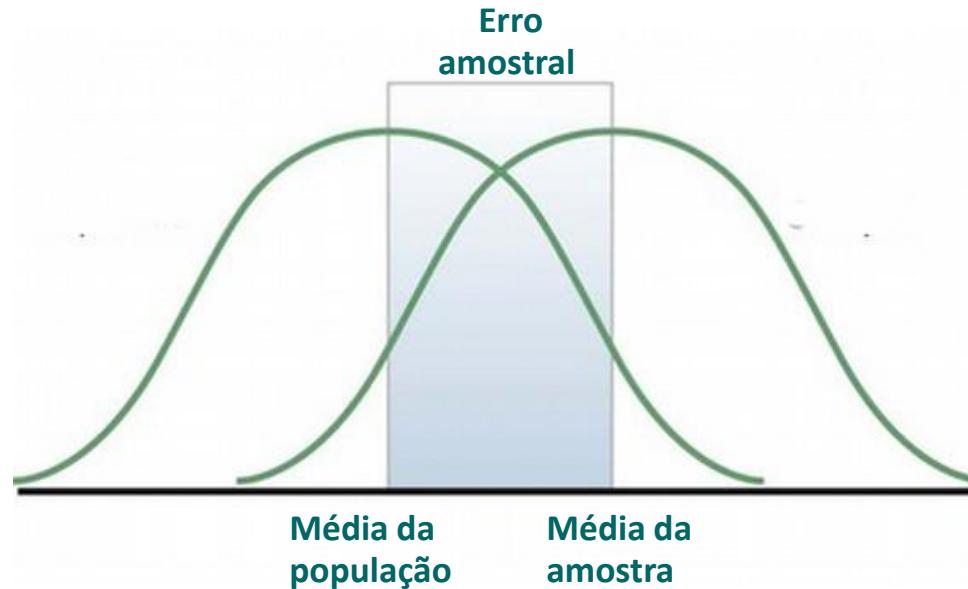
Qualidade da amostra

Quando falamos em qualidade da amostra consideramos especificamente três características fundamentais:

- 1. Consistência:** Precisamos ter confiança de que qualquer comportamento ou mudança observada nos dados reflete um comportamento ou mudança real na população.
- 2. Diversidade:** Para ser verdadeiramente representativa, uma amostra deve ser tão diversificada como a própria população.
- 3. Transparência:** É fundamental que pesquisadores discutam as limitações de seus dados e mantenham a transparência sobre os procedimentos seguidos ao selecionar a amostra.

Erros amostrais

- Erro amostral é a diferença entre a estimativa obtida para um parâmetro e o seu verdadeiro valor.



- Decorrem da variabilidade natural das unidades amostrais; é plausível esperar que amostras tenham comportamentos similares aos da população, mas não idênticos.
- Dependem de diversos fatores: tamanho da amostra, variabilidade das características de interesse na população, projeto amostral e método de seleção das amostras.
- Podem ser medidos a partir da própria amostra: medidos através do Coeficiente de Variação (CV).
- Regra de ouro:

$CV < 10\%$ → amostras boas

$10\% < CV < 20\%$ → amostras aceitáveis

$20\% < CV$ → amostras a serem descartadas

Erros não amostrais

- São classificados em quatro tipos principais:
 - Erros de valores ausentes;
 - Erros de cobertura;
 - Erros de medição;
 - Erros de processamento.

Erros de valores ausentes

- Consistem na incapacidade de se medir todas as variáveis de interesse em todas as unidades amostrais.
- Produzem erros nas estimativas da pesquisa de duas maneiras:
 1. **Bias amostral:** os valores ausentes muitas vezes possuem características que os diferem dos valores presentes nas amostras.
 2. **Tamanho efetivo da amostra:** ter um número maior de valores ausentes reduz o tamanho efetivo da amostra. Como resultado, a precisão das estimativas diminui.

Erros de cobertura

- Ocorrem quando há diferenças entre a população-alvo e a população amostrada
 - A cobertura aumentada, geralmente, ocorre quando a população amostrada é maior que a população-alvo, não configurando um problema.
 - A cobertura reduzida, em que a população amostrada torna-se menor que a alvo, é problemática.



Erros de medição

- Ocorrem quando uma resposta fornecida difere do valor real.
- Existem várias fontes de erro de medição:
 - Viés social sobre o comportamento;
 - Efeito entrevistador;
 - Erro no questionário;
 - Imprecisão.

Erros de processamento

- Podem ocorrer erros em várias etapas do processamento e armazenamento das amostras, incluindo captura, codificação e transformação dos dados.
- Neste caso, é necessário definir um conjunto de procedimentos completos e sistematizados para avaliar a qualidade de cada variável de interesse e corrigir os erros identificados.

- As características que definem qualidade de amostras.
- Erro amostral vs. não amostral.
- Os principais tipos de erros ao longo do processo de amostragem.

- ❑ Importância do pré-processamento de dados.
- ❑ Tipos de dados.
- ❑ Mensuração de qualidade de dados.
- ❑ Tratamento de dados numéricos & categóricos.



Modelos Preditivos Séries Temporais

Capítulo 3. Tratamento de Dados

Prof. Fernando Mourão e Prof. Túlio Vieira



Aula 3.1.1. Dados numéricos e categóricos (Parte 1)

- Discutir a importância do tratamento de dados.
- Diferenciar os principais tipos de dados.
- Iniciar a discussão sobre os principais passos de tratamento aplicados a dados numéricos e categóricos.

Por que pré-processar os dados?

- Consistência e confiabilidade são fundamentais para que possamos usar qualquer tipo de dados em modelagem preditiva.
- Porém, dados do mundo real são:
 - **Incompletos**: falta de valores de atributo, falta de certos atributos de interesse ou contêm apenas dados agregados.
 - **Ruidosos**: com erros ou outliers.
 - **Inconsistentes**: contendo discrepâncias entre valores esperados e observados.
 - **Duplicação**: mesmos registros podem ser coletados mais de uma vez.

Por que pré-processar os dados?

- **Dados incompletos decorrem de:**

- Valores de dados não aplicáveis a algumas observações durante a coleta.
- Diferentes considerações e metodologias entre o tempo de coleta e o tempo de análise.

- **Dados ruidosos decorrem de:**

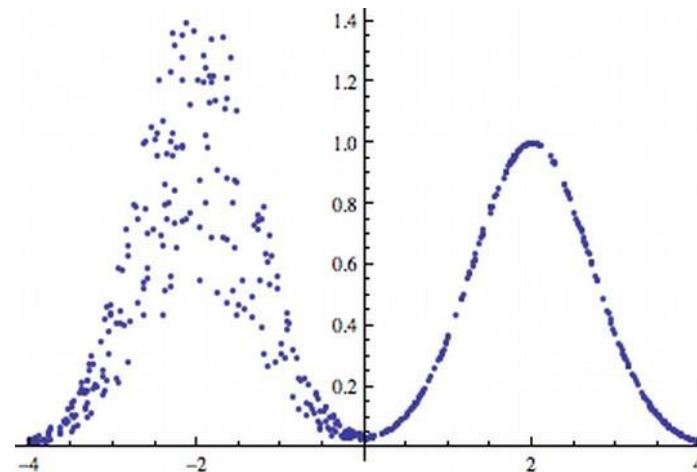
- Falha nos instrumentos de coleta.
- Erros humanos ou erros na entrada de dados.
- Erros na transmissão de dados.

■ Por que pré-processar os dados?

- **Dados inconsistentes decorrem de:**
 - Diferentes fontes de dados.
 - Violação de dependências funcionais.
- **Dados duplicados decorrem de:**
 - Metodologia de coleta.
 - Erros de verificação.

Por que pré-processar os dados?

- A ocorrência excessiva de ruídos pode ocultar o real comportamento de uma população.

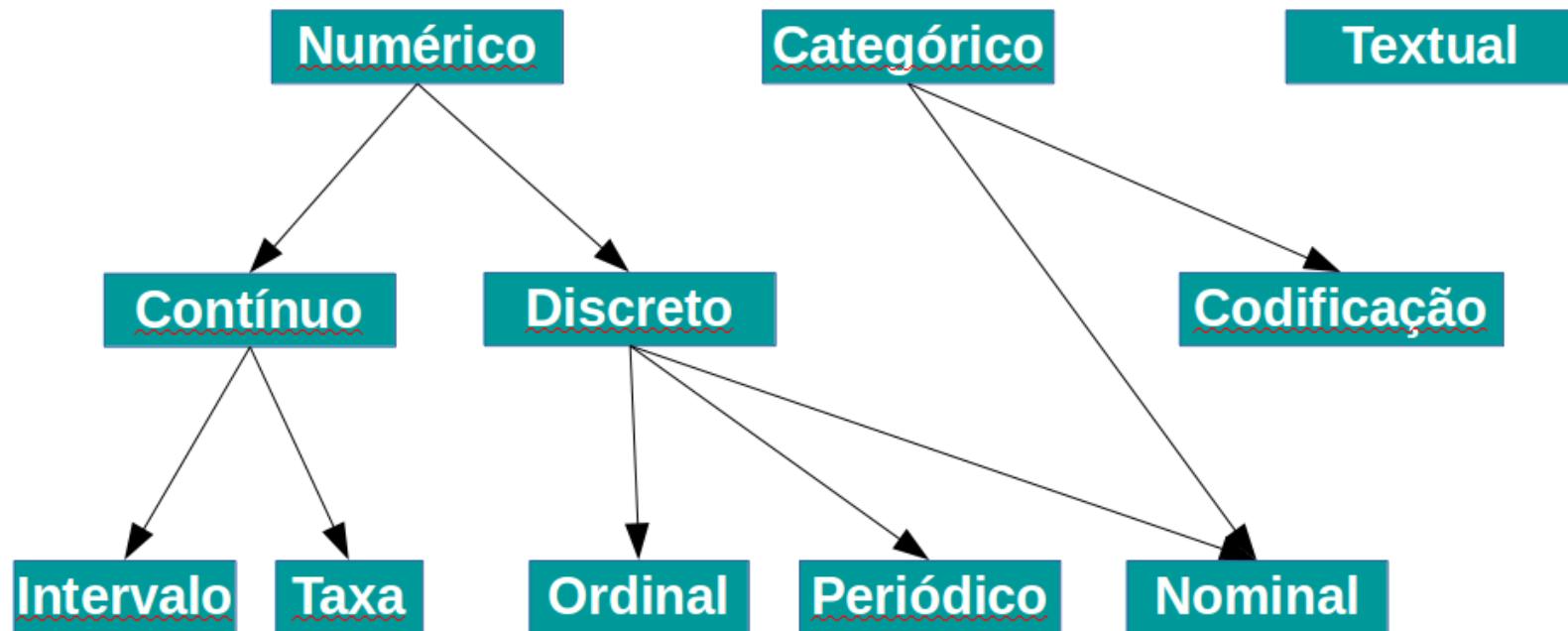


- Decisões qualificadas se baseiam em dados de qualidade!

Mensuração de qualidade dos dados

- Pode-se considerar diversas dimensões de análise de qualidade:
 - Acurácia;
 - Completeza;
 - Consistência;
 - Confiabilidade;
 - Interpretabilidade;
 - Valor agregado.
- Tratar os dados de forma a se garantir altos níveis de qualidade é, muitas vezes, a etapa mais trabalhosa e demorada do processo de modelagem preditiva.

Tipos de dados



- **Nominal:** CEP, cor dos olhos, número de identidade.
- **Ordinal:** Notas de avaliação, altura de uma pessoa, peso corporal.
- **Intervalo:** Datas, temperatura em Celsius.
- **Razão:** Idade, quantidades monetárias, massa.
- **Periódico:** Semana, meses, horas.
- **Codificação:** Dummy coding ou Effects coding.
- **Textual:** Posts, avaliações textuais de apps, páginas web.

Modelos de dados

- **Registros – conjunto fixo de atributos:**

- Matriz de dados.
- Documentos.
- Transações.

- **Grafos – representação de relações:**

- Web.
- Estruturas moleculares.

- **Ordenados – ordem de ocorrência importa:**

- Dados espaciais.
- Dados temporais.
- DNA.

- **Principais etapas:**

- Limpeza dos dados.
- Integração.
- Transformação.
- Redução.
- Discretização.

- Consiste no processo de se preencher valores ausentes, atenuar distorções em dados ruidosos, identificar e remover outliers e duplicações, bem como resolver inconsistências observadas nos dados.
- Dados numéricos, categóricos e textuais requerem estratégias distintas de se realizar tal limpeza.
- Grande parte do esforço depende do domínio de análise e objetivos de estudo.

- Observações com valores ausentes em uma ou mais variáveis.
- Estratégias de solução:
 1. Ignorar observações.
 2. Preencher manualmente.
 3. Usar uma constante global ou valores médios.
 4. Preencher automaticamente:
 - Inferência bayesiana.
 - EM.
 - Árvores de Decisão.

- Consiste em erros ou variações randômicas nos valores das variáveis.
- Estratégias de solução:
 - Análise de variabilidade das variáveis.
 - Detecção manual de valores suspeitos.
 - Métodos de binificação e atenuação de valores.
- Binificação:
 - Ordena-se os dados e, em seguida, os particiona em bins.
 - Aplicam-se alguma função de agregação sobre os valores de cada bin gerando um único valor por bin.

- Observações inconsistentes com o restante dos dados.
- Exemplos:
 - Salários de CEO
 - Idade = 350
 - Altura física: 287 cm
- Métodos de identificação de outliers:
 - Algoritmos de agrupamento.
 - Fitting de curvas.
 - Teste de hipótese com algum modelo.

- A importância do tratamento de dados.
- Os principais tipos de dados.
- Os principais passos de tratamento aplicados a dados numéricos e categóricos.

- ❑ Integração dos dados.
- ❑ Transformação sobre dados numéricos e categóricos.
- ❑ Redução de dados.
- ❑ O problema do “mal da dimensionalidade”.
- ❑ Técnicas de discretização.



Aula 3.1.2. Dados numéricos e categóricos (Parte 2)

- ❑ Discutir sobre as etapas de integração, transformação e discretização do dados.
- ❑ Apresentar o problema do “mal da dimensionalidade”.
- ❑ Revisar os principais conceitos por trás da tarefa de redução de dimensionalidade.

Integração de dados

- Objetiva combinar dados de múltiplas fontes de dados e metadados em uma única fonte coerente.
- Aborda o problema de identificação de entidade: identifica entidades do mundo real de múltiplas fontes de dados (e.g. Lula vs. Luís Inácio).
- Envolve a detecção e resolução de conflitos de valores: para uma mesma entidade do mundo real, valores oriundos de diferentes fontes podem divergir.
- Uma integração cuidadosa dos dados de várias fontes pode ajudar a reduzir/evitar redundâncias e inconsistências, bem como melhorar a qualidade dos dados.

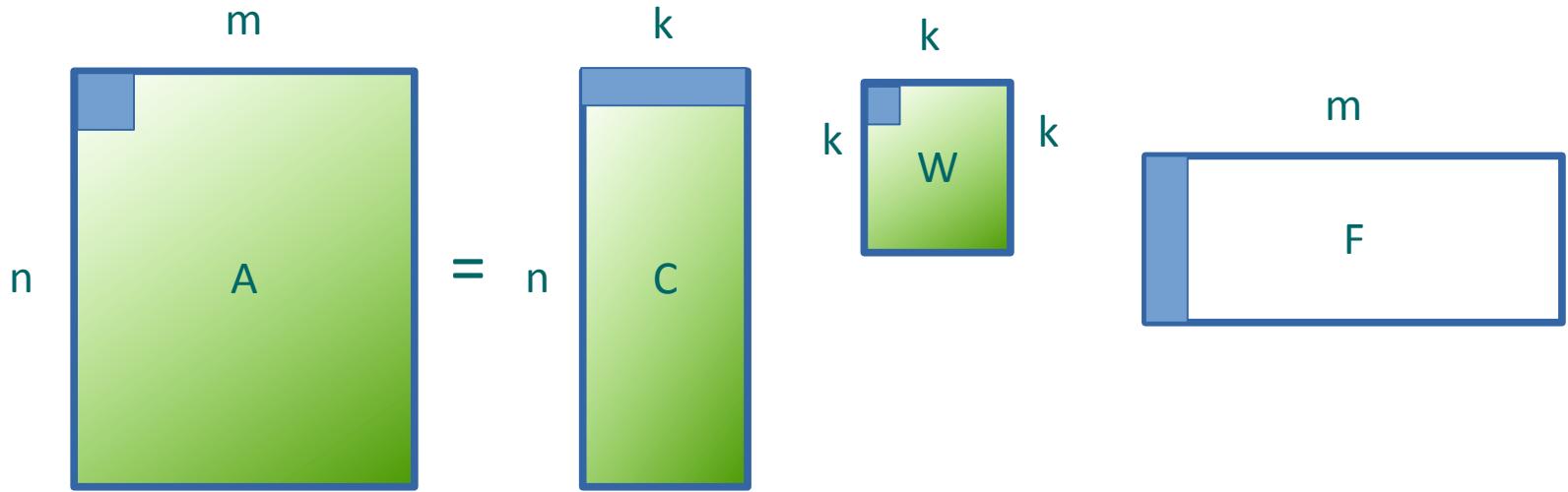
- Consiste na aplicação de uma ou mais das seguintes tarefas:
 - **Agregação ou sumarização de valores:** Aplicação de alguma função de agregação de distintos valores em um único. Exemplo: uso de média, mediana.
 - **Generalização:** Uso de ontologia ou hierarquia de conceitos para substituir conceitos específicos por outros mais genéricos. Exemplo: ontologia de cargos para representar empregos.
 - **Normalização:** Alteração dos valores dos atributos para que caiam sobre um determinado intervalo de interesse. Exemplo: uso do z-score, normalização min-max.
 - **Construção de features:** Criação de novos atributos/variáveis a partir dos dados. Exemplo: geração de tuplas a partir dos dados.

- Consiste em transformar valores numéricos contínuos em valores discretos.
- Discretização não supervisionada:
 - Intervalos iguais: divido os números em intervalos de mesmo tamanho.
 - Intervalos uniformes: use intervalos contendo o mesmo número de valores.
- Discretização supervisionada:
 - Limites de classe.
 - Entropia.

Redução de dimensionalidade

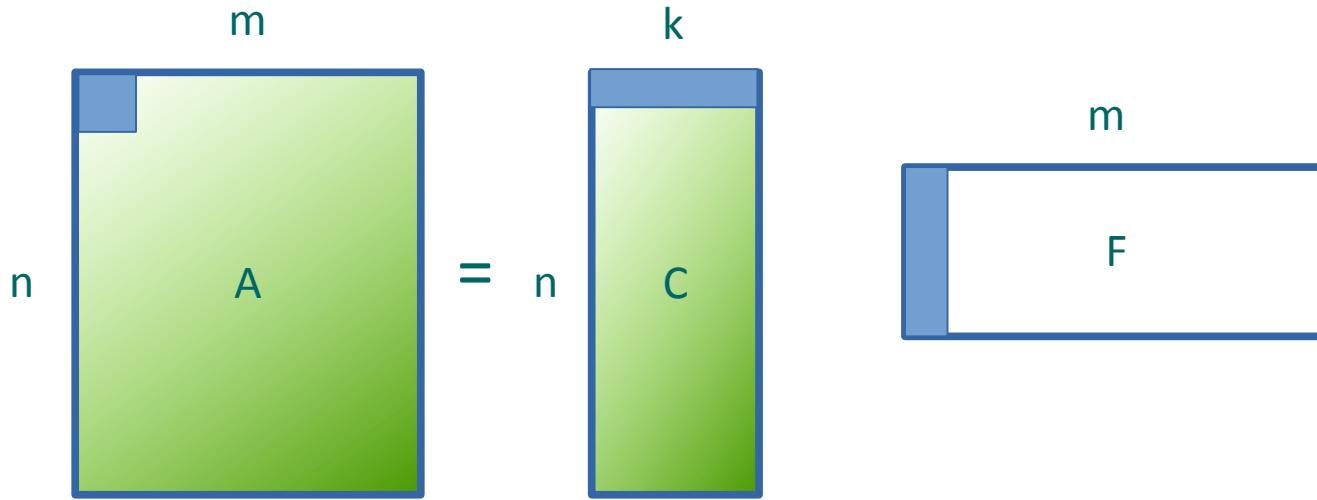
- Técnicas de redução da dimensionalidade se tornaram fundamentais para diversas aplicações.
- Reduzir consiste em representar de uma forma mais compacta sem perda de informações importantes.
- Podemos dividir técnicas de redução em três grandes grupos:
 - Seleção de características (Feature selection): Information Gain.
 - Redução de características (Feature Reduction): Fatoração de Matrizes.
 - Compressão de dados: Amostragem.

Fatoração de matrizes



- Cada entrada de A é um tipo de combinação ponderada de pedaços de informação de C e F .
- Esses pedaços de informação são denominados fatores latentes.
- O papel de k é forçar uma representação mais compacta. Em geral, $k \ll m$.

Fatoração de matrizes



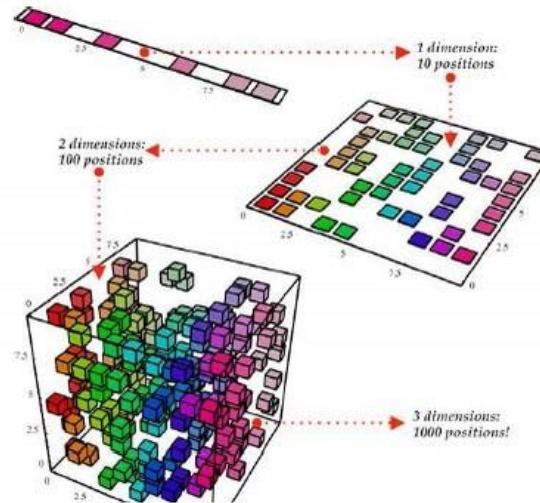
- A matriz W é sempre diagonal.
- Algumas técnicas de decomposição não geram a matriz W .
- Essas técnicas diferem-se sobre as premissas feitas sobre os fatores latentes.

Fatoração de matrizes

- Fatoração de matrizes possui duas funções importantes na análise de dados:
 - **Limpeza dos dados:** Esse conjunto de técnicas nos permite separar automaticamente os diferentes processos subjacentes aos dados.
 - **Reducir redundâncias:** Representa uma maneira elegante de identificar similaridades e redundâncias entre objetos ou atributos em coleções de dados.

Mal da dimensionalidade

- O mal da dimensionalidade refere-se a um problema resultante do grande número de variáveis (i.e., atributos) envolvidas.



- Esse problema resulta do fato de um número fixo de pontos se tornar crescentemente “esparso” a medida que o número de dimensões aumenta.

- Muitas técnicas da modelagem preditiva dependem criticamente de medidas baseadas em distância ou similaridade de objetos no espaço.
- Estudos demonstraram que tais medidas são significativamente afetadas em espaços com altas dimensões:

$$\lim_{d \rightarrow \infty} \frac{\text{MaxDist} - \text{MinDist}}{\text{MinDist}} \rightarrow 0$$

- Em espaços com alta dimensão, distâncias entre pontos se tornam relativamente uniformes!

- As etapas de integração, transformação e discretização do dados.
- O problema do “mal da dimensionalidade”.
- Principais conceitos por trás da tarefa de redução de dimensionalidade.

- Tratamento de dados textuais.
- Codificação de dados.
- Limpeza de dados textuais.



Aula 3.2.1. Dados textuais (Parte 1)

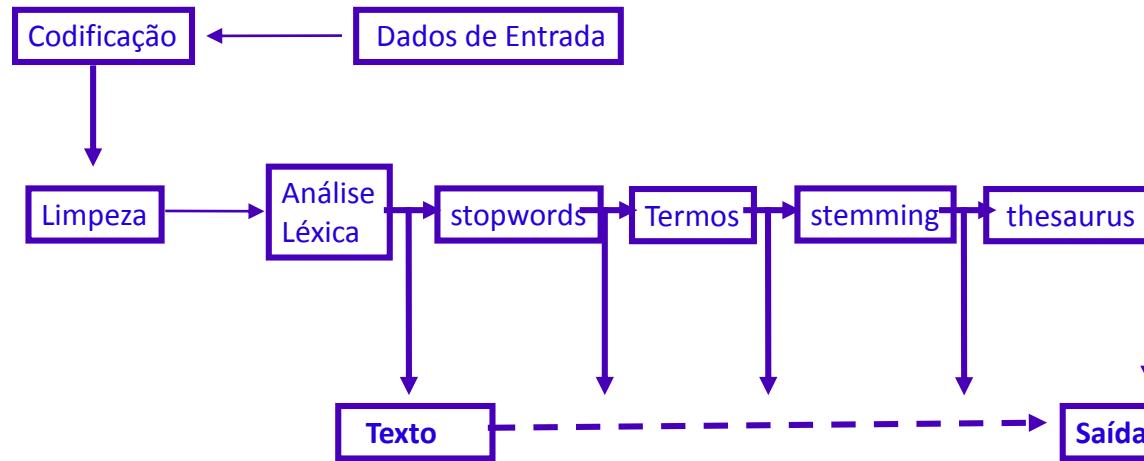
- ❑ Apresentar os principais passos de pré-processamento de dados textuais.
- ❑ Discutir sobre codificação de dados.
- ❑ Revisar o processo de limpeza de dados textuais.

- Nem sempre documentos estão representados de maneira adequada e precisamos transformá-los a fim de melhorar a precisão na posterior manipulação dos dados.
- Existem vários passos neste tratamento:
 - Identificação da codificação.
 - Limpeza do texto.
 - Análise léxica/transformações.
 - Eliminação de Stopwords.
 - Stemming.
 - Indexação.
 - Aplicação de Thesaurus.

- **Identificação da codificação:** Identificar corretamente a codificação em que o texto se encontra.
- **Limpeza dos dados:** Análise de dados incompletos, inconsistentes ou mesmo errados.
- **Análise léxica:** Reconhecimento de dígitos, hífens, pontuações, caixa alta. Transformações são aplicadas, em geral, a dados numéricos.

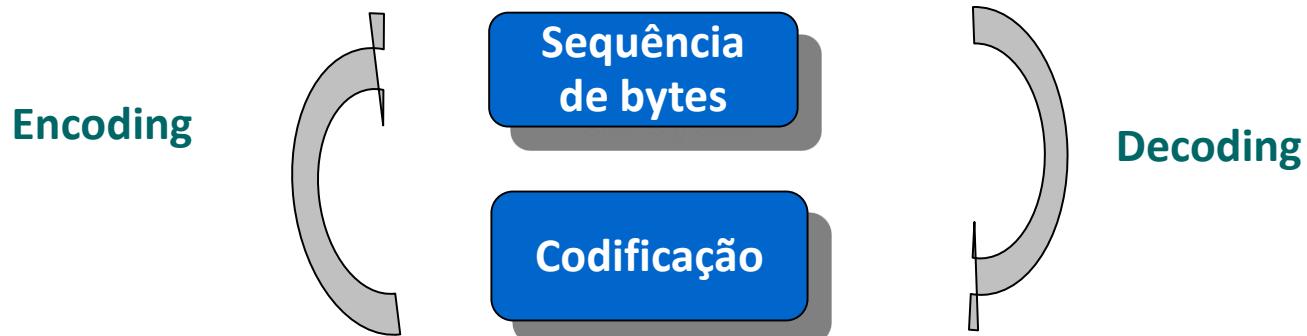
- **Stemming:** Tratar a variação sintática de termos.
- **Indexação:** Definir índices para posterior recuperação dos termos.
- **Aplicação de Thesaurus:** Expansão dos termos originais em termos semanticamente correlatos.

Tratamento de dados



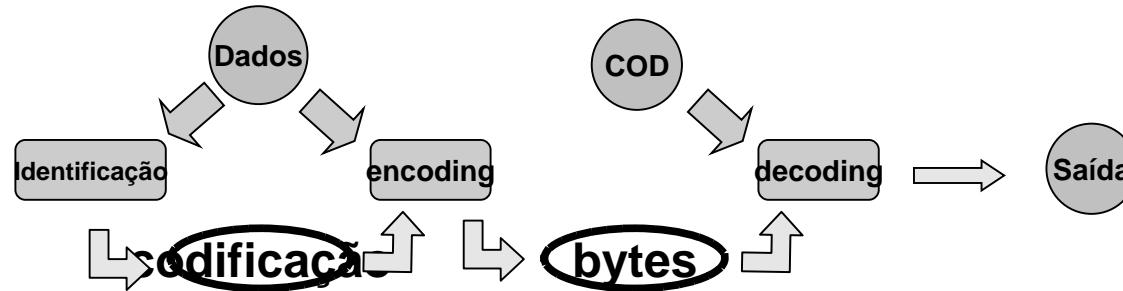
Codificação de dados

- Ao falar sobre codificação de dados, temos dois processos:
 - **Encoding**: Transforma um conjunto de caracteres de uma codificação específica para uma sequência de bytes.
 - **Decoding**: Transforma uma sequência de bytes em um conjunto de codificação qualquer.



Codificação de dados

- Como nem sempre sabemos a codificação que uma página adota, precisamos:
 - Tentar identificar esta codificação.
 - Realizar o processo de **encoding** sobre os dados, transformando os dados textuais em sequencias de bytes.
 - Realizar o processo de **decoding** sobre os dados, transformando as sequência de bytes na codificação textual desejada.

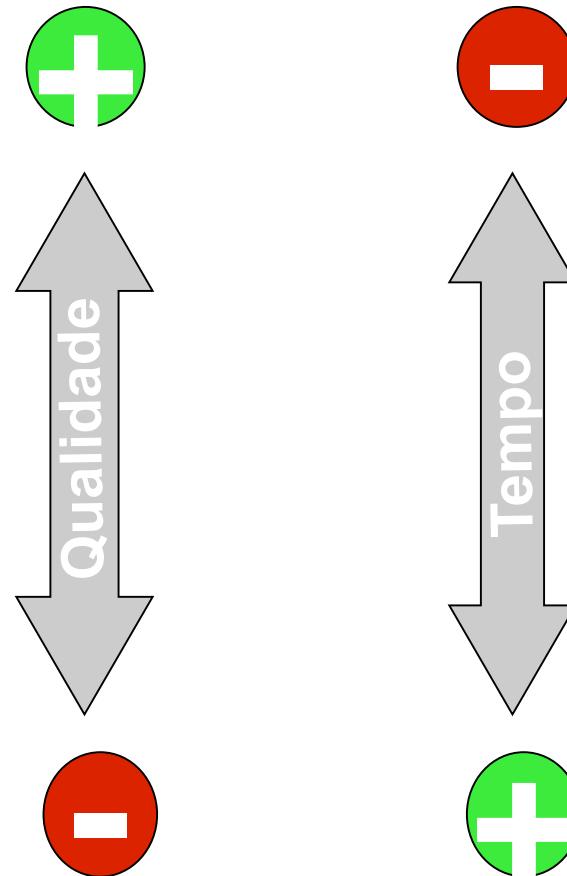


- Decisões de qualidade devem estar baseadas em dados de qualidade.
- Preenchimento de valores inexistentes, atenuação de dados ruidosos, identificação e remoção de desvios, resolução de inconsistências, etc.
- Pode corresponder a grande parte do esforço em análise dos dados.
- Há duas etapas distintas: identificação e correção.

- **Valor em falta:** falta de preenchimento de atributos obrigatórios.
- **Erro ortográfico:** anomalia encontra-se associada a atributos textuais.
Exemplo: **cidade = 'Brga'**
- **Utilização de sinônimos:** informações sintaticamente distintas, mas semanticamente iguais. **Exemplo profiss = 'professor' profiss = 'docente'**
- **Inexistência de representação padrão:** o valor do atributo aparece sob variados formatos. Exemplo **data = 04/05/2004, data = 4 de Julho de 2004**

Limpeza de dados

- Detecção Manual
- Detecção Automática
- Correção Manual
- Correção Automática



- Algoritmos específicos para cada contexto vs. abordagens genéricas.
- Sistematizar a limpeza de dados, adaptando-se a diferentes domínios e a uma variedade de tipos de atributos.
- Ferramentas comerciais/acadêmicas disponíveis.
- Dada a complexidade e o custo, muitas vezes estratégias simples são adotadas.

Conclusão

- Os principais passos de pré-processamento de dados textuais.
- Codificação de dados.
- O processo de limpeza de dados textuais.

- Análise léxica e transformações sobre os dados.
- Remoção de Stopwords.
- Stemming.
- Aplicação de Thesaurus.



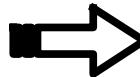
Aula 3.2.2. Dados textuais (Parte 2)

- Revisar análise léxica e transformações sobre os dados.
- Analisar o processo de remoção de Stopwords.
- Discutir o processo de Stemming.
- Introduzir o uso de Thesaurus.

- Consiste, basicamente, na remoção de caracteres especiais em texto.
- Tratamento de numerais
 - Usualmente não são bons para a análise, dada a variedade de valores que podem assumir.
 - Exemplos: 1000 D.C., 12/02/2000
- Tratamento de hifenação
 - Separar palavras constituídas por hífen pode ser útil em alguns casos.
 - Porém pode distorcer a semântica das palavras.
 - É necessário definir uma regra geral, bem como as exceções.
 - Exemplos: guarda-chuva, guarda chuva ou guardachuva?

- Tratamento de pontuações e caracteres especiais
 - Usualmente remove-se todos os caracteres de pontuação.
 - Exemplo: 1000 D.C → 1000 DC; Faça! → façā
 - . ; : ? / \ ! * & % \$ # () { } [] ^a ^o “ ”
 - Para documentos HTML remoção das tags.
- Tratamento de caracteres
 - Desabilita a caixa alta de todas as letras.
 - Remove caracteres não pertencentes ao vocabulário desejado.
 - Remove acentos.

```
<HTML>
<BODY>
...
<p> Na maioria das vezes, os
documentos retornados pelas
ferramentas de Recuperação de
Informação envolvem um
contexto mais amplo, fazendo
com que garimpar, ou seja, ...
</p>
</BODY>
</HTML>
```



na maioria das vezes os
documentos retornados pelas
ferramentas de recuperacao de
Informacao envolvem um
contexto mais amplo, fazendo
com que garimpar, ou seja, ...

Remoção de Stopwords

- Conceito básico:
 - Remover do texto palavras com baixo poder discriminativo.
 - Exemplo: Um, uma, de, hoje, a, o, embora, ser, não.
- Vantagem:
 - Reduz o tamanho da estrutura de indexação.
 - Foca a análise em termos semanticamente mais relevantes.
- Desvantagem:
 - Processo de definição das stopwords é usualmente manual e trabalhoso.
 - Reduz a qualidade da análise.
 - Exemplo: “Ser ou não ser, eis a questão”.

Remoção de Stopwords

último	dois	maioria	sem
é	dos	maiorias	ser
acerca	e	mais	seu
agora	ela	mas	somente
algmas	ele	mesmo	têm
alguns	eles	meu	tal
ali	em	muito	também
ambos	enquanto	muitos	tem
antes	então	nós	tempo
apontar	está	não	tenho
aquela	estão	nome	tentar
aquelas	estado	nosso	tentaram
aquele	estar estará	novo	tente
aqueles	este	o	tentei
aqui	estes	onde	teu
atrás	esteve	os	teve
bem	estive	ou	tipo
bom	estivemos	outro	tive
cada	estiveram	para	todos
caminho	eu	parte	trabalhar
cima	fará	pegar	trabalho
com	faz	pelo	tu

Remoção de Stopwords

... na maioria das vezes os documentos retornados pelas ferramentas de recuperação de informações evolvem um contexto mais amplo fazendo com que o usuário tenha que garimpar ou seja especificar ou filtrar estes documentos e que demanda tempo e conhecimento a fim de obter a informação que ele realmente necessita ...

Documento normalizado



... maioria vezes documentos retornados ferramentas recuperação informações evolvem contexto amplo fazendo usuário tenha garimpar especificar filtrar documentos demanda tempo conhecimento fim obter informação realmente necessita ...

Documento sem stopwords

- Stemming é o processo de extrair o prefixo de uma palavra, baseando-se em seu radical ou forma raiz. Exemplo:

Nomeação Presidente Estudante

- Stemming é útil, por exemplo, para máquinas de busca, facilitando:
 - Expansão de consultas. Exemplo:
 - “indexação de texto”, “index de texto”, “indexar de texto”...
 - Indexação
 - Ter, teríamos, tivesse, tiver, terá...

- Para realizar o stemming, usualmente criamos uma tabela que mapeia cada inflexão para um radical.
- Com isso, basta procurar cada palavra do texto na tabela e substituí-la pelo seu radical.
- A desvantagem desta estratégia consiste no tempo dispendido para construir essa tabela.
- Além disso, remover o sufixo de uma palavra pode acarretar em perda semântica. Exemplo:
 - conjugue vs. conjugado

- O que é um Tesauro?
 - Lista de palavras relevantes em um dado domínio.
 - Agrupa as palavras com mesmo significado.
 - É visto como um vocabulário controlado utilizado para as tarefas de indexação e busca.
 - Basicamente pode ser visto como um dicionário com algumas informações adicionais.
- Objetivos:
 - Prover um vocabulário padronizado.
 - Realizar uma “classificação” semântica dos termos.

Aplicação de Thesaurus

IGTI

Google  

All Images News Maps Videos More Settings Tools

About 21,400,000 results (0.50 seconds)

Página Inicial — Planalto
www2.planalto.gov.br/  Translate this page
Todas as informações sobre a Presidência da República do Brasil e as atividades diárias do presidente Michel Temer.

Legislação
Vice-Presidência · Agendas · Residência Oficial · Bandeira ...

Presidência
Acesso à Informação · Institucional · Presidência · Agendas de ...
[More results from planalto.gov.br »](#)

Agenda de Presidente Michel ...
Solenidade de Transmissão do Cargo de Diretor-Geral da ...

Agenda do Presidente
Antonio Imbassahy, ministro-chefe da Secretaria de Governo da ...

Top stories



 
See outside

Gabinete da Segunda Vice-Presidência do TJMG ★

Government office in Belo Horizonte, Brazil - 1.0 km

[Website](#) [Directions](#)

Address: Av. Afonso Pena, 4001 - Serra, Belo Horizonte - MG

Hours: Open today · 8AM–6PM ▾

Phone: (31) 3306-3100

[Suggest an edit](#) · [Own this business?](#)

Know this place? Answer quick questions

Conclusão

- Análise léxica e transformações sobre os dados.
- O processo de remoção de Stopwords.
- O processo de Stemming.
- O uso de Thesaurus.

- Definição do problema de regressão linear.
- Revisar o problema dos quadrados mínimos.



Modelos Preditivos Séries Temporais

Capítulo 4. Aprendizado de Modelos Preditivos

Prof. Fernando Mourão e Prof. Túlio Vieira



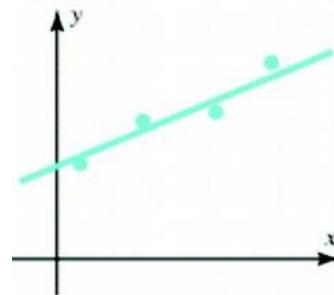
Aula 4.1.1. Regressão linear (Parte 1)

- ❑ Definir o problema de regressão linear.
- ❑ Apresentar a base teórica do problema.
- ❑ Discutir como solucionar este problema.
- ❑ Revisar o problema de quadrados mínimos.

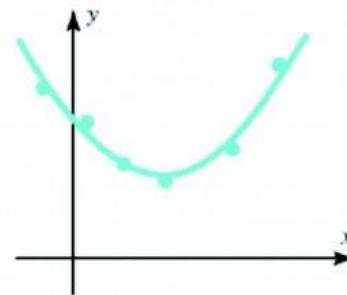
- Análise de regressão é um processo de estatística, que visa estimar a relação entre variáveis.
- A regressão nos ajuda a entender como o valor de uma variável dependente, muda quando uma das variáveis independentes se altera.
- O objetivo de predição é uma função de variáveis independentes, denominada de função de regressão.

Modelos de regressão linear

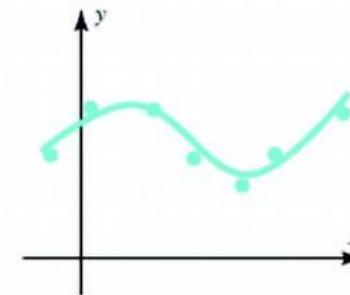
- Em regressão linear, dados são modelados usando-se funções de predição lineares.
- Parâmetros desconhecidos do modelo são estimados a partir dos dados.



(a) $y = a + bx$



(b) $y = a + bx + cx^2$



(c) $y = a + bx + cx^2 + dx^3$

- Um sistema linear é uma coleção de duas ou mais equações lineares, envolvendo o mesmo conjunto de variáveis.
- Uma equação linear é uma equação algébrica, na qual cada termo é uma constante, ou o produto de uma constante é uma variável simples (operadores lineares).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i$$

- O termo “linear” refere-se à ocorrência dos coeficientes de regressão β_j .

Modelos de regressão linear

- Relação entre as variáveis é uma função linear:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Diagram illustrating the components of a linear regression equation:

- Y-Intercept**: Points to the term β_0 .
- Slope**: Points to the term β_1 .
- Variável dependente (resposta)**: Points to the term Y_i .
- Variável independente**: Points to the term X_i .
- Erro randômico**: Points to the term ε_i .

Resolvendo regressão linear

- Uma vez que entendemos os objetivos relacionados ao problema de regressão, algumas perguntas simples surgem:
 - Como resolver este problema?
 - Como modelar os dados de entrada?
 - Existe uma solução ideal para o problema?

Resolvendo regressão linear

- Para se determinar a solução de um problema de regressão linear, precisamos responder duas questões:
 - Como determinar os coeficientes de regressão β_j ? (problema dos quadrados mínimos)
 - Como determinar o número de variáveis independentes a se considerar na função final? (seleção de modelos)

■ Problema dos quadrados mínimos

- Dado um sistema linear de m equações sobre n variáveis, encontre o vetor x que minimiza a equação abaixo com respeito ao produto euclidiana em R^m :

$$Y = AX + E$$

- Denominados o vetor x como a solução quadrados mínimos do sistema.
- Denominados E como o vetor de erros quadrados.
- A função acima é denominada linha de regressão.

■ Problema dos quadrados mínimos

- O termo “solução quadrados mínimos”, resulta do fato de que minimizar:

$$\|Y - Ax - E\|$$

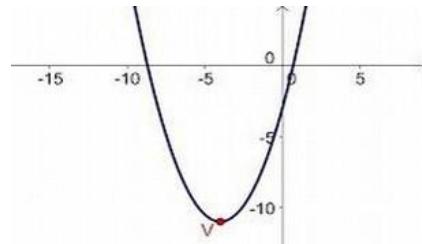
- Representa minimizar:

$$\|Y - Ax\|^2 = e_1^2 + e_2^2 + \dots + e_m^2$$

- A questão é: por que minimizar o quadrado?

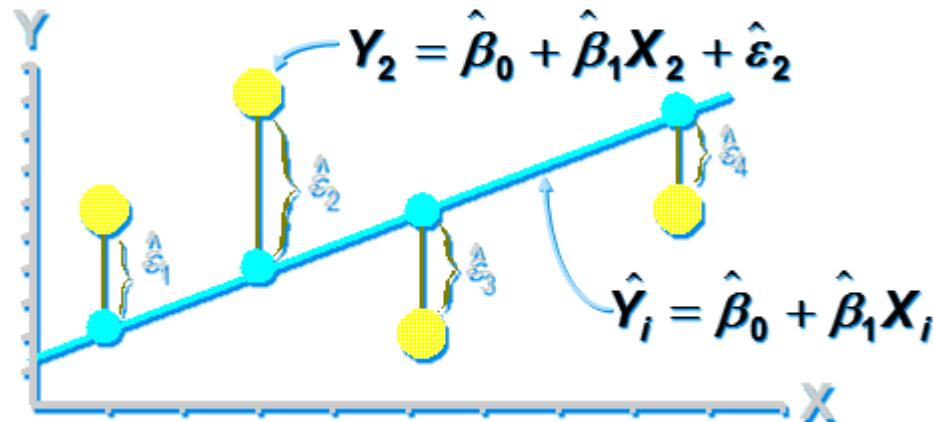
■ Problema dos quadrados mínimos

- Listamos quatro razões principais para se minimizar quadrados:
 1. Ao somar quadrados, evitamos de misturar valores positivos e negativos;
 2. Funções quadráticas enfatizam magnitudes maiores e, consequentemente, penalizam erros maiores;
 3. Forma quadrática é a mais simples função, que possui exatamente um único ponto de mínimo (ou de máximo);
 4. Achar o mínimo da função quadrático é extremamente simples e barato, basta derivar a função e igualar a zero.



■ Problema dos quadrados mínimos

LS minimizes $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



■ Problema dos quadrados mínimos

Solução Analítica

$$\begin{aligned}\sum_{i=1}^n \varepsilon_i^2 &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \\ 0 &= \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} \\ &= -2 \left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x} \right) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

$$\begin{aligned}0 &= \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} \\ &= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) \\ &= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) \\ \beta_1 \sum x_i (x_i - \bar{x}) &= \sum x_i (y_i - \bar{y}) \\ \beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}}\end{aligned}$$

Conclusão

- Definir o problema de regressão linear.
- Apresentar a base teórica do problema.
- Discutir como solucionar este problema.
- Revisar o problema de quadrados mínimos.

- Continuaremos a discutir a resolução de problemas de regressão linear.
- Definiremos o problema de modelo de seleção.



Aula 4.1.2. Regressão linear (Parte 2)

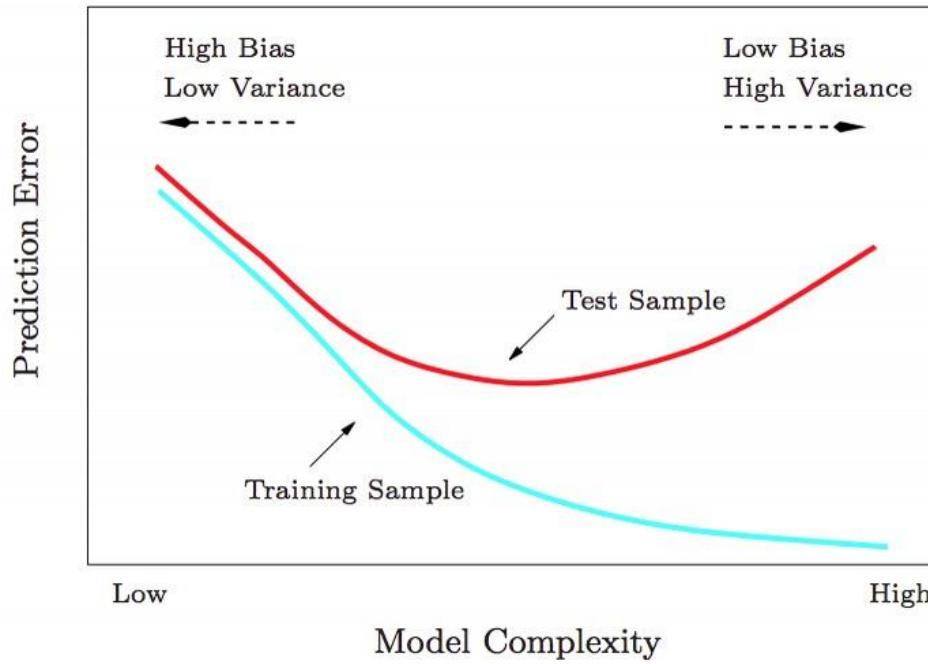
- ❑ Continuaremos a discutir a resolução de problemas de regressão linear.
- ❑ Definiremos o problema de modelo de seleção.
- ❑ Discutir um exemplo real.

- Em termos de consolidação de modelo, nossa próxima questão é:
como escolher a ordem M do polinômio que representa nossa função de regressão?
- Este problema é denominado de seleção de modelos.
- Ao projetar modelos preditivos, almejamos atingir alta generalização para dados novos.
- Entretanto, polinomios de ordem elevada (i.e., com valor de M alto) não são bons em generalização.

Ordem de polinômio

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

Viés vs. Variância

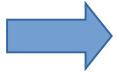


Exemplo - Base de cigarros

	X1	X2	X3	Y
Alpine	14.1	.86	.9853	13.6
Benson&Hedges	16.0	1.06	1.0938	16.6
BullDurham	29.8	2.03	1.1650	23.5
CamelLights	8.0	.67	.9280	10.2
Carlton	4.1	.40	.9462	5.4
Chesterfield	15.0	1.04	.8885	15.0
GoldenLights	8.8	.76	1.0267	9.0
Kent	12.4	.95	.9225	12.3
Kool	16.6	1.12	.9372	16.3
L&M	14.9	1.02	.8858	15.4
LarkLights	13.7	1.01	.9643	13.0
Marlboro	15.1	.90	.9316	14.4
Merit	7.8	.57	.9705	10.0
MultiFilter	11.4	.78	1.1240	10.2
NewportLights	9.0	.74	.8517	9.5
Now	1.0	.13	.7851	1.5
OldGold	17.0	1.26	.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	.96	.9573	17.5
SalemUltra	4.5	.42	.9106	4.9
Tareyton	14.5	1.01	1.0070	15.9
True	7.3	.61	.9806	8.5
ViceroyRichLight	8.6	.69	.9693	10.6
VirginiaSlims	15.2	1.02	.9496	13.9
WinstonLights	12.0	.82	1.1184	14.9

Exemplo - Base de cigarros

X1	X2	X3	Y
14.1	.86	.9853	13.6
16.0	1.06	1.0938	16.6
29.8	2.03	1.1650	23.5
8.0	.67	.9280	10.2
4.1	.40	.9462	5.4
15.0	1.04	.8885	15.0
8.8	.76	1.0267	9.0
12.4	.95	.9225	12.3
16.6	1.12	.9372	16.3
14.9	1.02	.8858	15.4
13.7	1.01	.9643	13.0
15.1	.90	.9316	14.4
7.8	.57	.9705	10.0
11.4	.78	1.1240	10.2
9.0	.74	.8517	9.5
1.0	.13	.7851	1.5
17.0	1.26	.9186	18.5
12.8	1.08	1.0395	12.6
15.8	.96	.9573	17.5
4.5	.42	.9106	4.9
14.5	1.01	1.0070	15.9
7.3	.61	.9806	8.5
8.6	.69	.9693	10.6
15.2	1.02	.9496	13.9
12.0	.82	1.1184	14.9



14.1	.86	.9853
16.0	1.06	1.0938
29.8	2.03	1.1650
8.0	.67	.9280
4.1	.40	.9462
15.0	1.04	.8885
8.8	.76	1.0267
12.4	.95	.9225
16.6	1.12	.9372
14.9	1.02	.8858
13.7	1.01	.9643
15.1	.90	.9316
7.8	.57	.9705
11.4	.78	1.1240
9.0	.74	.8517
1.0	.13	.7851
17.0	1.26	.9186
12.8	1.08	1.0395
15.8	.96	.9573
4.5	.42	.9106
14.5	1.01	1.0070
7.3	.61	.9806
8.6	.69	.9693
15.2	1.02	.9496
12.0	.82	1.1184

13.6
16.6
23.5
10.2
5.4
15.0
9.0
12.3
16.3
15.4
13.0
14.4
10.0
10.2
9.5
1.5
18.5
12.6
7.5
4.9
15.9
8.5
10.6
13.9
14.9

X1

x2

x3

- 8838
- 8642

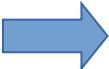
x2

X3

X

Exemplo - Base de cigarros

X1	X2	X3	Y
14.1	.86	.9853	13.6
16.0	1.06	1.0938	16.6
29.8	2.03	1.1650	23.5
8.0	.67	.9280	10.2
4.1	.40	.9462	5.4
15.0	1.04	.8885	15.0
8.8	.76	1.0267	9.0
12.4	.95	.9225	12.3
16.6	1.12	.9372	16.3
14.9	1.02	.8858	15.4
13.7	1.01	.9643	13.0
15.1	.90	.9316	14.4
7.8	.57	.9705	10.0
11.4	.78	1.1240	10.2
9.0	.74	.8517	9.5
1.0	.13	.7851	1.5
17.0	1.26	.9186	18.5
12.8	1.08	1.0395	12.6
15.8	.96	.9573	17.5
4.5	.42	.9106	4.9
14.5	1.01	1.0070	15.9
7.3	.61	.9806	8.5
8.6	.69	.9693	10.6
15.2	1.02	.9496	13.9
12.0	.82	1.1184	14.9



14.1	.86	.9853	13.6
16.0	1.06	1.0938	16.6
29.8	2.03	1.1650	23.5
8.0	.67	.9280	10.2
4.1	.40	.9462	5.4
15.0	1.04	.8885	15.0
8.8	.76	1.0267	9.0
12.4	.95	.9225	12.3
16.6	1.12	.9372	16.3
14.9	1.01	.9643	13.0
15.1	.90	.9316	14.4
7.8	.57	.9705	10.0
11.4	.78	1.1240	10.2
9.0	.74	.8517	9.5
1.0	.13	.7851	1.5
17.0	1.26	.9186	18.5
12.8	1.08	1.0395	12.6
15.8	.96	.9573	17.5
4.5	.42	.9106	4.9
14.5	1.01	1.0070	15.9
7.3	.61	.9806	8.5
8.6	.69	.9693	10.6
15.2	1.02	.9496	13.9
12.0	.82	1.1184	14.9

$$\begin{matrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{matrix} = \begin{matrix} 13.6 \\ 16.6 \\ 23.5 \\ 10.2 \\ 5.4 \\ 15.0 \\ 9.0 \\ 12.3 \\ 16.3 \\ 15.4 \\ 13.0 \\ 14.4 \\ 10.0 \\ 10.2 \\ 9.5 \\ 1.5 \\ 18.5 \\ 12.6 \\ 7.5 \\ 4.9 \\ 15.9 \\ 8.5 \\ 10.6 \\ 13.9 \\ 14.9 \end{matrix}$$

$$\mathbf{A} * \mathbf{X} = \mathbf{Y}$$

Exemplo - Base de cigarros

14.1	.86	.9853
16.0	1.06	1.0938
29.8	2.03	1.1650
8.0	.67	.9280
4.1	.40	.9462
15.0	1.04	.8885
8.8	.76	1.0267
12.4	.95	.9225
16.6	1.12	.9372
14.9	1.02	.8858
13.7	1.01	.9643
15.1	.90	.9316
7.8	.57	.9705
11.4	.78	1.1240
9.0	.74	.8517
1.0	.13	.7851
17.0	1.26	.9186
12.8	1.08	1.0395
15.8	.96	.9573
4.5	.42	.9106
14.5	1.01	1.0070
7.3	.61	.9806
8.6	.69	.9693
15.2	1.02	.9496
12.0	.82	1.1184

X1

X2

X3

*

13.6
16.6
23.5
10.2
5.4
15.0
9.0
12.3
16.3
15.4
13.0
14.4
10.0
10.2
9.5
1.5
18.5
12.6
7.5
4.9
15.9
8.5
10.6
13.9
14.9



ε1	14.1	.86	.9853	13.6
ε2	16.0	1.06	1.0938	16.6
ε3	29.8	2.03	1.1650	23.5
ε4	8.0	.67	.9280	10.2
ε5	4.1	.40	.9462	5.4
ε6	15.0	1.04	.8885	15.0
ε7	8.8	.76	1.0267	9.0
ε8	12.4	.95	.9225	12.3
ε9	16.6	1.12	.9372	16.3
ε10	14.9	1.02	.8858	15.4
ε11	13.7	1.01	.9643	13.0
ε12	15.1	.90	.9316	14.4
ε13	7.8	.57	.9705	10.0
ε14	11.4	.78	1.1240	10.2
ε15	9.0	.74	.8517	9.5
ε16	1.0	.13	.7851	1.5
ε17	17.0	1.26	.9186	18.5
ε18	12.8	1.08	1.0395	12.6
ε19	15.8	.96	.9573	7.5
ε20	4.5	.42	.9106	4.9
ε21	14.5	1.01	1.0070	15.9
ε22	7.3	.61	.9806	8.5
ε23	8.6	.69	.9693	10.6
ε24	15.2	1.02	.9496	13.9
ε25	12.0	.82	1.1184	14.9

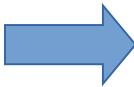
A

* X = Y

E + A * X = Y

Exemplo - Base de cigarros

A Constance **c** representa
comportamentos não relacionados às
variaveis independentes.



$$Y = c + X_1 + X_2 + X_3$$

ϵ_1	1	14.1	.86	.9853	13.6
ϵ_2	1	16.0	1.06	1.0938	16.6
ϵ_3	1	29.8	2.03	1.1650	23.5
ϵ_4	1	8.0	.67	.9280	10.2
ϵ_5	1	4.1	.40	.9462	5.4
ϵ_6	1	15.0	1.04	.8885	15.0
ϵ_7	1	8.8	.76	1.0267	9.0
ϵ_8	1	12.4	.95	.9225	12.3
ϵ_9	1	16.6	1.12	.9372	16.3
ϵ_{10}	1	14.9	1.02	.8858	15.4
ϵ_{11}	1	13.7	1.01	.9643	13.0
ϵ_{12}	1	15.1	.90	.9316	14.4
ϵ_{13}	1	7.8	.57	.9705	10.0
ϵ_{14}	1	11.4	.78	1.1240	10.2
ϵ_{15}	1	9.0	.74	.8517	9.5
ϵ_{16}	1	1.0	.13	.7851	1.5
ϵ_{17}	1	17.0	1.26	.9186	18.5
ϵ_{18}	1	12.8	1.08	1.0395	12.6
ϵ_{19}	1	15.8	.96	.9573	7.5
ϵ_{20}	1	4.5	.42	.9106	4.9
ϵ_{21}	1	14.5	1.01	1.0070	15.9
ϵ_{22}	1	7.3	.61	.9806	8.5
ϵ_{23}	1	8.6	.69	.9693	10.6
ϵ_{24}	1	15.2	1.02	.9496	13.9
ϵ_{25}	1	12.0	.82	1.1184	14.9

$$* \quad \begin{matrix} \textbf{c} \\ \textbf{X}_1 \\ \textbf{X}_2 \\ \textbf{X}_3 \end{matrix} =$$

$$\mathbf{E} + \mathbf{A} * \mathbf{X} = \mathbf{Y}$$

Conclusão

- Continuaremos a discutir a resolução de problemas de regressão linear.
- Definiremos o problema de modelo de seleção.
- Discutir um exemplo real.

■ Próxima aula

- ❑ Regressão logística.
- ❑ Quando utilizar regressão logística ou linear?
- ❑ Como resolver regressão logística?



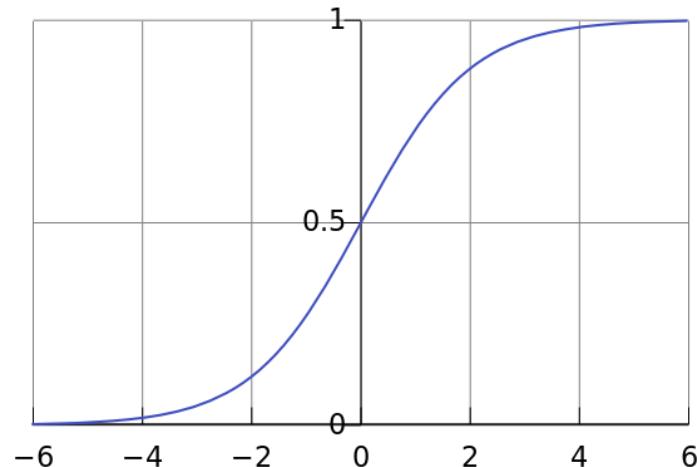
Aula 4.2. Regressão logística

- ❑ Apresentaremos modelos de regressão logística.
- ❑ Discutiremos os principais conceitos relacionados a este método.
- ❑ Revisaremos o método de geração de modelos para regressão logística.

Regressão logística

- Método recomendado para situações em que a variável dependente é de natureza dicotômica ou binária.
- Produz uma estimativa de probabilidade para cada objeto pertencente a uma classe:
 - Os resultados da análise ficam contidos no intervalo de zero a um.
- Funciona para conjuntos de dados relativamente grandes.
- Rápido de se aplicar.

Função Logit



$$w \cdot x = B_0 + B_1 X_1 + \dots + B_p X_p$$

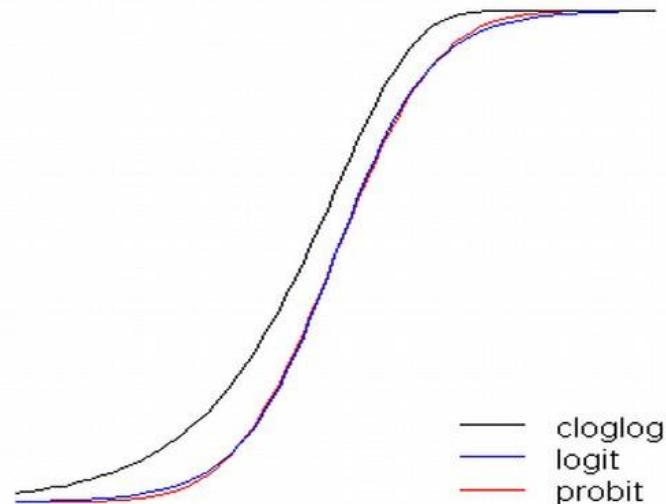
- **Previsão de risco na área tributária:** calcular a probabilidade do contribuinte ser inadimplente após o parcelamento de tributos.
- **Classificação de empresas:** classificar se a empresa encontra-se no grupo de empresas solvente, ou insolvente.
- **Identificação de pacientes doentes:** determinar os fatores que caracterizam um grupo de indivíduos doentes em relação aos indivíduos sãos.

Função Logit - Aprendizado

- Os coeficientes B_0 , B_1 , ... , B_p , são estimados a partir do conjunto dados, pelo método da máxima verossimilhança (MLE).
- Tais coeficientes maximizam a probabilidade da amostra ter sido observada.
- Não deve haver outliers nos dados, os quais podem ser removidos, normalizando-se as variáveis independentes via z-score e removendo valores maiores que 3.29 e menores que -3.29.
- Não deve haver intercorrelações elevadas entre as variáveis independentes.

- Às vezes, em vez de um modelo logit para regressão logística, um modelo probit é usado.
- Na maioria dos casos, um modelo é construído com ambas as funções e a função com melhor ajuste é escolhida.
- A função Probit assume a distribuição normal da probabilidade do evento, enquanto a função logit assume a distribuição do log.
- Na função logit a distribuição condicional $y|x$ é uma distribuição de Bernoulli, em vez de uma distribuição Gaussiana.

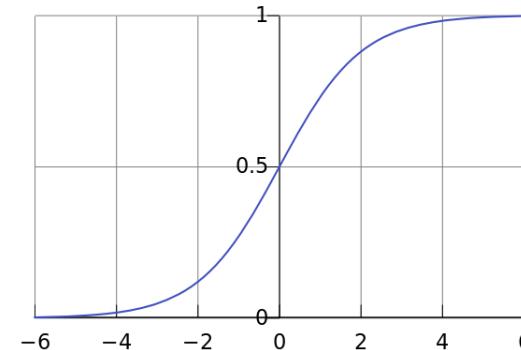
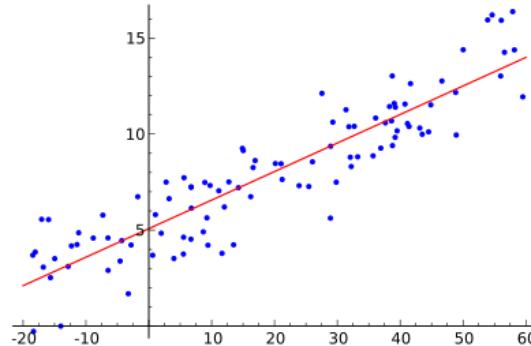
Função Logit vs. Profit



- Facilidade para lidar com variáveis independentes categóricas.
- Fornece resultados em termos de probabilidade.
- Facilidade de classificação de indivíduos em categorias.
- Requer pequeno número de suposições.
- Alto grau de confiabilidade.

Regressão logística vs. Linear

- **Linear:** Utiliza o método dos mínimos quadrados.
- **Logística:** Utiliza o método da máxima verossimilhança.



Conclusão

- Apresentaremos modelos de regressão logística.
- Discutiremos os principais conceitos relacionados a este método.
- Revisaremos o método de geração de modelos para regressão logística.

■ Próxima aula

- KNN.
- Árvores de regressão.
- Árvores de classificação.



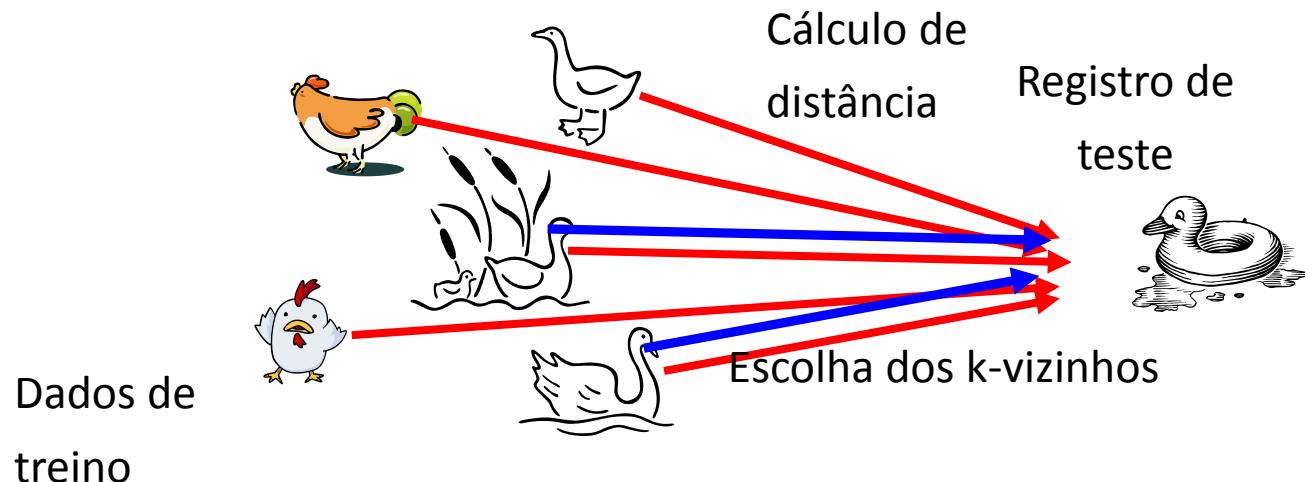
Aula 4.3. KNN & CART

- Discutir o algoritmo KNN.
- Apresentar o método CART.

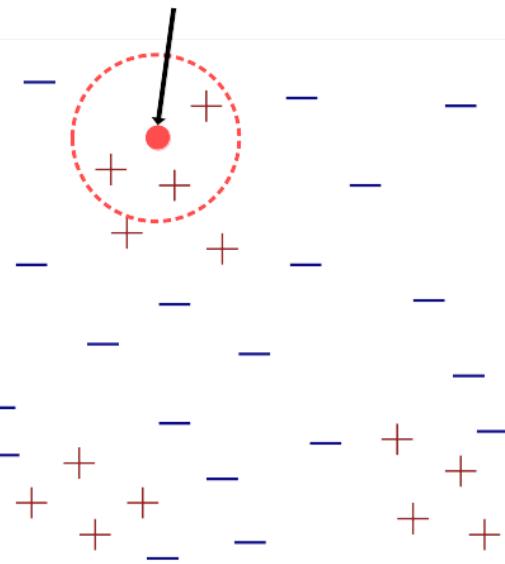
- Aprendizado supervisionado.
- O método mais básico de classificação baseado em instâncias.
- Os dados são representados em um espaço vetorial e os vizinhos são definidos em termos de distância.
- No cálculo do vizinho mais próximo, a função alvo pode retornar valores discretos ou reais.
- Para valores discretos, o KNN retorna a moda entre os K exemplos de treinamento mais próximos.

KNN - Princípio

- Princípio: se anda como um pato, granya como um pato, então é um pato!

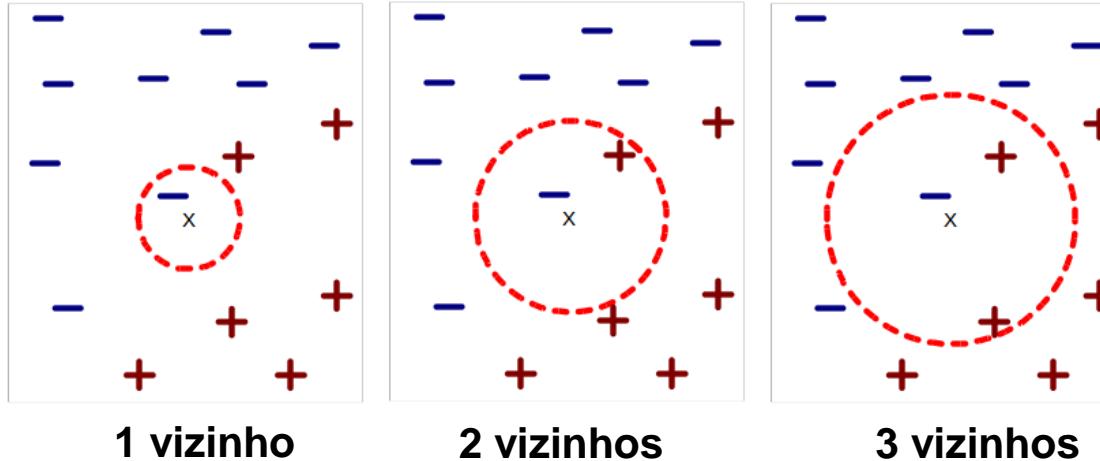


Objeto alvo



- Três requisitos básicos:
 - Conjunto de treinamento;
 - Métrica de distância;
 - Número de vizinhos a inspecionar.
- Como funciona:
 - Computa a distância do objeto alvo e todos os outros objetos de treino;
 - Seleciona os K mais próximos;
 - Usa a classe dos K vizinhos para prever a classe do objeto alvo.

KNN - Vizinhança



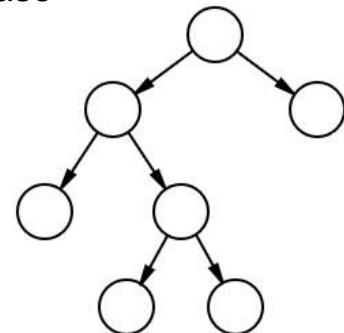
Quantos vizinhos são necessários?

- O valor de K é determinado experimentalmente.
- Heurística de seleção do K ótimo:
 - Comece com $K = 3$ e use um conjunto de teste para validar a taxa de erro do classificador.
 - Repita com $K = K + 2$.
 - Escolha o valor de K para o qual a taxa de erro é mínima.
- Tente sempre utilizar um número ímpar para K, de forma a reduzir as chances de empates durante a classificação.

- Árvores de classificação e regressão (Classification And Regression Trees - CART).
- As árvores de classificação e regressão (CART) são árvores binárias não-paramétricas, que produzem árvores de classificação ou regressão dependendo se a variável dependente é categórica ou numérica, respectivamente.
- Desenvolvido por Breiman, Friedman, Olshen e Stone no início dos anos 80.
- Introduziu modelagem estatística baseada em árvores
- Abordagem rigorosa envolvendo validação cruzada para selecionar a árvore ideal.

CART - Particionamento recursivo

- O funcionamento consiste na construção iterativa de árvores binárias:
 - Comece pelo nodo raiz;
 - Analisa-se todos os dados de entrada;
 - Inspeciona todos os possíveis valores de todas as variáveis (força-bruta);
 - Seleciona a variável/valor ($X=vi$) que produz a melhor separação dos dados, de acordo com a função objetivo;
 - Para cada observação, se $X < vi$, coloque a observação no nodo esquerdo, caso contrário assinale-a para o nodo direito;
 - Repita o processo para cada nodo.



- O principal ponto do algoritmo é, a cada iteração, selecionar o valor de variável que produza a melhor separação dos dados de entrada, de acordo com a função objetivo.
- Separação pode ser definida de diversas formas:
 - Árvores de regressão: possui uma função objetiva continua, e a medida de separação usada é a soma dos erros quadrados (SSE).
 - Árvores de classificação: possui uma função objetivo categórica ,e as principais medidas de separação são: Gini measure e Entropia, dentre outras que visam estimar a pureza de cada nodo na árvore.

Conclusão

- Discutir o algoritmo KNN.
- Apresentar o método CART.

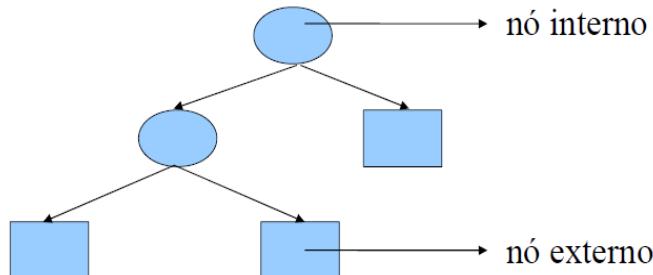
- Árvores de decisão.
- Boosted Trees.
- Random Forests.



Aula 4.4. Árvores de decisão e Ensemble Trees

- Revisão das árvores de decisão.
- Apresentar os métodos de Ensemble Trees.
- Discutir sobre Random Forests.
- Discutir sobre Boosted Trees.

- Uma árvore de decisão é uma ferramenta de suporte à decisão, que usa estrutura hierárquica para representar visualmente decisões.
- Árvores n-árias
- Dois tipos: árvores de classificação e regressão.



Árvores de decisão - Exemplo

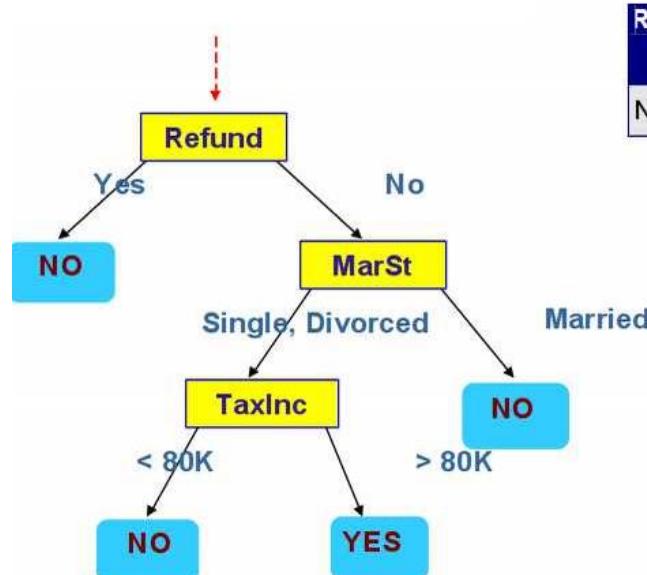
Tid	Refund	Marital Status	Taxable Income	Cheat	
				categorical	categorical
				continuous	dass
1	Yes	Single	125K	No	
2	No	Married	100K	No	
3	No	Single	70K	No	
4	Yes	Married	120K	No	
5	No	Divorced	95K	Yes	
6	No	Married	60K	No	
7	Yes	Divorced	220K	No	
8	No	Single	85K	Yes	
9	No	Married	75K	No	
10	No	Single	90K	Yes	

Training Data



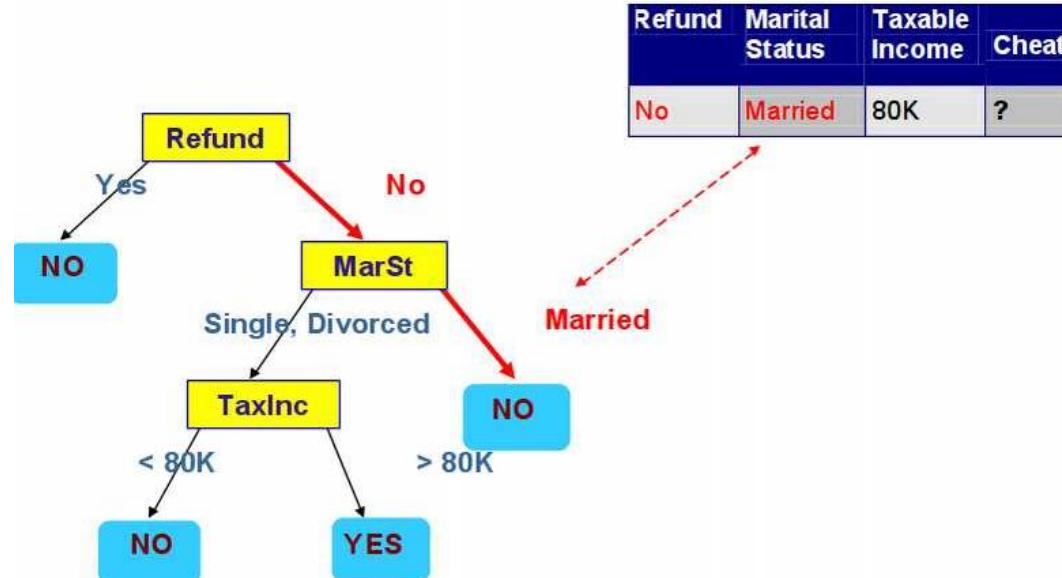
Model: Decision Tree

Árvores de decisão - Como classificar novos itens?



Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Árvores de decisão - Como classificar novos itens?

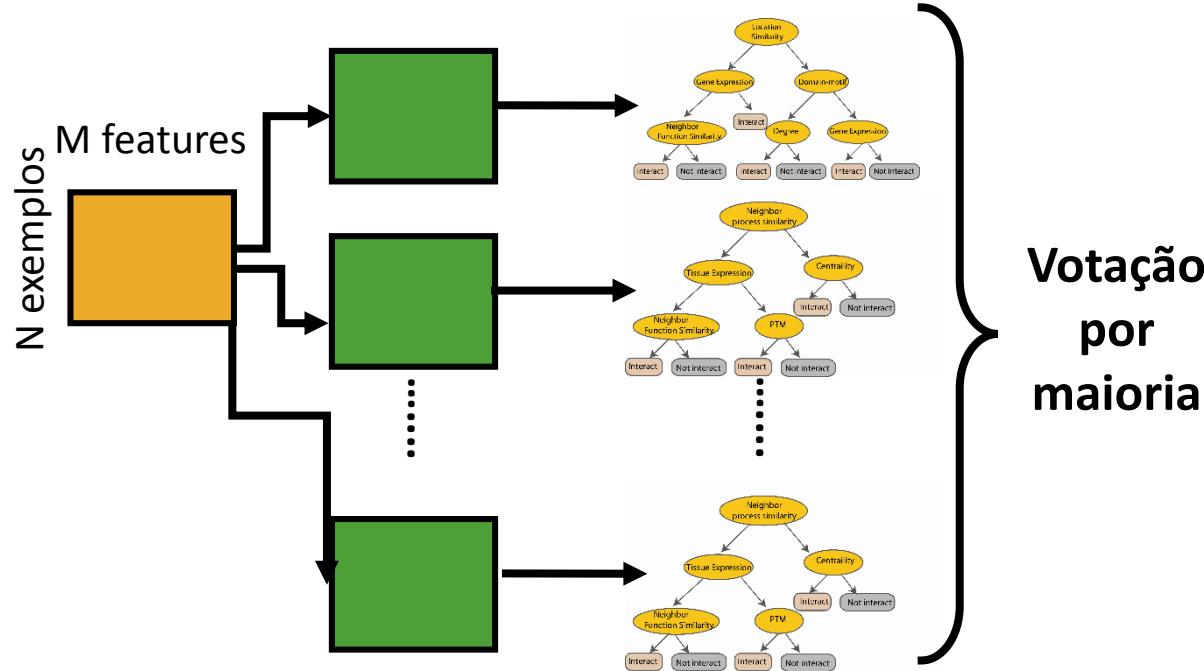


- Um grande problema no aprendizado de árvores de decisão é sobre a árvore ótima.
- Em geral, quanto maior a árvore, menor é a sua capacidade de generalização.
- O problema de overfitting ocorre com frequência no aprendizado de modelos de árvores de decisão.
- Como solucionar este problema?

- Em vez de aprender um único classificador (fraco), aprendemos muitos classificadores fracos que são bons em diferentes partes do espaço de entrada.
- Classe de saída: (ponderada) voto de cada classificador.
- Classificadores que estão mais “certos” votarão com mais convicção.
- Na média, esta estratégia se torna melhor que um único classificador forte!
- Mas, como fazer isso?
 - Forçar classificadores a aprender sobre diferentes partes do espaço de entrada?
 - Pesar os votos de diferentes classificadores?

- Princípio: combinar diferentes modelos que exploram características e dados distintos.
- Considere o conjunto de treinamento contendo N observações:
 - Gere K amostras distintas de tamanho N, a partir dos dados de entrada com repetição.
 - Aprenda uma árvore de decisão sobre cada amostra, selecionando-se aleatoriamente m das p variáveis existentes nos dados originais.
 - Cada árvore de decisão realizará previsões sobre as mesmas observações de teste.
 - Agregue as previsões dadas por cada modelo, de forma a gerar uma única previsão final para cada observação.

Random Forests

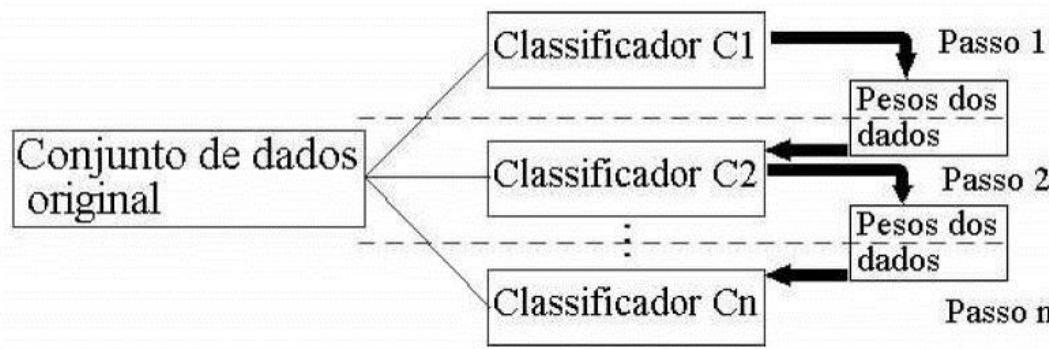


- Um dos classificadores mais populares para dados densos.
- Fácil de implementar.
- Facilmente paralelizável.
- Não é o método estado-da-arte.
- Precisa de várias iterações sobre os dados .

- Consiste em escolher o conjunto de treinamento usado por cada classificador base, o qual é uma sequência de classificadores, de acordo com o desempenho dos classificadores que obtiveram o melhor resultado.
- É necessário que os classificadores base do algoritmo sejam treinados sequencialmente, visando definir os padrões que irão constituir os próximos conjuntos de treinamento.
- O classificador final é obtido por um esquema de votação a partir dos resultados obtidos pelos classificadores base.

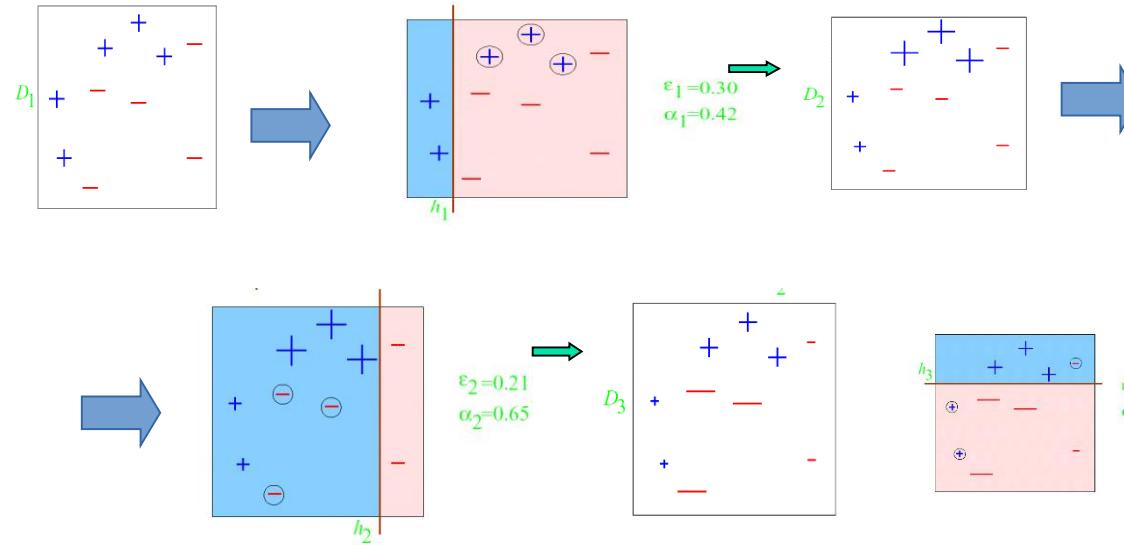
Boosted Decision Trees

- O mesmo algoritmo é executado de forma recursiva até convergir.



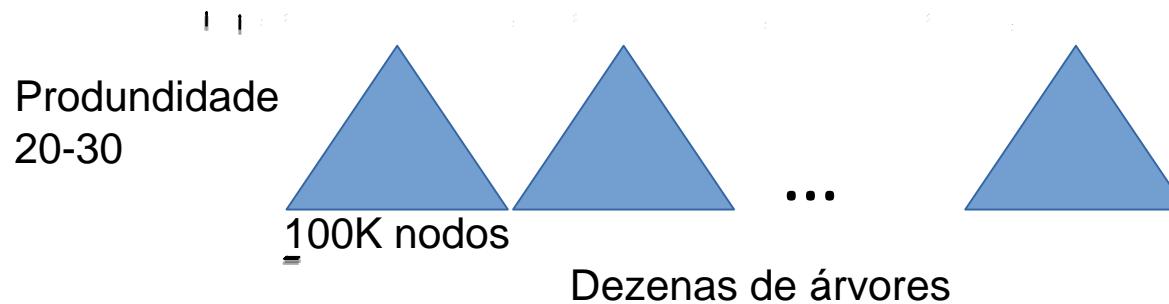
Boosted Decision Trees

- Os objetos “difíceis” são ponderados e o algoritmo executa nesse novo cenário.

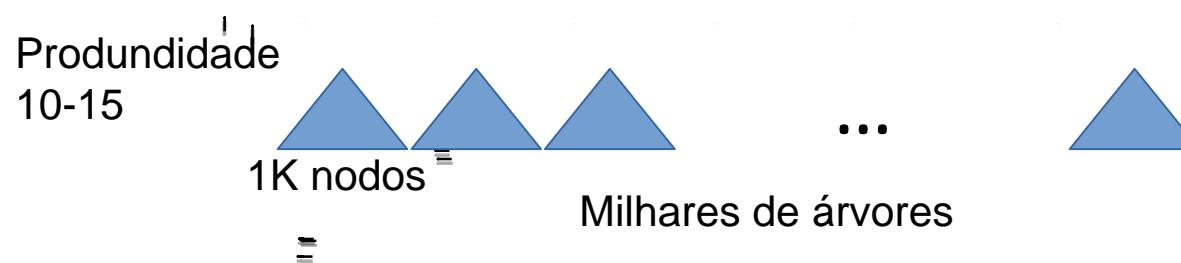


Random Forests vs. Boosted Decision Trees

Random Forests



Boosted Trees



- Revisão das árvores de decisão.
- Apresentar os métodos de Ensemble Trees.
- Discutir sobre Random Forests.
- Discutir sobre Boosted Trees.

- SVM.
- Kernel Trick.



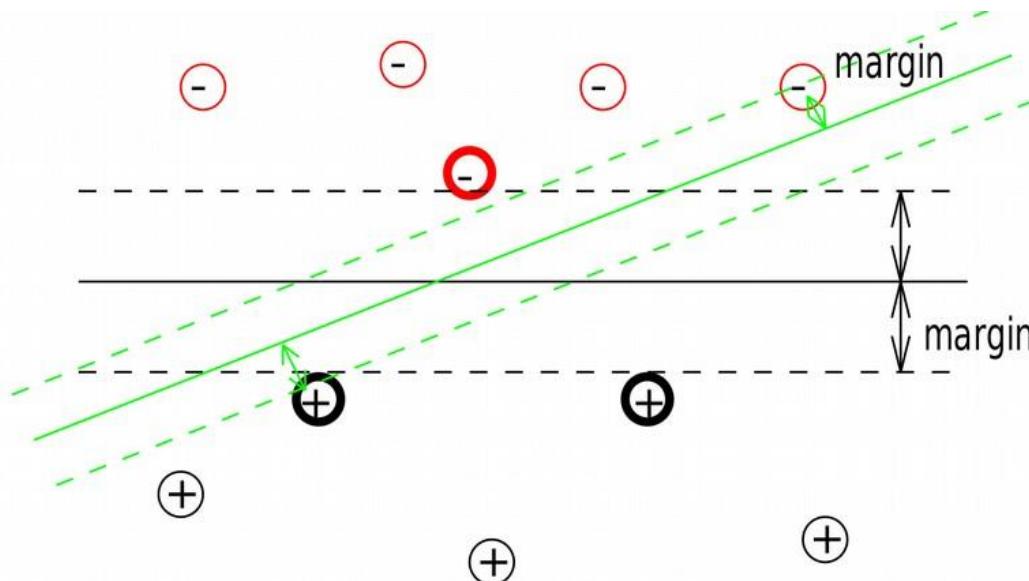
Aula 4.5. SVM

- Apresentar o método SVM.
- Discutir a ideia básica por trás deste algoritmo.
- Apresentar a estratégia utilizada para o SVM resolver problemas não lineares.

Máquinas de vetor de suporte – SVM

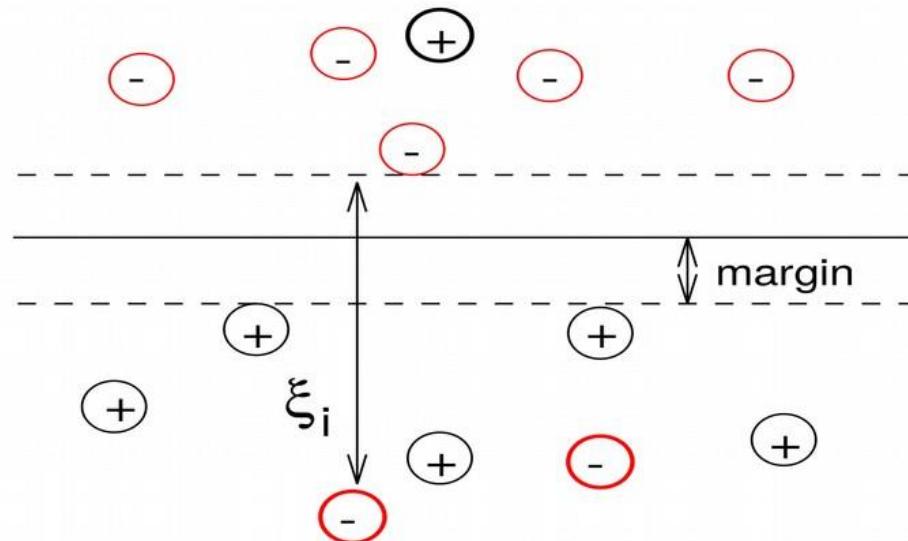
- Utilizam um algoritmo de treinamento eficiente.
- Amplamente utilizado para classificação de documentos, manuscritos, imagens, dentre outros.
- Capaz de produzir apenas modelos lineares.

Máquinas de vetor de suporte – SVM



- É fácil encontrar um separador linear no espaço correspondente de dimensão elevada?
 - Infelizmente não.
- Um separador linear em um espaço de d dimensões é definido por uma equação com d parâmetros.
- SVM encontram o separador linear ótimo:
 - Aquele que tem a maior **margem** entre ele e os exemplos positivos de um lado e os negativos de outro.

Máquinas de vetor de suporte – SVM



- Como encontrar o separador linear?
 - Problema de otimização de **programação quadrática**.
- Suponha exemplos \mathbf{x}_i com classificações $y_i = +1$:
 - Encontrar um separador ótimo no espaço de entrada;
 - O problema de programação quadrática é encontrar valores dos parâmetros α_i que maximizem a expressão:

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

- Sujeita às restrições $\alpha_i \geq 0$ e $\sum \alpha_i y_i = 0$.

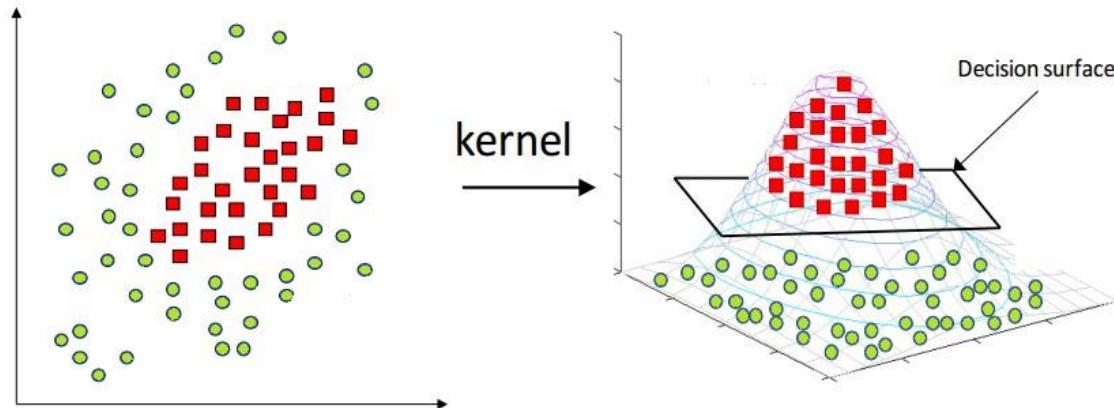
- A derivação da expressão anterior tem duas propriedades importantes:
 - Tem um único máximo global que pode ser encontrado de forma eficiente.
 - Os dados entram na expressão apenas sob a forma de produtos pontuais de pares de pontos.
- Uma vez que os valores ótimos de α_i foram calculados, a equação correspondente ao próprio separador é:

$$h(x) = \text{sign} \left(\sum_i \alpha_i y_i (x \cdot x_i) \right)$$

■ Problemas não lineares

- É possível usar SVM pra resolver problemas não lineares?
 - Sim.
- Se os dados forem mapeados em um espaço de dimensão suficientemente alta, então eles sempre serão linearmente separáveis.
- É fácil encontrar um separador linear no espaço correspondente de dimensão elevada?
 - Infelizmente não, devido ao mal da dimensionalidade.

Problemas não lineares



- Existe outra solução além de se tentar projetar os dados em uma dimensão arbitrariamente grande.
- Funções Kernel: são funções que retornam o produto escalar das imagens de seus argumentos.
- Em outras palavras, podemos dizer que funções Kernel nos permitem calcular o produto escalar de pontos no espaço euclidiano original, como se estivessem em um espaço maior (ou menor).
- Como o SVM usa apenas o produto escalar entre pares de pontos para definir as margens, o uso de Kernels se tornou predominante.

Conclusão

- Apresentar o método SVM.
- Discutir a ideia básica por trás deste algoritmo.
- Apresentar a estratégia utilizada para o SVM resolver problemas não lineares.

- Revisar o algoritmo de redes neurais.
- Discutir o processo de treinamento de redes neurais.
- Discutir os principais conceitos relacionados a este tópico.

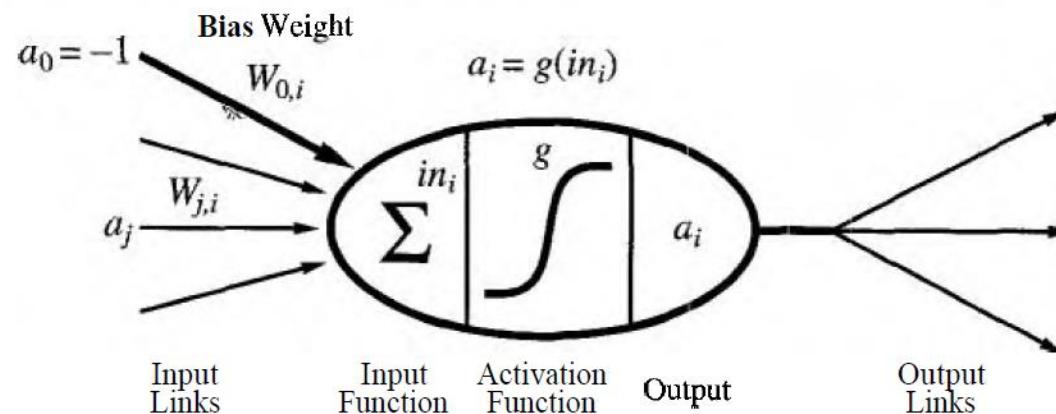


Aula 4.6. Redes neurais

- Apresentar os principais conceitos relacionados as redes neurais.
- Revisar o funcionamento deste tipo de algoritmo.
- Discutir sobre o processo de treinamento de redes neurais.

- Acredita-se que a capacidade de processamento de informações do cérebro emerge de *redes* de neurônios.
- Por essa razão, uma parte do trabalho inicial de IA teve como objetivo criar **redes neurais** artificiais.
- Este campo também é chamado de **conexionismo, processamento distribuído paralelo e computação neural**.

Modelo de neurônio



- As redes neurais são compostas de nós ou unidades conectadas por vínculos orientados.
- Um vínculo da unidade j para a unidade i , serve para propagar a ativação a_j desde j até i .
- Cada vínculo tem um peso numérico W_{ji} associado a ele, o qual determina a intensidade e o sinal da conexão.
- Cada unidade i calcula primeiro uma soma ponderada de suas entradas:

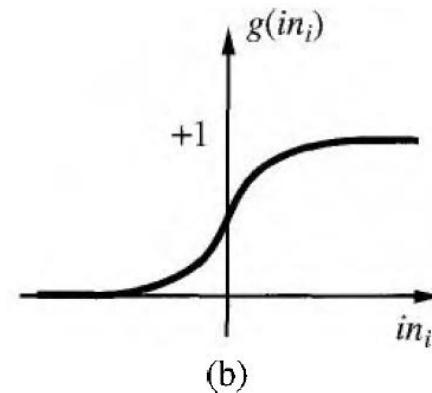
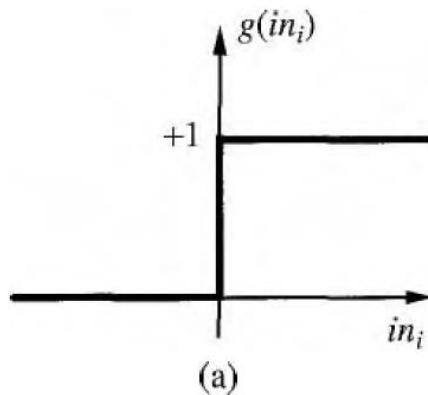
$$in_i = \sum_{j=0}^n W_{j,i} a_j$$

- Então, ela aplica uma função de ativação g a essa soma, para derivar a saída:

$$a_i = g(in_i) = g\left(\sum_{j=0}^n W_{j,i} a_j\right)$$

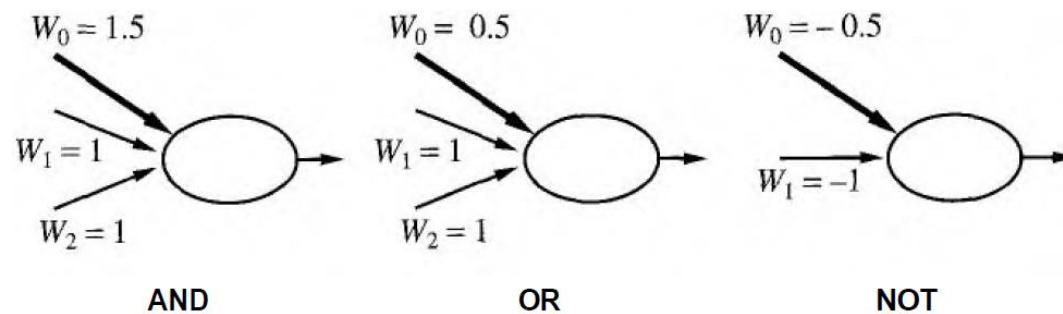
- A função de ativação g é projetada para atender a duas aspirações:
 - Que a unidade seja “ativa” (1) quando as entradas “corretas” forem recebidas, e que a unidade seja “inativa” (0) quando as entradas “erradas” forem recebidas.
 - A ativação precisa ser não linear, caso contrário a rede neural inteira entrará em colapso, tornando-se uma função linear simples.

Exemplo de função de ativação



Unidades individuais

- Motivação inicial para as unidades individuais: sua habilidade para representar funções booleanas básicas.



Estrutura da rede

- Existem duas categorias de estruturas de redes neurais:
 - Redes acíclicas ou **redes de alimentação direta**:
 - Representa uma função de sua entrada atual.
 - Não tem nenhum estado interno além dos pesos propriamente ditos.
 - Redes cíclicas ou **redes recorrentes**:
 - Utiliza suas saídas para alimentar de volta suas próprias entradas.
 - Os níveis de ativação da rede formam um sistema dinâmico que podem atingir um estado estável ou exibir oscilações, ou mesmo apresentar um comportamento caótico.
 - A resposta da rede a uma determinada entrada depende de seu estado inicial, que pode depender de entradas anteriores.

Conclusão

- Apresentar os principais conceitos relacionados as redes neurais.
- Revisar o funcionamento deste tipo de algoritmo.
- Discutir sobre o processo de treinamento de redes neurais.

- Meta-aprendizado.
- Bagging.
- Boosting.
- Stacking.



Aula 4.8. Meta-aprendizado

- Introduzir o conceito de meta-aprendizado.
- Revisar a estratégia de Bagging.
- Revisar a estratégia de Boosting.
- Revisar a estratégia de Stacking.

- Processo de combinação de algoritmos ou dados:
 - Melhora a acurácia.
 - Executa múltiplos algoritmos.
 - Executa o mesmo algoritmo múltiplas vezes.
 - Mas, quando isso funciona?
 - Quando os algoritmos são “bons” e independentes.

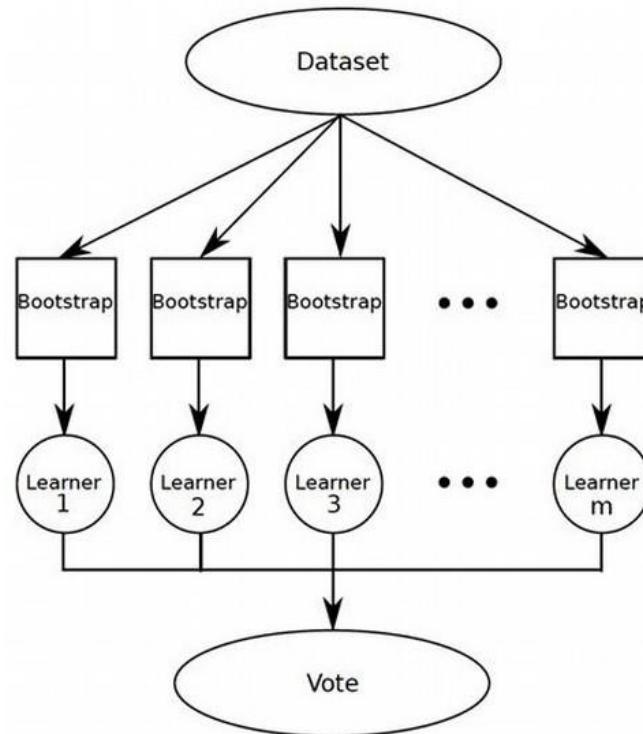
- Um sistema de meta-aprendizagem combina três requisitos:
 - O sistema deve incluir um subsistema de aprendizagem;
 - A experiência é adquirida pelo uso de meta-conhecimento, extraído de uma coleção de dados ou de domínios diferentes;
 - O aprendizado deve ser definido dinamicamente.

- Por que funciona?
 - Porque erros não correlacionados podem ser eliminados.
 - A função real pode não ser diretamente implementável por um único algoritmo.
 - Erro é decorrente de viés e variância.
 - Viés → culpa do algoritmo.
 - Variância → discrepância entre treino e teste.

- Como combinar os algoritmos?
 - Não-supervisionadas: o desempenho de cada algoritmo não é levado em conta.
 - Supervisionadas: existe uma ou mais fases de treinamento.

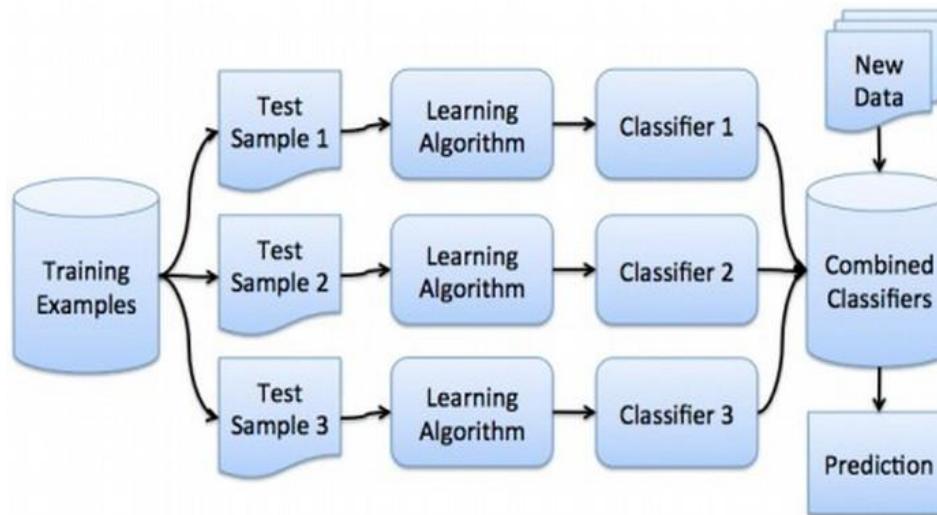
- Não supervisionado.
- Implementa votação.
- Executa o mesmo algoritmo:
 - Cada um utiliza diferentes amostras dos dados com repetição (i.e., Bootstrap).
- As previsões são combinadas para criar a função de mapeamento:
 - Média.
 - Votação majoritária.
 - Média da probabilidades.

Bagging



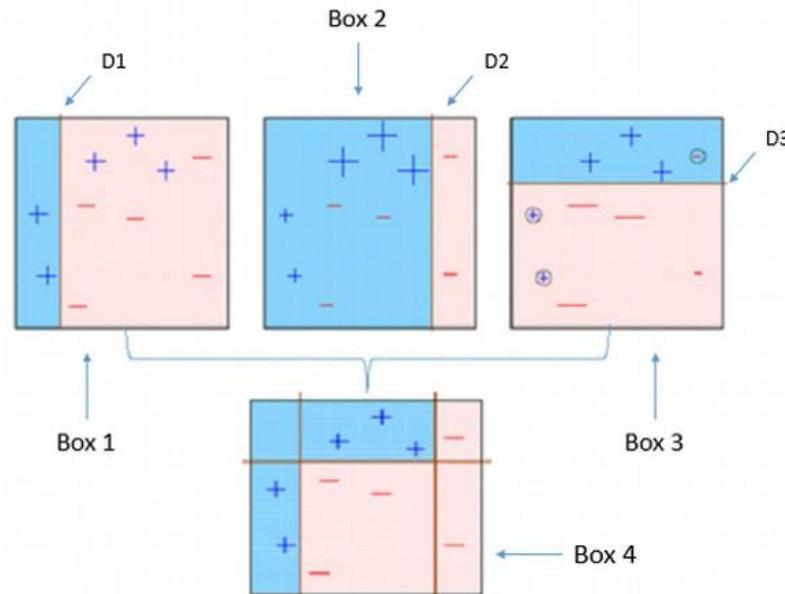
- Supervisionado.
- O mesmo algoritmo é executado de forma recursiva, até haver convergência.
- Os objetos “difíceis” são ponderados e o algoritmo executa nesse novo cenário:
 - Objetos para os quais o algoritmo erra, são inseridos mais vezes nos dados de treino.

Boosting



- Processo iterativo:
 - Algoritmo gera a função;
 - A função geralmente é muito simples (i.e., weak learner).
 - Avalia-se os erros e os acertos no treino;
 - No conjunto de treino/validação.
 - Pondera-se os erros.
 - Objetos errados aparecem mais vezes.
- No final, a função é dada por uma combinação de cada iteração.

Boosting



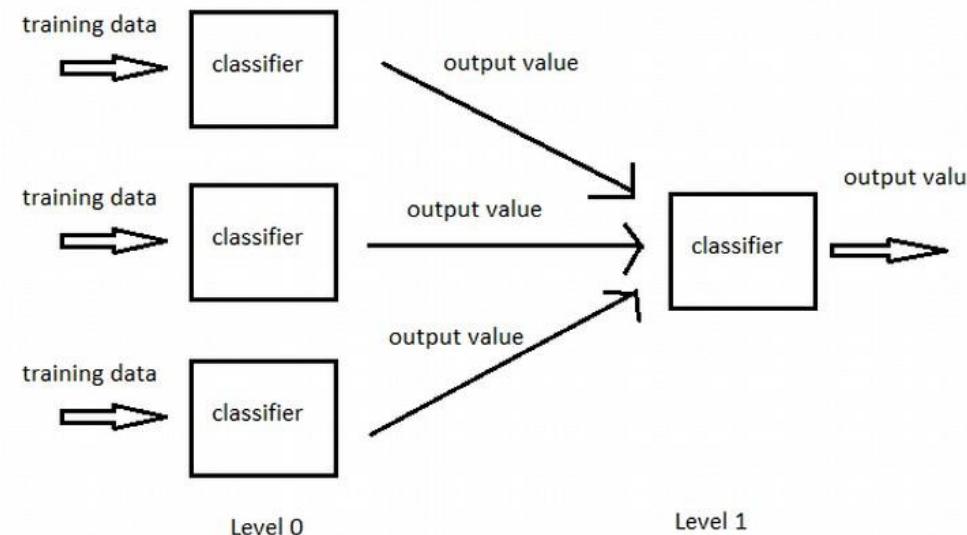
Bagging vs. Boosting

- Bagging sempre usa reamostragem ao invés de responderação.
- Bagging não modifica a distribuição de exemplos, mas usa sempre a distribuição uniforme.
- No voto final, bagging dá igual peso a cada um dos modelos fracos.

- Supervisionado.
- Usa múltiplos algoritmos e combina os resultados.
- A saída de uma iteração serve como entrada para a próxima.
- Ajuda a reduzir problemas de um algoritmo específico.

Stacking

Concept Diagram of Stacking



Conclusão

- Introduzir o conceito de meta-aprendizado.
- Revisar a estratégia de Bagging.
- Revisar a estratégia de Boosting.
- Revisar a estratégia de Stacking.

- Reamostragem.
- Hold-out.
- Cross-Validation.
- Jackknife.
- Bootstrap.

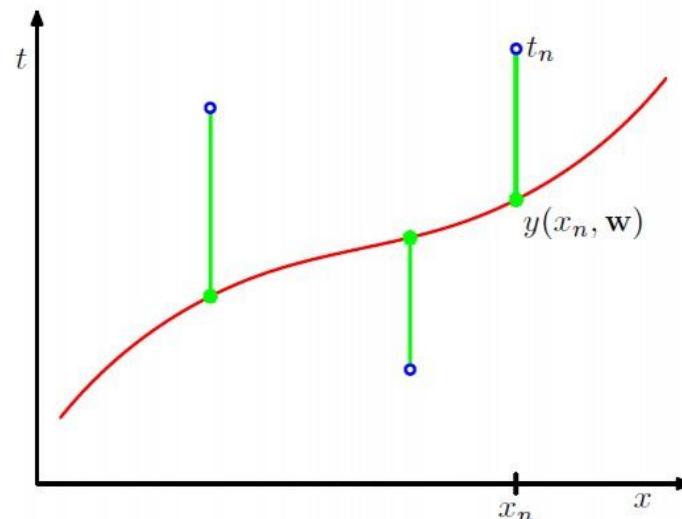


Aula 4.7. Seleção e ajuste de modelos

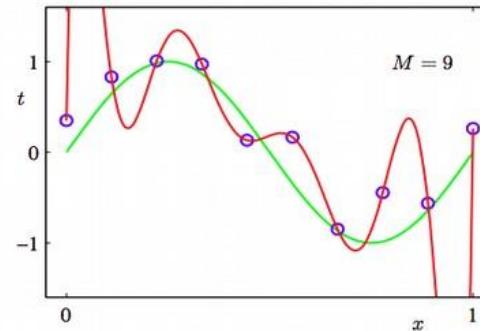
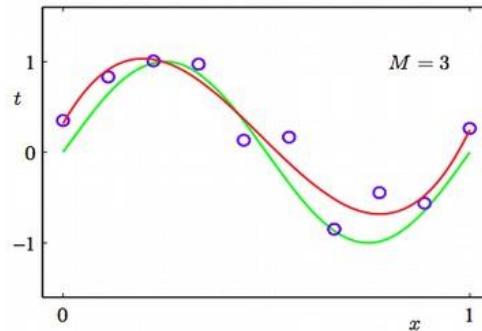
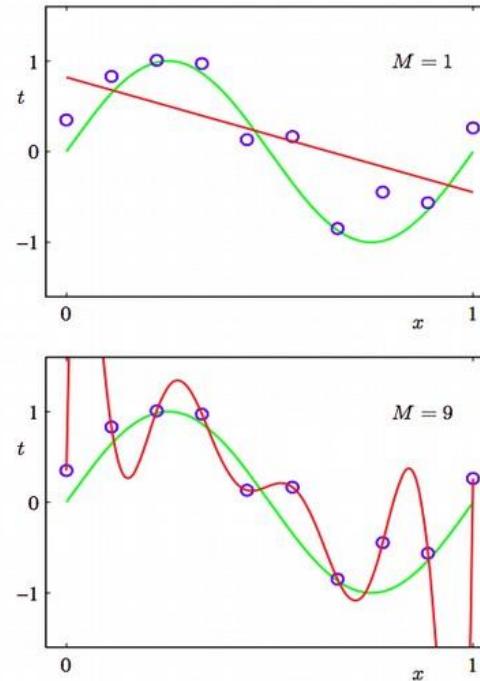
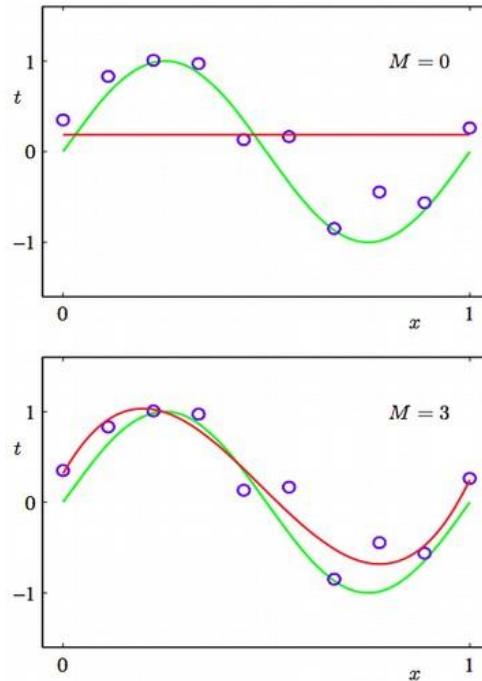
- Revisar função objetivo e seleção de modelo.
- Discutir sobre overfitting.
- Discutir sobre underfitting.
- Revisar o processo de regularização.

Função objetivo

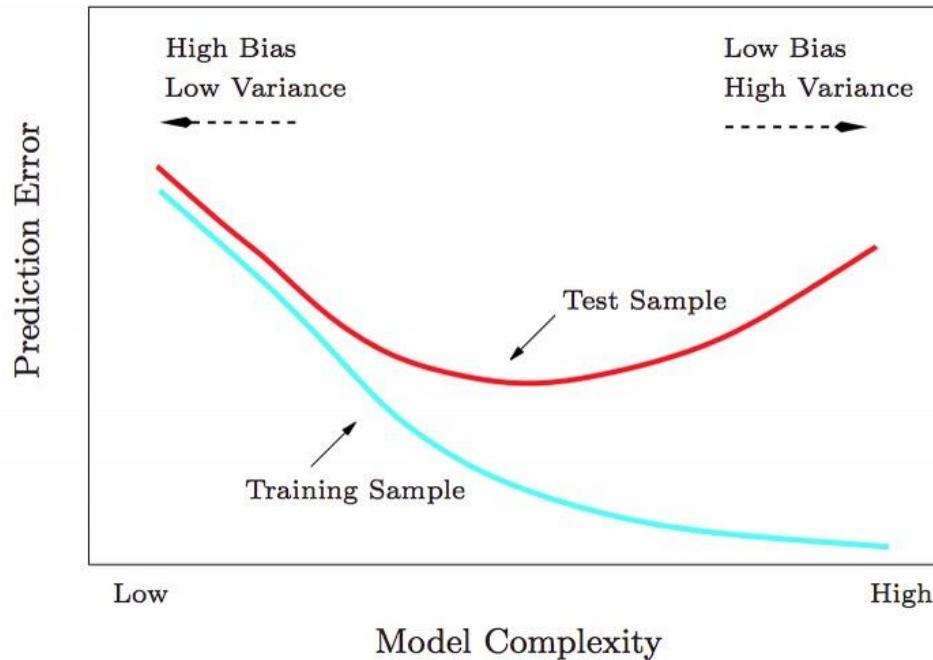
- O objetivo do aprendizado é minimizar os erros da função objetivo (função de erro):
$$E(\mathbf{w}) = \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$
- O mínimo corresponde a derivadas com respeito aos coeficientes.



Seleção de modelo



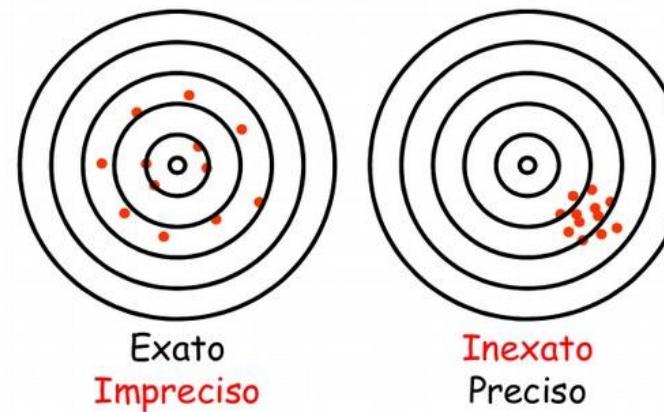
Viés vs. Variância



Incerteza em predições

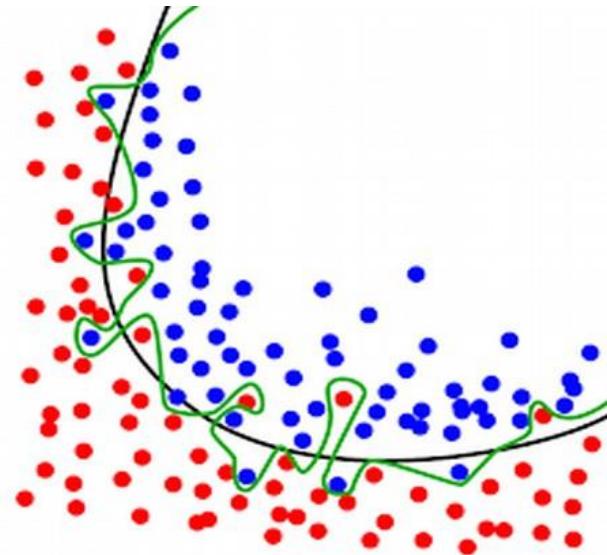
Característica ideal de um estimador

- não tendencioso ↔ exatidão/acurácia
- variância mínima ↔ precisão/incerteza



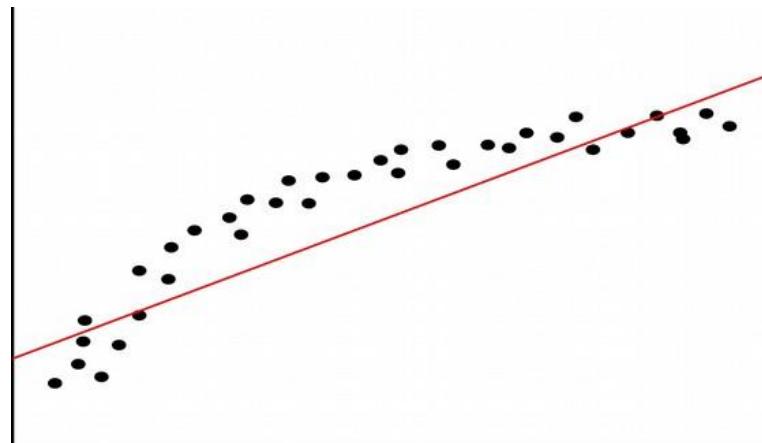
Overfitting

- Overfitting: o modelo memoriza os dados, ao invés de aprender a partir deles (alta variância).



Underfitting

- Underfitting: o modelo ignora o que os dados estão dizendo (alto viés bias).

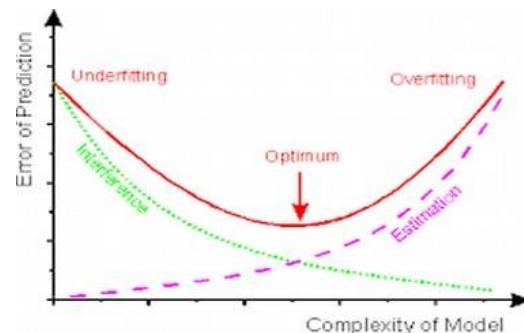
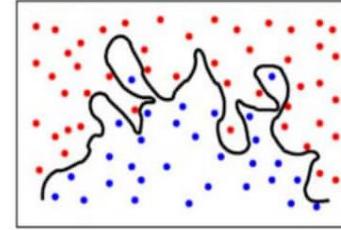
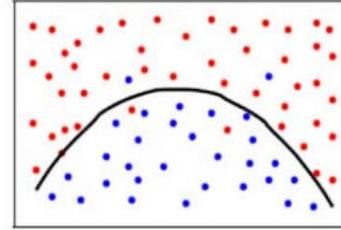
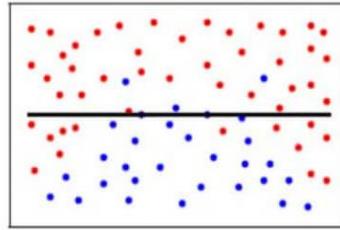


Modelo Ótimo

Underfitting

↔

Overfitting



Regularização

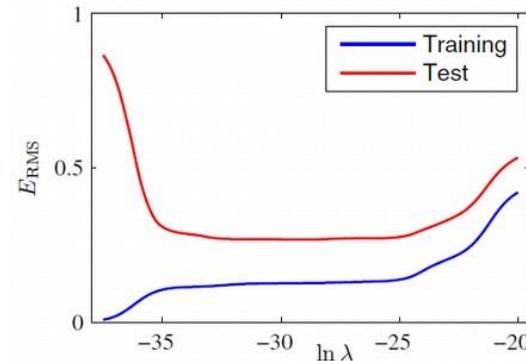
- Regularização é uma técnica para controlar overfitting
- Basicamente, adiciona-se um fator de penalização à função de erro, de forma a penalizar coeficientes com magnitudes altas.

$$\tilde{E}(\mathbf{w}) = \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \dots + w_M^2$$

Regularização

- λ determina a importância relativa do termo de regularização.
- λ controla a complexidade do modelo e o grau de overfitting.
- λ deve ser aprendizado através de um conjunto de validação (hold-out).



Conclusão

- Revisar função objetivo e seleção de modelo.
- Discutir sobre overfitting.
- Discutir sobre underfitting.
- Revisar o processo de regularização.

■ Próxima aula

- Discutir overfitting.
- Discutir underfitting.
- Apresentar o método de regularização.



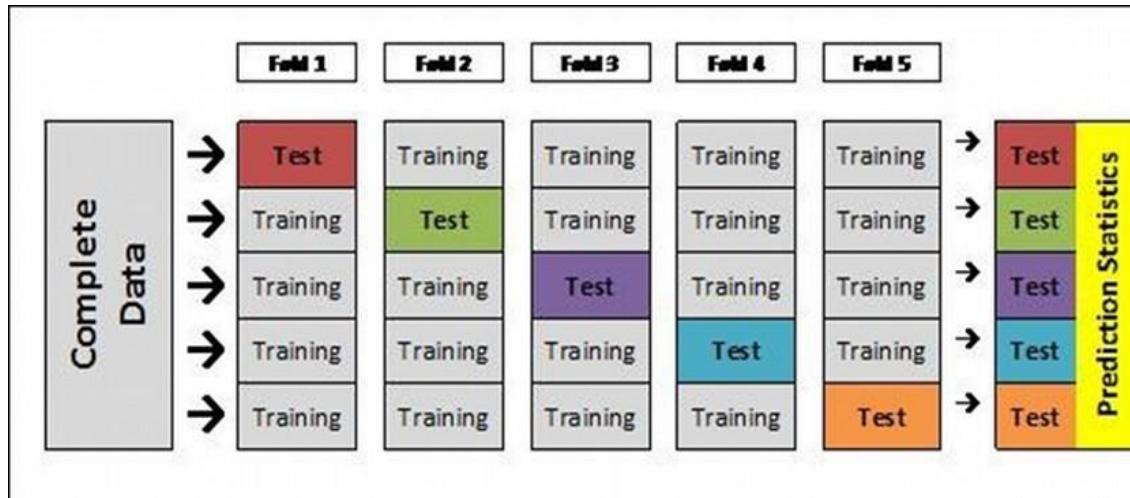
Aula 4.9. Resampling

- Introduzir o conceito de reamostragem.
- Apresentar o método hold-out.
- Apresentar o método cross-validation.
- Apresentar o método jackknife.
- Apresentar o método bootstrap.

- Testes paramétricos clássicos comparam estatísticas calculadas a partir de uma amostra à distribuições amostrais teóricas.
- Reamostragem é um conjunto de técnicas, ou métodos, que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra.
- Tipos de **reamostragem**:
 - Cross-validation.
 - Hold-out.
 - Jackknife.
 - Bootstrap.

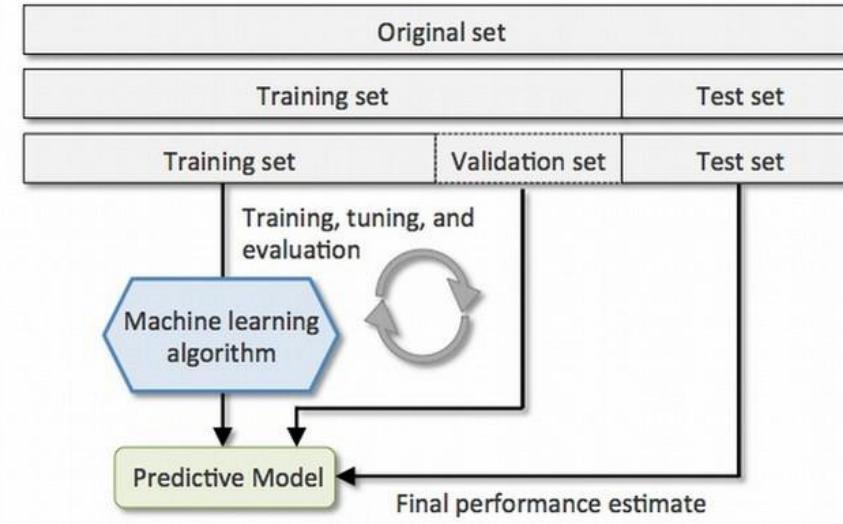
- Tipicamente, na validação cruzada, a amostra é dividida aleatoriamente em N subconjuntos disjuntos.
- Cada subconjunto possui o mesmo tamanho.
- É uma das estratégias mais utilizadas em modelagem preditiva.
- Esta análise pode ficar comprometida quando a amostra é muito pequena.

Cross-validation



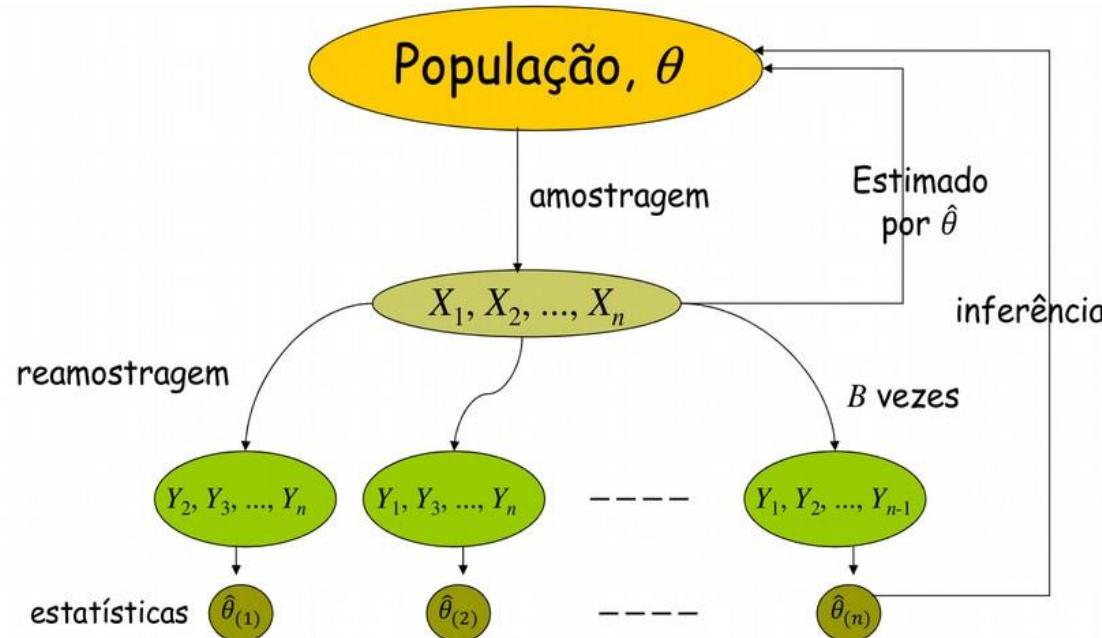
- Particiona o conjunto de dados em dois conjuntos.
- Os conjuntos são disjuntos e obtidos de forma aleatória.
- Geralmente, usa-se 2/3 dos dados para o conjunto de treino e é utilizado para a etapa de aprendizado dos modelos.
- O segundo conjunto, teste, compreende 1/3 dos dados e é utilizado para avaliação do modelo construído.

Hold-out



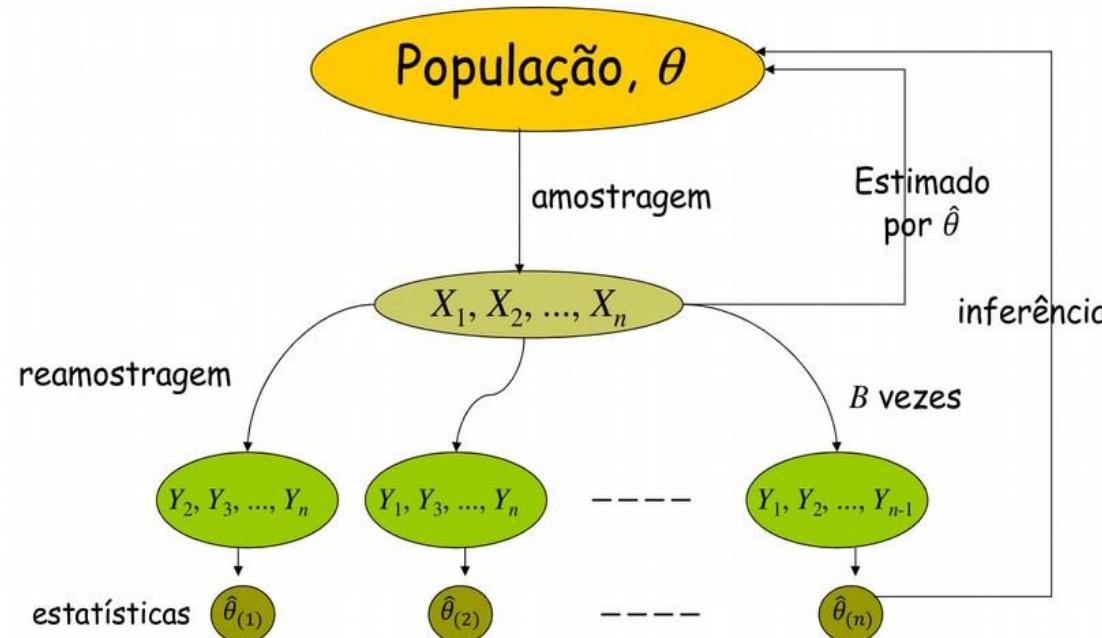
- Também chamado “leave-one-out”.
- Usado para estimar a variância e a tendência de um estimador qualquer.
- Baseia-se na remoção de uma amostra (podendo ser mais) do conjunto total observado, recalculando-se o estimador a partir dos valores restantes.
- É de fácil implementação e possui número fixo de iterações (caso se retire apenas uma amostra por vez).

Jackknife



- Pode ser considerado uma estratégia mais abrangente que o jackknife, por permitir um maior número de replicações. Também é usado para estimar a variância e a tendência de um estimador qualquer.
- Baseia-se na geração de uma nova amostra, de mesmo tamanho da amostra original, a partir do sorteio aleatório com reposição de seus elementos.

Bootstrap



Comparação dos métodos

- **Hold-out:** ideal para coleções muito grande de dados.
- **Cross-validation:** útil para coleções de tamanho intermediário e grande, em que deseja-se estimar o erro verdadeiro, mas com elevada variância
- **Jackknife:** aplicável em coleções pequenas. Método mais caro de todos, porém que traz as estimativas de erro mais precisas.
- **Bootsrap:** Método aplicável em coleções pequenas ou médias. Permite uma estimativa do erro real com uma variância mais baixa que o cross-validation.

- Introduzir o conceito de reamostragem.
- Apresentar o método hold-out.
- Apresentar o método cross-validation.
- Apresentar o método jackknife.
- Apresentar o método bootstrap.

- Métricas de qualidade em modelagem preditiva.
- Avaliação nas tarefas de predição e ranking.
- Conceitos de avaliação: acurácia, precisão e revocação.



Modelos Preditivos Séries Temporais

**Capítulo 5. Avaliação e comparação de Modelos
Preditivos**

Prof. Fernando Mourão e Prof. Túlio Vieira



Aula 5.1.1. Métricas de qualidade (Parte 1)

- Discutir a distinção entre as tarefas de predição e ranking.
- Apresentar as principais métricas de avaliação de modelos preditivos relacionados aos acertos.
- Discutir sobre os conceitos de precisão e revocação.

- Encontramos na literatura diversas métricas para avaliar modelos de predição.
- Tais métricas são podem ser agrupadas de acordo com os objetivos de análise:
 - **Medidas de erro:** MAE, MSE e RMSE.
 - **Métricas centradas em acertos:** acurácia, ROC AUC, Breese score e precision/recall.
 - **Métricas centradas em usuários:** cobertura, user retention, satisfação e reversals.

- Em muitos cenários reais, acurácia não é muito relevante na prática.
 - Exemplo: recomendação de supermercado.
- Lift, cross-sale, up-sales e ROI:
 - Um bom preditor é aquele que traz mais dinheiro!
- Novas métricas focadas não apenas no objetivo da previsão, mas na experiência do usuário.

- A seleção de métricas de avaliação depende fundamentalmente do tipo de tarefa-alvo:
 - Predição foca mais em acurácia e pouco em suporte a decisão; análise local
 - Exemplo: classificação de documentos.
 - Top-N foca mais em ranking e muito mais em suporte a decisão; análise comparativa.
 - Exemplos: recomendação de filmes.

MAE – Mean Absolute Error

- O que é erro?
 - Diferença entre a predição P e a avaliação real R (rating): $P - R$
- Absolute error remove a direção do erro:
 - $|P - R|$
- MAE = Média ($|P - R|$).

MSE – Mean Squared Error

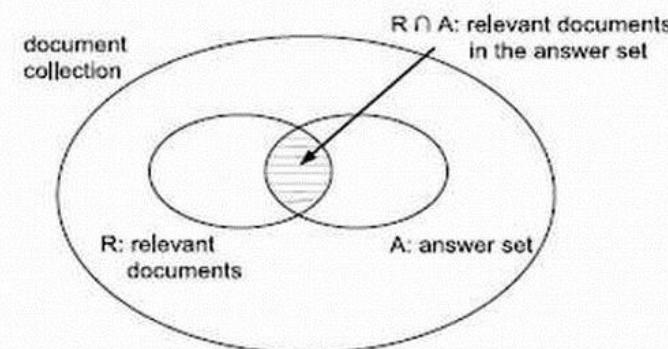
- Por que o quadrado do erro?
 - Remove o sinal (direção) – evita de usarmos módulo.
 - Penaliza mais erros maiores a erros menores.
- $\text{MSE} = \text{Média} ((P-R)^2)$
 - Uma desvantagem: quadrado do erro não é uma escala intuitiva!

RMSE – Root Mean Squared Root

- Apenas aplica raiz quadrada ao MSE.
- Coloca a métrica na mesma escala das avaliações.
- Muito mais intuitiva.
- Mesmas vantagens do MSE.

Precisão e revocação

- Considere:
 - I: uma informação requisitada.
 - R: o conjunto de objetos relevantes para I.
 - A: o conjunto resposta para I, gerada pelo modelo.
 - $R \cap A$: a interseção dos conjuntos R e A.



- Precisão e revocação são definidas como segue:
 - Revocação: fração de objetos relevantes (o conjunto R) que foram recuperados:

$$Recall = \frac{|R \cap A|}{|R|}$$

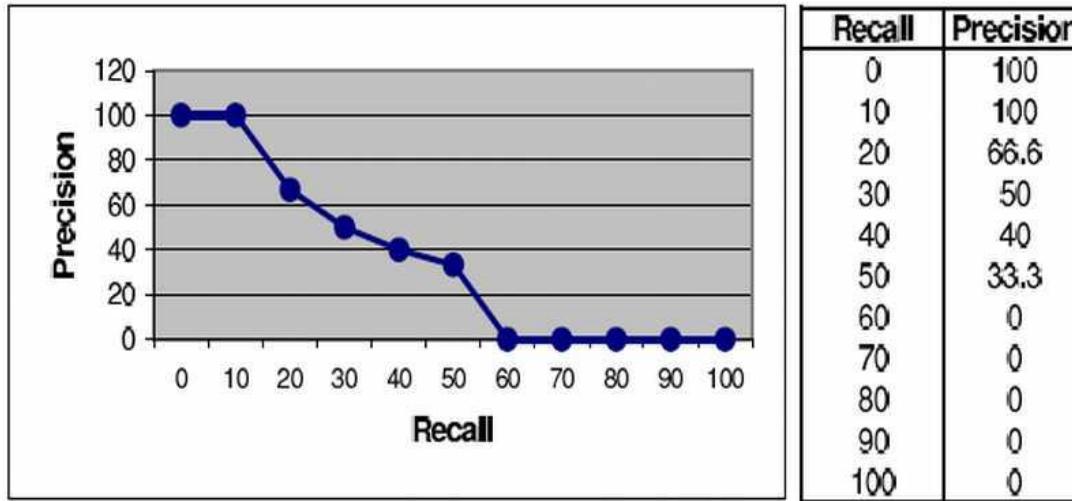
- Precisão: a fração de objetos recuperados (o conjunto A) que são relevantes:

$$Precision = \frac{|R \cap A|}{|A|}$$

Precisão e revocação

- Precisão e revocação assumem que todos os objetos de A foram examinados previamente.
- Usuários não estão interessados em todos os objetos retornados pelo modelo:
 - Avaliam apenas os top-N objetos.
 - N é um número usualmente baixo.
- Precisão e revocação devem variar a medida que o número de objetos avaliados aumenta.
- Com isso, torna-se mais apropriado apresentar curvas de precisão e revocação.

Curva de precisão vs. revocação

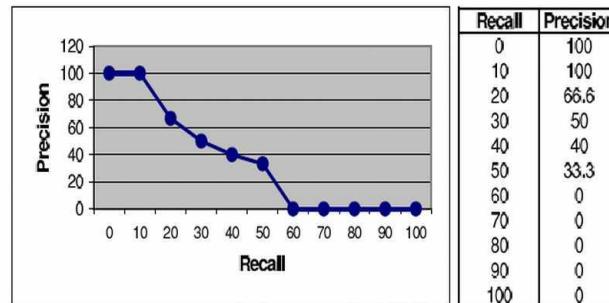


- Alta revocação nem sempre é um requisito forte.
- Quanto maior o número de objetos relevantes nos primeiros ranks, melhor a impressão.
- Precision at 5 ([P@5](#)) e Precision at 10 ([P@10](#)) medem, respectivamente, a precisão, quando temos 5 e 10 objetos retornados.
- Verificam se os usuários estão tendo acesso aos objetos relevantes no topo do rank.

- Apesar da popularidade de uso, há alguns problemas com essas métricas:
 - Estimar a revocação máxima requer um conhecimento detalhado sobre toda a coleção.
 - Em vários cenários o uso de uma única métrica seria mais intuitivo e claro.
 - Para sistemas com um requisito fraco sobre ordenação, revocação e precisão podem ser inadequados.

MAP - Mean Average Precision

- A ideia aqui é calcular a média de precisão após recuperar cada objeto relevante.
- Para ilustrar, considere a curva de precisão vs. revocação para abaixo.



$$MAP_1 = \frac{1 + 0.66 + 0.5 + 0.4 + 0.33 + 0 + 0 + 0 + 0}{10} = 0.28$$

- Métrica simples que combina precisão e revocação:

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

- Onde:
 - $r(j)$ é a revocação da j-ésima posição do rank.
 - $P(j)$ é a precisão da j-ésima posição do rank.
 - $F(j)$ é a média harmônica da j-ésima posição do rank.

- A função F assume valores entre 0 e 1.
- Ela é zero quando nenhum objeto relevante é retornado, e um quando todos os documentos retornados forem relevantes.
- Média harmônica é alta apenas quando precisão e revocação forem altas.

Conclusão

- Discutir a distinção entre as tarefas de predição e ranking.
- Apresentar as principais métricas de avaliação de modelos preditivos relacionados aos acertos.
- Discutir sobre os conceitos de precisão e revocação.

- Revisar métricas focadas em ranking.
- Revisar métricas de correlação.
- Discutir algumas considerações práticas sobre seleção e uso de métricas.



Aula 5.1.2. Métricas de qualidade (Parte 2)

- Revisar métricas de correlação.
- Revisar métricas focadas em ranking.
- Discutir algumas considerações práticas sobre seleção e uso de métricas.

- Há situações em que:
 - Não podemos diretamente mensurar a relevância das previsões.
 - Estamos mais interessados em saber quanto uma função de ranking varia em relação a uma outra.
- Nestes casos, estamos interessados em comparar a ordem relativa produzida por duas funções de ranking.
- Isso é feito através de métricas estatísticas denominadas **rank correlation metrics**.

Métricas de correlação

- Seja R_1 e R_2 .
- Uma métrica de rank correlation produz um coeficiente de correlação $C(R_1, R_2)$ com as seguintes propriedades:
 - $-1 \leq C(R_1, R_2) \leq 1$.
 - Se $C(R_1, R_2) = 1$ ambas funções produzem o mesmo ranking.
 - Se $C(R_1, R_2) = -1$ os ranking são exatamente opostos.
 - Se $C(R_1, R_2) = 0$ as funções são independentes.

Spearman correlation

- O coeficiente Spearman é a métrica de correlação mais utilizada.
- Ela se baseia na diferença entre as posições de um mesmo documento em dois rankings.
- Seja:
 - Ranks de tamanho K.
 - s_1, j a posição do objeto d_j no ranking R1.
 - s_2, j a posição do objeto d_j no ranking R2.

$$S(\mathcal{R}_1, \mathcal{R}_2) = 1 - \frac{6 \times \sum_{j=1}^K (s_{1,j} - s_{2,j})^2}{K \times (K^2 - 1)}$$

- Precisão e revocação permitem apenas medições binárias de relevância.
- Ou seja, não existe distinção entre objetos altamente relevantes e pouco relevantes.
- Essas limitações podem ser superadas adotando definições de relevâncias graduais.
- Uma forma de se definir relevâncias graduais é através da métrica DCG.

- Ao examinarmos os resultados de uma tarefa de ranking, dois aspectos devem ser considerados:
 - Um resultado bom é o que apresenta objetos altamente relevantes nas primeiras posições do rank.
 - Objetos relevantes presentes nas últimas posições do rank são menos úteis.

- Considere que os objetos de análise foram graduados em três níveis de relevância (0-3).
- Por exemplo, a relevância dos objetos para os usuários q1 e q2, são:

$$\begin{aligned} R_{q_1} &= \{ [d_3, 3], [d_5, 3], [d_9, 3], [d_{25}, 2], [d_{39}, 2], \\ &\quad [d_{44}, 2], [d_{56}, 1], [d_{71}, 1], [d_{89}, 1], [d_{123}, 1] \} \\ R_{q_2} &= \{ [d_3, 3], [d_{56}, 2], [d_{129}, 1] \} \end{aligned}$$

- Ou seja, enquanto o objeto d3 é muito relevante para o usuário q1, o objeto d56 é apenas relevante para o usuário q2.

- Primeiro, especialistas associam um nível de relevância a cada objeto dos top-N retornados pelo algoritmo.
 - A lista de scores de relevância é definida como vetor de ganho (G).
- Por exemplo, vetores de ganhos para dois usuários:

$$G_1 = (1, 0, 1, 0, 0, 3, 0, 0, 0, 2, 0, 0, 0, 0, 3)$$

$$G_2 = (0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 3)$$

DCG – Discounted Cumulated Gain

- Somando-se os ganhos até um determinado ponto no ranking, temos o ganho acumulado (CG).
- Por exemplo, os vetores de ganho acumulado do exemplo anterior:

$$CG_1 = (1, 1, 2, 2, 2, 5, 5, 5, 5, 7, 7, 7, 7, 7, 10)$$

$$CG_2 = (0, 0, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 6)$$

- Ou seja, o ganho acumulado na posição 8 de CG1 é 5.

DCG – Discounted Cumulated Gain

- Dado um vetor de ganho G_j para um teste q_j , o vetor DCG j correspondente é dado por:

$$DCG_j[i] = \begin{cases} G_j[1] & \text{if } i=1; \\ \frac{G_j[i]}{\log_2 i} + DCG_j[i-1] & \text{otherwise} \end{cases}$$

- Em geral, inclui-se também um fator de desconto para atenuar o ganho acumulado a medida que andamos no ranking.
- Um fator de desconto muito utilizado é a função log2.

Considerações práticas

- Medidas de erro se baseiam em summarizações:
 - Modelo padrão: média sobre todas as avaliações.
 - Modelo alternativo: média sobre os usuários ou sobre os itens.
- E qual a diferença?
 - O que acontece se um usuário tem 3000 avaliações e outro apenas 10?
- De maneira prática é bom comparar ambos modelos.

Considerações práticas

- O que fazer quando calculamos MAE, considerando diferentes algoritmos?
 - Lembre-se: devemos considerar o mesmo conjunto de dados sempre!
Se a cobertura dos algoritmos for diferente?
 - Calcule o MAE considerando o conjunto em comum.
 - Defina previsões default, forçando a produzir previsões para todos os itens.

Considerações práticas

- Grande parte das medidas de erro são equivalents.
- Erros quadrados podem ser mais apropriados para domínios com escalas de avaliação maiores, permitindo capturar erros de maior magnitude.
- Grande parte dos trabalhos usam MAE.
 - Facilita a comparação com o que já foi feito.
- A desvantagem deste tipo de avaliação é que erros podem ser dominados por itens irrelevantes.

Conclusão

- Revisar métricas focadas em ranking.
- Revisar métricas de correlação.
- Discutir algumas considerações práticas sobre seleção e uso de métricas.

- ❑ Como realizar a comparação de modelos?
- ❑ Revisaremos os principais conceitos estatísticos relacionados a esta tarefa.
- ❑ Discutiremos os principais desafios relacionados.

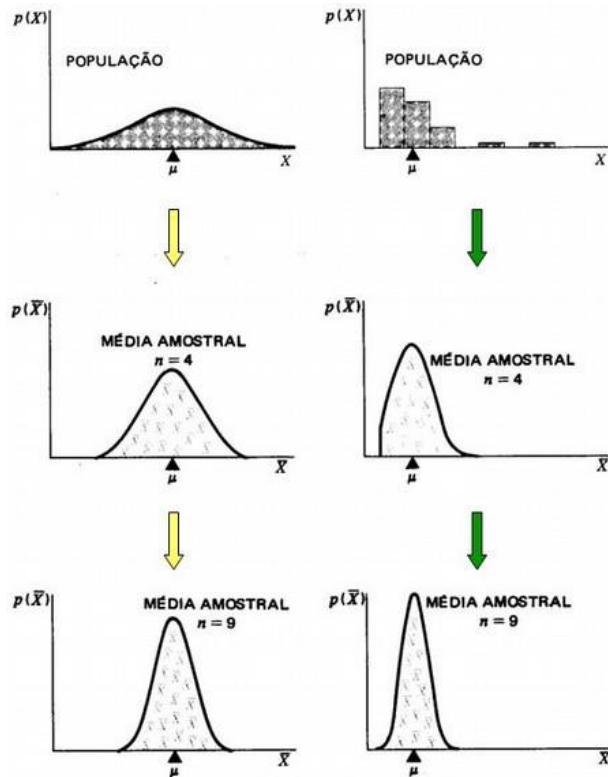


Aula 5.2. Comparação de modelos

Nesta aula

- ❑ Projetar experimentos para comparação entre modelos.
- ❑ Revisar conceitos estatísticos relacionados à avaliação comparativa.
- ❑ Discutir algumas considerações práticas sobre a comparação de modelos.

Teorema do limite central



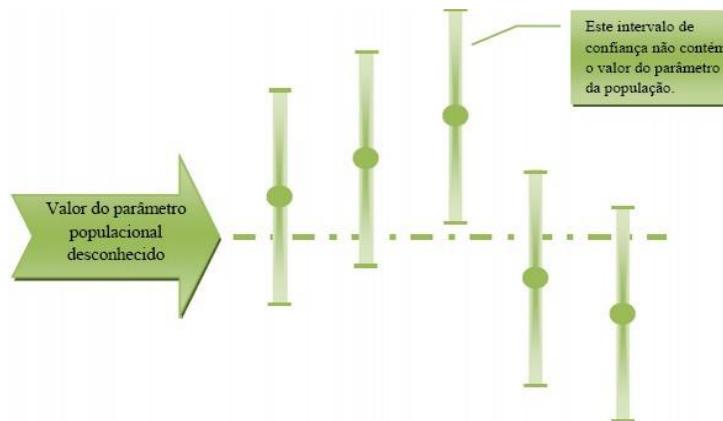
Intervalo de confiança

- Intervalo de confiança (IC) é um tipo de estimativa por intervalo de um parâmetro populacional desconhecido.
- Tomando-se qualquer amostra particular, o parâmetro populacional desconhecido pode ou não pode estar no intervalo de confiança observado.
- Quando tem-se um IC com 95% de confiança, significa que em 95% das vezes que amostrarmos os dados e mensurar o parâmetro de interesse em tais amostras, o valor encontrado estará dentro do intervalo.



Intervalo de confiança

- Intervalos de confiança são usados para indicar a confiabilidade de uma estimativa.



Intervalo de confiança

- Duas fórmulas para intervalo de confiança:
 - Acima de 30 amostras de qualquer distribuição: distribuição-z (normal).
 - Pequenas amostras de populações normalmente distribuídas: distribuição-t (Student).
- Um erro comum é usar distribuição-t para populações normalmente não distribuídas.

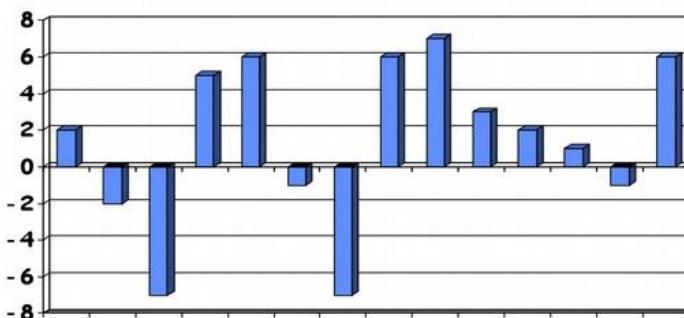
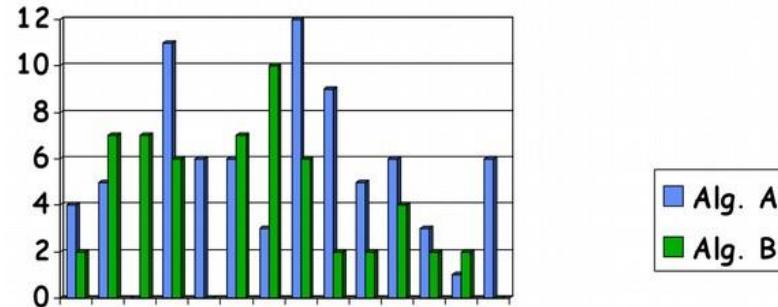
Comparando modelos

- Ao se propor um novo modelo é importante compará-lo com outro já existente, a fim de se mensurar o impacto das melhorias propostas.
- Este impacto pode ser mensurado de diversas formas:
 - Determinar qual modelo executa mais rápido.
 - Determinar qual modelo produz mais acertos.
 - Determinar qual modelo possui melhor QoSxPreço – Quality of Service.
- Precisamos de estratégias de comparação distintas quando temos observações **pareadas** e **não pareadas**.

Observações pareadas

- Cenário em que o n-ésimo teste, em cada modelo, foi o mesmo.
- Estratégia de comparação:
 - Tratar o problema como uma única amostra de n pares.
 - Para cada par: calcule a diferença dos resultados.
 - Calcule o intervalo de confiança para a diferença média.
 - Se o intervalo inclui 0, os modelos não são diferentes com a confiança utilizada.
 - Se o intervalo não inclui 0, o sinal da diferença indica qual modelo é melhor.

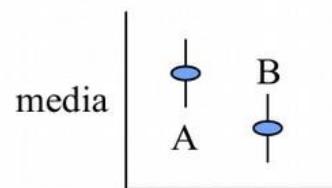
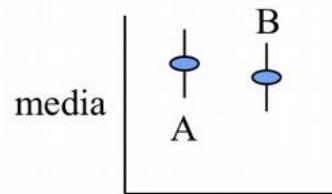
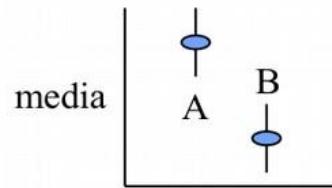
Observações pareadas



Observações não pareadas

- Cenário em que os testes são diferentes.
- Estratégia de comparação:
 - Calcule as médias das amostras para cada uma das alternativas.
 - Calcule o intervalo de confiança para cada alternativa.
 - Compare os intervalos de confiança:
 - Se não houver sobreposição: algoritmos são diferentes e a maior média é melhor.
 - Se houver sobreposição e cada IC contém a outra média: algoritmos não são diferentes neste nível.
 - Se houver sobreposição e uma média não está no outro IC: tem de fazer o teste-t.

Observações não pareadas



Testes de hipótese

- Método para testar uma reivindicação ou hipótese sobre um parâmetro em uma população, usando dados medidos em uma amostra:
 - Pode ser confuso em negativas duplas.
 - Provê menos informação que intervalos de confiança.
 - Em geral, é mais difícil de interpretar/entender.
- Começamos assumindo que a hipótese nula é verdadeira.
- O objetivo é provar que ela é falsa.

- O método possui quatro passos principais:
 - **Indique as hipóteses** (hipótese nula e hipótese alternativa);
 - **Defina os critérios para uma decisão** (nível de significância);
 - **Calcule a estatística de teste** (teste estatístico);
 - **Tome uma decisão** (p-value).

Conclusão

- Projetar experimentos para comparação entre modelos.
- Revisar conceitos estatísticos relacionados à avaliação comparativa.
- Discutir algumas considerações práticas sobre a comparação de modelos.

- Projeto de experimentos off-line.
- Desafios da experimentação off-line.
- Limitações relacionadas a este tipo de avaliação.



Aula 5.3. Avaliação off-line

- Experimentos off-line vs. on-line.
- Projeto de experimentação off-line.
- Análise de fatores.

Experimentos off-line vs. on-line

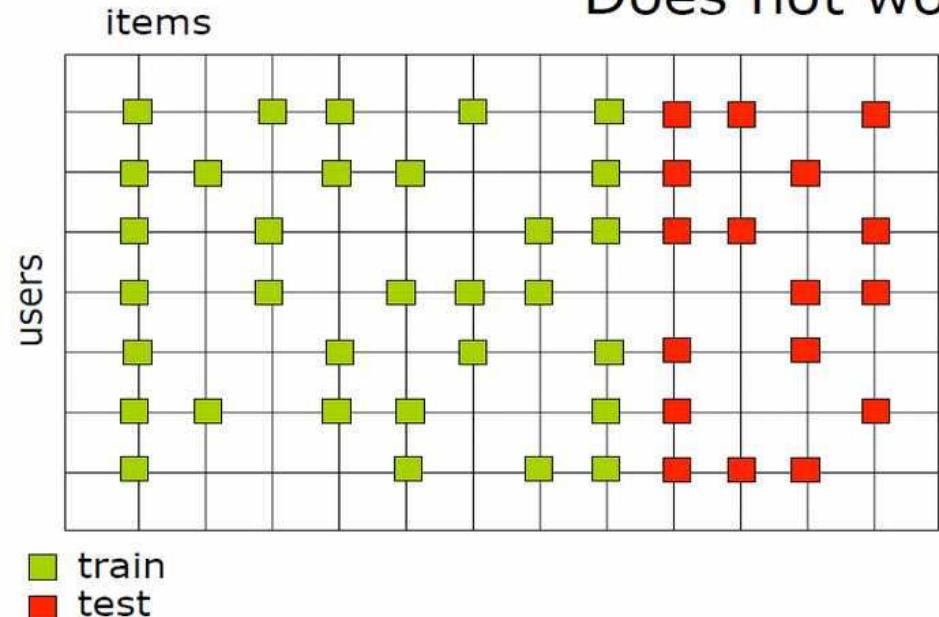
- Avaliação **retrospectiva** (off-line) olha apenas como o modelo gerou previsões para itens já consumidos e avaliados.
- Avaliação **prospectiva** (on-line) olha efetivamente como previsões são recebidas pelos usuários.
- Diferença fundamental:
 - Off-line: abundância de dados passados, mas não pode responder a principal pergunta:
 - Um modelo é bom para identificar novos itens?
 - On-line: avaliação mais completa, mas pode ser complexa e pode incomodar os usuários.

- Um pouco de intuição:
 - Medidas de erro/acurácia são usualmente feitas através de uma metodologia 'leave one out'.
 - 'Esconde' uma avaliação e tenta prever seu valor.
 - Muitas vezes este processo é muito caro e complicado. Então adota-se uma simplificação:
 - Esconde '10%' da base e tenta-se prever cada valor escondido.
 - Projeto denominado de teste/treino.

- Divide-se os dados em treino/teste (via hold-out ou cross-validation).
- Gere as previsões usando apenas o conjunto de treinamento.
- Compare as previsões:
 - Para cada previsão, verifique se você acertou ou errou, ou o quanto distante a previsão está do valor real.

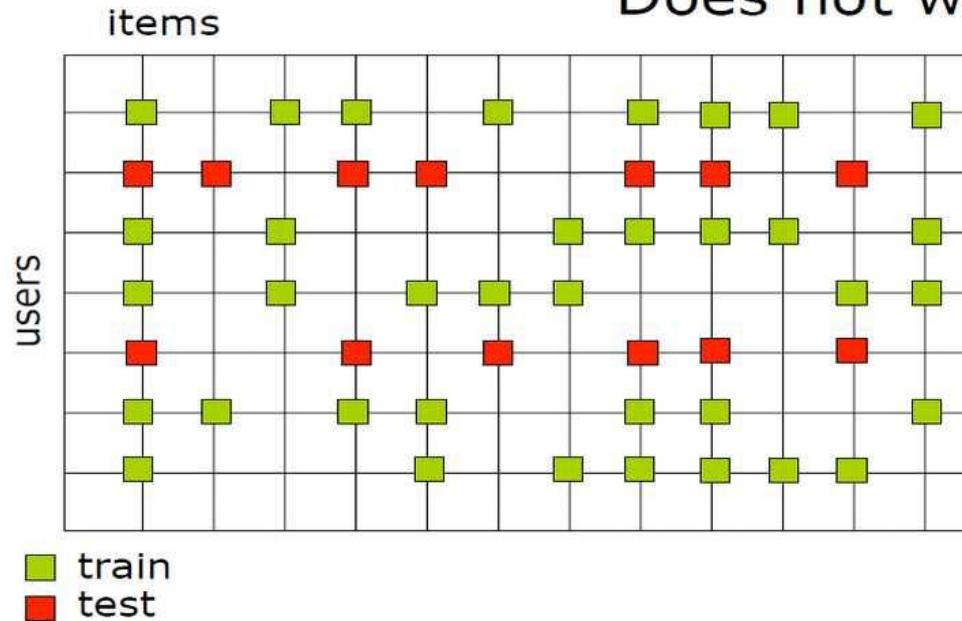
Dividindo os dados em teste/treino

Does not work

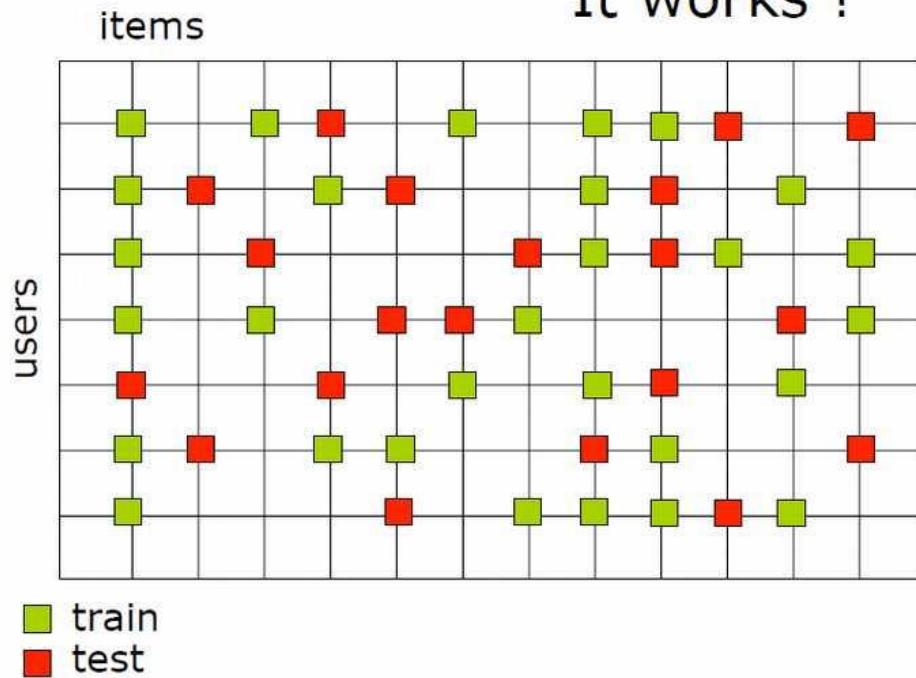


Dividindo os dados em teste/treino

Does not work



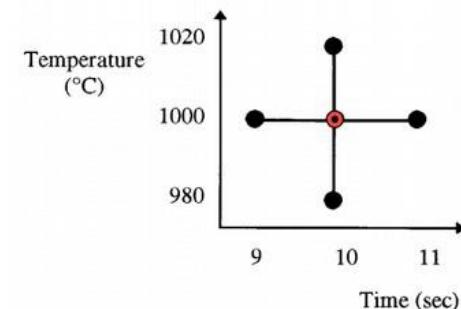
It works !



- Muitas vezes gostaríamos de entender quais fatores são responsáveis por um modelo ser melhor do que o outro.
- Principais estratégias:
 - Educated guesses.
 - Um fator por vez (projeto simples).
 - Fatorial completo.
 - Fatorial fracionado.
 - Fatorial 2K (e variantes).
 - Quasi-experimentos.
 - Field experiment.

Projeto simples

- Principais passos:
 - Selecione um ponto de referência.
 - Altere um fator individualmente e mantenha os demais constantes.
- Amplamente utilizado na prática (principalmente em pesquisas sobre algoritmos).
- Provê bons resultados quando não há interação entre os fatores (independência).



- k fatores distintos.
- Cada fator i , possui n_i níveis.
- r replicações.
- n – número de experimentação.

$$\begin{array}{c} \text{Fator 1} & & \text{Fator 2} \\ (\ell_{1,0}, \ell_{1,1}, \dots, \ell_{1,n_1-1}) \times (\ell_{2,0}, \ell_{2,1}, \dots, \ell_{2,n_2-1}) \times \dots \\ \times (\ell_{k,0}, \ell_{k,1}, \dots, \ell_{k,n_k-1}) & \text{Fator } k \end{array}$$

- Assume que os fatores não interagem.
- Usualmente, requer mais esforço que se pensa.
- Tente evitar esse enfoque de experimentação.

Conclusão

- Experimentos off-line vs. on-line.
- Projeto de experimentação off-line.
- Análise de fatores.

■ Próxima aula

- Projeto de experimentos on-line.
- Teste A/B.
- Interleaving.



Aula 5.4. Avaliação on-line

- Discutir as principais características da experimentação centrada no usuário.
- Planejamento de testes com usuário.
- Elaboração de surveys/questionários.
- Teste A/B.
- Interleaving.

Avaliação user-centered

- Muitas vezes, o novo alvo de experimentação são os usuários, ou sistemas que interagem diretamente com usuários.
- Nestes casos, os produtos, tanto quanto possível, devem ser projetados para satisfazer os usuários.
- Assim, para algumas organizações, experimentos randomizados desempenham um importante papel no projeto destes produtos e na tomada de decisão.

Avaliação user-centered

- Experimentos podem ser usados para:
 - Explorar diversas opções de projeto.
 - Entender como usuários reagem a mudanças.
- Empresas estão cientes da importância da experimentação, neste caso.
- Especialmente empresas que oferecem produtos e serviços na web, conduzem, frequentemente, experimentos com usuários.

- Não pode ser repetida, de forma trivial.
- Deve-se ter um cuidado maior para que o usuário comprehenda o que está sendo avaliado.
- Não escala tanto quanto experimentos centrados em dados.
- Envolve, muitas vezes, questões mais qualitativas.
- As perguntas a serem feitas são completamente diferentes das perguntas de experimentação centradas em dados.

- Quando projetamos experimentos centrados nos usuários, os dados vêm, essencialmente, de observações e entrevistas.
- Observações podem ser conduzidas em três momentos distintos no processo:
 - Antes do projeto do produto;
 - Durante o projeto e a prototipação da aplicação;
 - Quando o produto já está pronto.

Tipos de experimentos

- Há diversos tipos distintos de experimentos que podemos conduzir centrado nos usuários.
- Experimento **presencial** vs. **on-line**.
- Focaremos em três tipos principais:
 - Surveys/Questionários.
 - Teste A/B.
 - Interleaving.

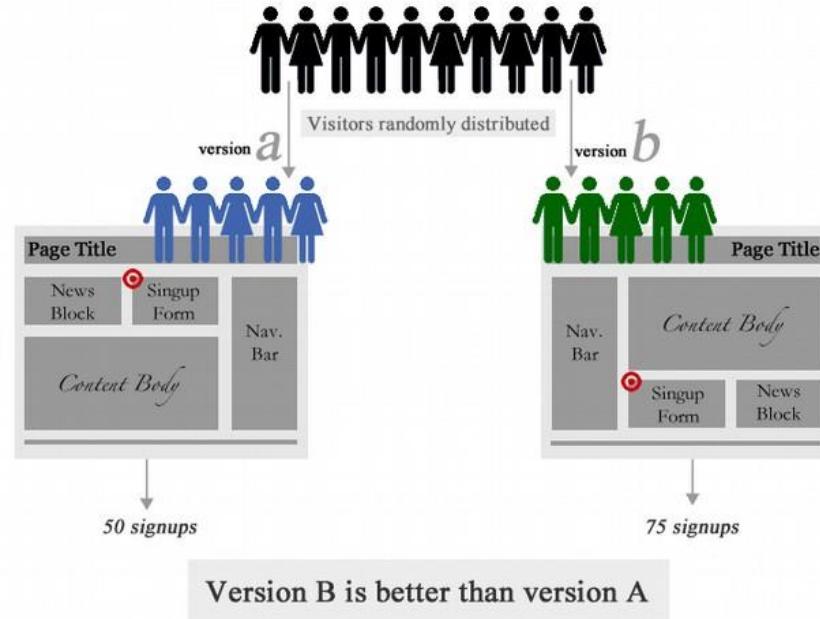
- Estudo retrospectivo de uma situação para documentar relacionamentos e resultados.
- Sempre feito depois do evento.
- Não há controle sobre a situação.
- Não há manipulação de variáveis.
- Feito com entrevistas e questionários:
 - Os dados vêm da memória dos entrevistados.

Planejamento de um survey

- Escolha de objetivos específicos e mensuráveis.
- Planejamento e escalonamento do survey.
- Obtenção dos recursos necessários.
- Design do survey.
- Preparação do instrumento de coleta de dados.
- Validação do instrument.
- Seleção de participantes.
- Execução.
- Análise de dados.
- Relatar os resultados.

- Você tem duas versões de um elemento (A e B) e uma métrica que define o sucesso.
- Para determinar qual versão é melhor, você avalia ambas as versões simultaneamente.
- Seleciona-se randomicamente usuários para usar cada versão.
- No final, mede-se qual versão foi mais bem sucedida e seleciona-se essa versão para o uso no mundo real.

Teste A/B

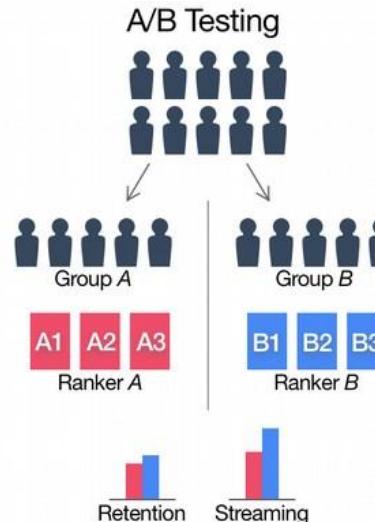


- Saiba quanto tempo para executar um teste antes de desistir.
- Mostre a mesma versão a visitantes que retornam.
- Torne o teste A/B consistente em todo o site.
- Faça vários testes A/B.

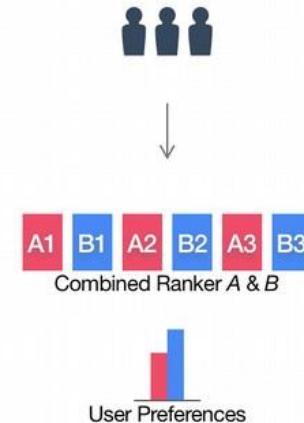
- Variação do teste A/B, em que cria-se um terceiro grupo que utiliza o mesmo algoritmo de A.
- Todos os três grupos possuem o mesmo número de usuários e são comparáveis.
- O objetivo é garantir que o teste A/B esteja configurado e executado de maneira adequada.
- O comportamento esperado é que A e C tenham o mesmo resultado, já que usam o mesmo algoritmo.
- A diferença entre A e B dá o intervalo de erro, bem como o tempo necessário para convergência do teste A/B.

- Interleaving é uma técnica poderosa que nos permite acelerar a avaliação de algoritmos de ranking.
- Consiste em apresentar, simultaneamente, para cada usuário, resultados de dois algoritmos de ranking, sem que o usuário perceba.
- Comece aleatoriamente com o ranking A, ou o ranking B, para minimizar o viés de apresentação.
- Conte o número de clicks nos resultados de A, versus resultados de B.
- O melhor ranking obterá (em média) mais clicks.

Interleaving



Interleaving



- Discutir as principais características da experimentação centrada no usuário.
- Planejamento de testes com usuário.
- Elaboração de surveys/questionários.
- Teste A/B.
- Interleaving.

■ Próxima aula

- Discussão de cenários reais.
- Revisão dos principais passos da modelagem preditiva.
- Desafios e decisões práticas.



Modelos Preditivos Séries Temporais

Capítulo 6. Aplicações de Modelos Preditivos

Prof. Fernando Mourão e Prof. Túlio Vieira



Aula 6.1.1. Aplicação prática de MPE (Parte 1)

- ❑ Cenário 1 - Riscos de incêndio.

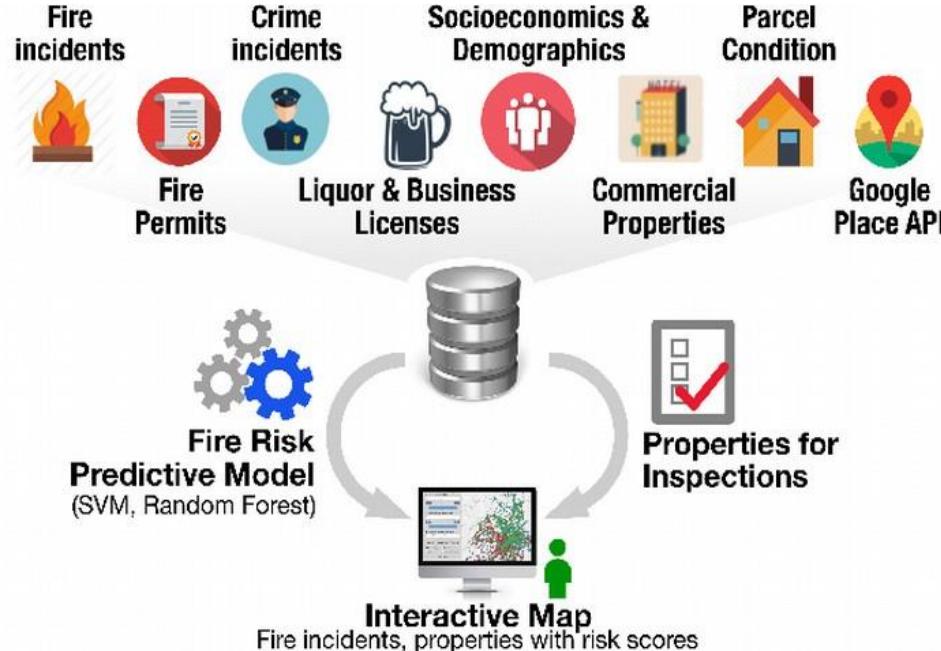
- Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta.
- Best Paper KDD Conference 2016.
- Trabalho concluído por grandes universidades americanas, em parceria com o Departamento de Resgate ao Incêndio da Cidade de Atlanta.

Todas as figuras aqui apresentadas foram extraídas do artigo original:

MADAIO, Michael et al. "Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.

- Incêndio em propriedades comerciais traz sérios riscos à vida de um grande número de pessoas.
- Em grandes cidades, há um número muito maior de propriedades a serem inspecionadas que a capacidade das autoridades responsáveis é capaz de efetivamente inspecionar.
- **Como selecionar quais as propriedades a serem inspecionadas, visando minimizar o risco de incêndio em toda a cidade?**

- Um novo framework (Firebird) para identificação e priorização de propriedades com maior risco de incêndio.
- Data-driven: baseada em critérios e históricos do Departamento de Resgate do Incêndio.
- Aborda dois desafios separadamente:
 1. Identificação de propriedades: **tarefa de Top-N**.
 2. predição de risco de incêndio: **tarefa de ranking**.



Consolidação dos dados



Consolidação dos dados

Source	Name	Description
Atlanta Fire Rescue Department	Fire Incidents	Fire incidents from 2011 - 2015
	Fire Permits	All permits filed by AFRD in 2012-2015
City of Atlanta	Parcel	Basic information for each parcel in Atlanta
	Strategic Community Investigation	Information regarding parcel conditions
	Business Licenses	All the business licenses issued in Atlanta
Atlanta Police Department	Crime	2014 crime in Atlanta
	Liquor Licenses	All filed liquor licenses by Police Department
Atlanta Regional Commission	Neighborhood Planning Unit	Boundary data for each Atlanta neighborhood
U.S. Census Bureau	Demographic	Household number, population by race and age
	Socioeconomic	Household median income
CoStar Group, Inc	CoStar Properties	Commercial property information
Google Place APIs	Google Place	Information regarding places from Google Maps

Feature selection

- Manualmente avaliou-se 252 variáveis distintas.
- Posteriormente, aplicou-se um processo de feature selection para se determinar a relevância de cada variável.

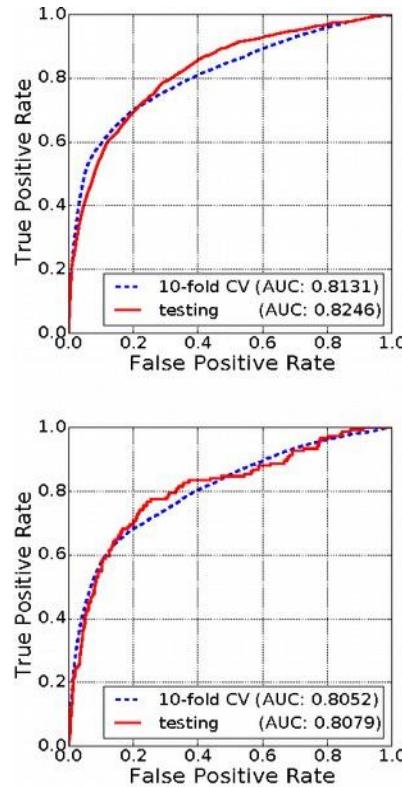
Top 10 features	
1	floor size
2	land area
3	number of units
4	appraised value
5	number of buildings
6	total taxes
7	property type is multi-family
8	lot size
9	number of living units
10	percent leased

Modelos de predição

- Testaram diversos métodos de predição:
 - Logistic Regression.
 - Gradient boosting.
 - Support Vector Machine.
 - Random forest.
- Métodos implementados pelo pacote scikit-learn de python.
- Métodos com melhor desempenho: SVM e random forest.

- Particionamento temporal dos dados.
- Seleção dos parâmetros via grid search com 10-fold cross validation.
- Métricas de avaliação:
 - True Positive Rate (TPR).
 - False Positive Rate (FPR).
 - Curvas ROC.
 - Area Under the Curve (AUC).

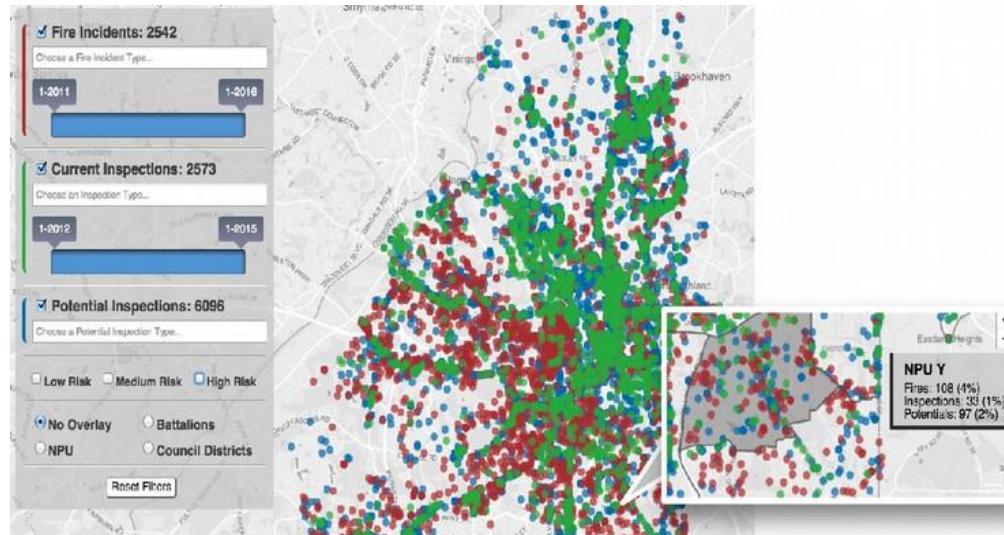
Resultados



(a) SVM

Training window	Testing AUC of the following year	
	Random Forest	SVM
2011-2012	0.7624	0.7614
2011-2013	0.8030	0.7914
2011-2014	0.8246	0.8079

Uso práctico



Conclusão

Cenário 1 - Riscos de incêndio.

■ Próxima aula

- Discussão de cenários reais.
- Revisão dos principais passos da modelagem preditiva.
- Desafios e decisões práticas.



Aula 6.1.2. Aplicação prática de MPE (Parte 2)

- ❑ Cenário 2 - Recomendação de eventos.

- Context-Aware Event Recommendation in Event-based Social Networks.
- Best Paper RecSys Conference 2015.
- Trabalho concluído por duas grandes universidades brasileiras:
UFCG e UFMG.

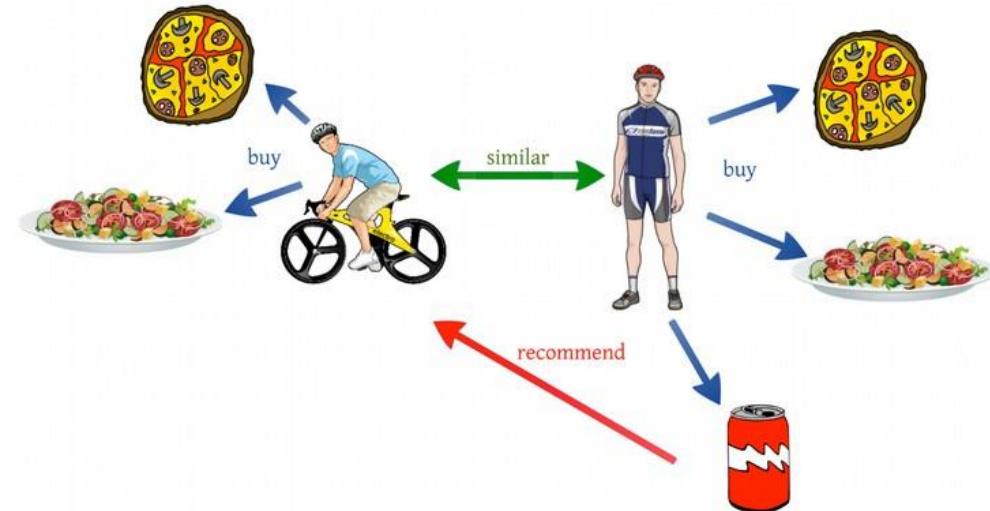
Todas as figuras aqui apresentadas foram extraídas do artigo original:
MACEDO, Augusto Q.; MARINHO, Leandro B; SANTOS, Rodrygo L. T. "Context-aware event recommendation in event-based social networks". Proceedings of the 9th ACM Conference on Recommender Systems. ACM, 2015.

- Através de redes sociais, somos notificados sobre diversos eventos sociais que ocorrem a nossa volta.
- Há um grande número e diversidade de eventos capazes de agradar aos mais distintos gostos.
- Tal diversidade, porém, dificulta a tomada de decisão por parte dos usuários.
- Como apresentar a cada usuário, apenas os eventos que ele gostaria de ir?

- Um sistema de recomendações personalizadas híbrido, que explora três tipos de informações contextuais:
 - Grupos de redes sociais que cada usuário participa;
 - Informações de localidade do usuário;
 - Informações temporais sobre as preferências dos usuários.
- O grande desafio neste caso é propor um sistema, capaz de prever preferências sobre eventos futuros, de curta duração, para o qual possuímos pouca ou nenhuma informação passada.

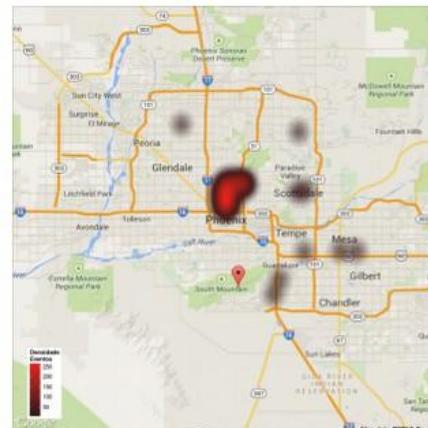
Consolidação dos dados

- Método Content-based explora a descrição textual dos eventos, de forma a identificar eventos similares.
- Adotou-se o modelo bag-of-words para representar as descrições dos eventos.

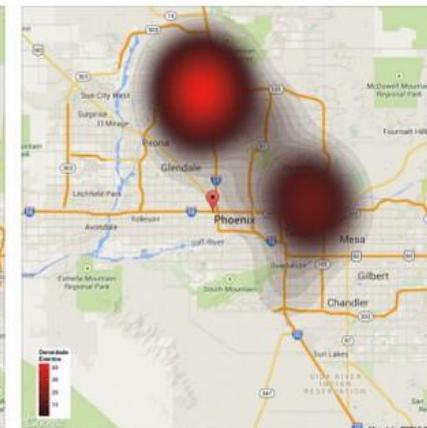


Consolidação dos dados

- Modela-se o padrão de mobilidade de cada usuário.
- Os autores propõem um novo método para estimar a densidade geográfica baseada em Kernels.



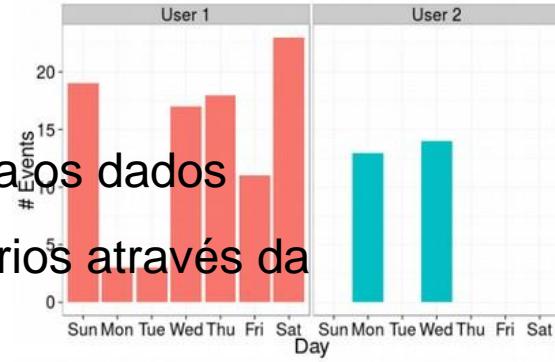
(a) User 1



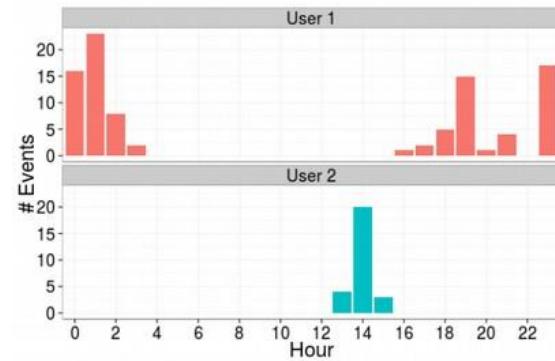
(b) User 2

Consolidação dos dados

- Modela-se o padrão temporal de cada usuário.
- Os autores propõem uma representação vetorial para os dados temporais, e calculam a similaridade entre dois usuários através da distância cosseno entre os vetores.



(a) Distribution per day.



(b) Distribution per hour.

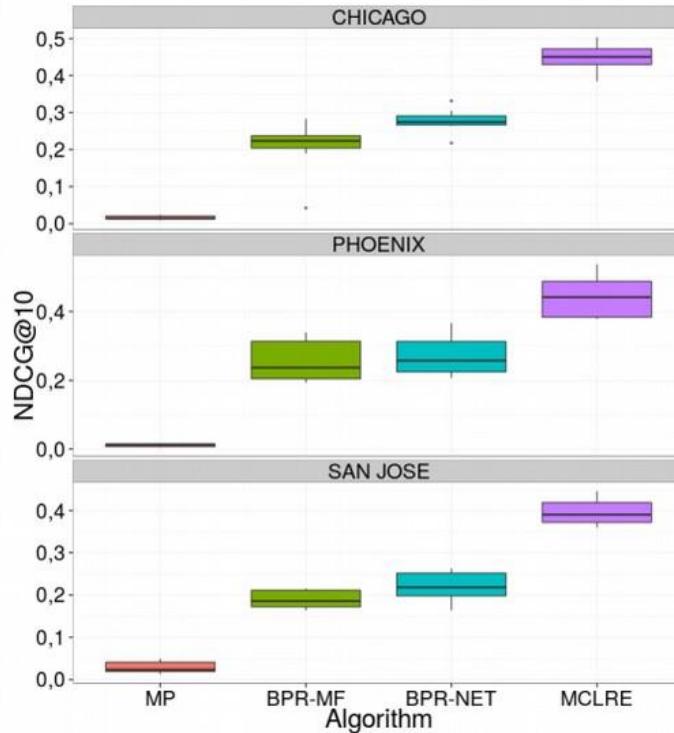
- Sistema de recomendação híbrido:
 - Combina informações de conteúdo, grupos sociais, mobilidade espacial e comportamento temporal.
- Modela o problema como uma tarefa de learning to rank:
 - Cada informações é entendida como uma feature de entrada passada para o algoritmo de learning to rank.
- Adota o método Coordinate Ascent listwise learning to rank.

- Coletaram dados de uma rede social de eventos, chamada de Meetup.com.
- Focaram em eventos que ocorreram nas cidades americanas de Phoenix, Chicago e San Jose.
- Estratégia de coleta: snowball a partir de grupos de interesses.
- Coleção altamente esparsa.

Avaliação

- Particionamento temporal dos dados baseado em janelas deslizantes.
- Métrica: Normalized Discounted Cumulative Gain (NDCG) sobre as top-10 recomendações.
- Baselines:
 - Most popular events.
 - Bayesian Personalized Ranking (BPR-MF).
 - BPR-NET: BPR-MF enriquecido com informações de redes sociais.
- Os parâmetros usados para todos os modelos, foram definidos usando-se grid search sobre conjuntos de validação.

Resultados



Cenário 2 - Recomendação de eventos.

■ Próxima aula

- Discussão de cenários reais.
- Revisão dos principais passos da modelagem preditiva.
- Desafios e decisões práticas.



Aula 6.1.3. Aplicação prática de MPE (Parte 3)

- ❑ Cenário 3 – Mudanças climáticas.

- Big Data in Climate: Opportunities and Challenges for Machine Learning and Data Mining.
- Talk feito por Vipin Kumar, em diversas conferências relacionadas à modelagem preditiva.
- Conjunto de esforços feitos pela Universidade de Minnesota e colaboradores, com o intuito de se modelar e entender mudanças climáticas.

Todas as figuras aqui apresentadas foram extraídas do talk original:

KUMAR, Vipin. “Applied Data Science Track of KDD-2017: The 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining - Big Data in Climate: Opportunities and Challenges for Machine Learning”.

- “Climate change research is now ‘big science,’ comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics.”
- É possível usar dados históricos para responder importantes questões relacionadas ao clima?
 - Monitoramento de redemoinhos oceânicos.
 - Queda de precipitações em escala global.
 - Modelagem de ondas de calor.
 - Predição de atividades de furacões.

- Dados de multiresolução e multiescala.
- Alta variabilidade temporal.
- Autocorrelação espaço-temporal.
- Heterogeneidade espacial e temporal.
- Grande quantidade de ruído e valores ausentes.
- Falta de valores verdadeiros e representativos.
- Desbalanceamento de classe (as mudanças são eventos raros).

1. Global mapping of forest fires:

- ❑ RAPT: Rare Class Prediction in Absence of Ground Truth (TKDE 2017)



2. Mapping of plantation dynamics in tropical forests:

- ❑ Recurrent Neural Networks to model space and time (IEEE Big Data 2016, SDM 2016, KDD 2017)



3. Global mapping of inland surface water dynamics

- ❑ Heterogeneous Ensemble Learning (SDM 2015, ICDM 2015)
- ❑ Physics-guided Labeling (ICDM 2015, RSE 2017)
- ❑ Information Transfer across Space and Time



Modelo preditivo

- Modelagem preditiva para uma classe de alvo rara, usando rótulos imperfeitos.
- O que são rótulos imperfeitos:
 - Classes ruidosas.
 - Baratas de se obter.
 - Disponíveis para todas as instâncias de teste.
- Exemplos:

Aplicação

Área de incêndio

Recomendação

Classe alvo

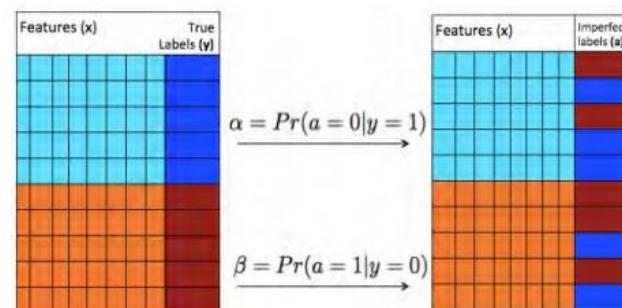
Fogo/Não

Interesse/Não

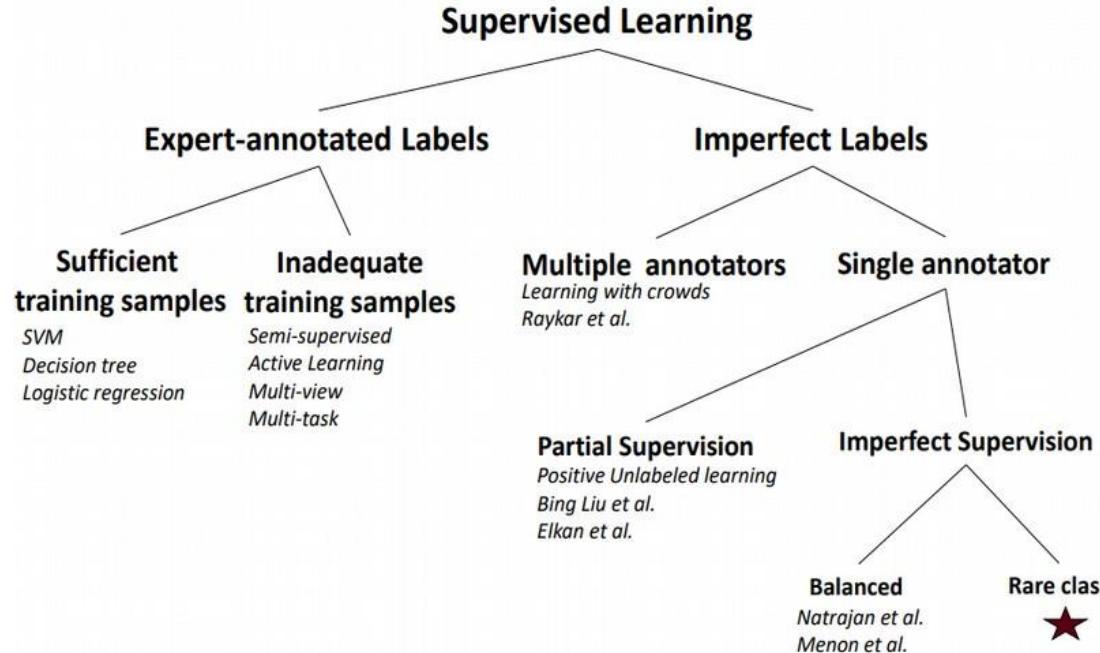
Rótulos imperfeitos

Anomalia termal

Interesse dos amigos



Aprendizado com rótulos imperfeitos

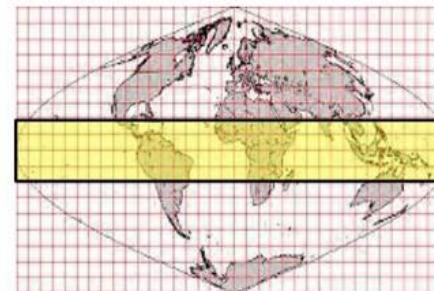
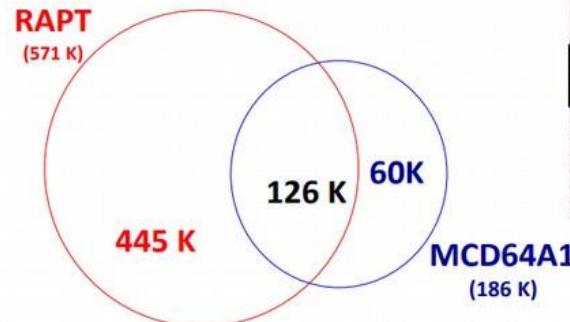


Aprendizado com rótulos imperfeitos

- Passo 1 - Aprende modelos de classificação usando rótulos imperfeitos.
 - Os autores propõem um novo método que otimiza precisão * revocação, usando rótulos imperfeitos.
- Passo 2 - Combine previsões do modelo de classificação com rótulos imperfeitos, para melhorar a precisão com alguma perda em revocação.
 - Para cenários de classe ultra-raros, o ganho de precisão após a agregação é significativamente maior em relação à perda de revocação.
- Passo 3 - Use o contexto espacial para melhorar a revocação.
 - O método explora a estrutura do relacionamento, como a vizinhança espacial, rede biológica ou rede social, para melhorar a cobertura (revocação) das instâncias da classe rara.

Aplicação – Queimadas em florestas tropicais

- O método proposto identificou, que entre 2001 e 2014, foram queimados 571 mil km² de florestas tropicais.
- O método estado da arte foi capaz de identificar apenas 186 mil km².



Conclusão

Cenário 3 – Mudanças climáticas.

- ❑ Discussão sobre melhores práticas em modelagem preditiva.
- ❑ Documentação do ciclo de projeto.
- ❑ Fases de implantação e monitoramento.

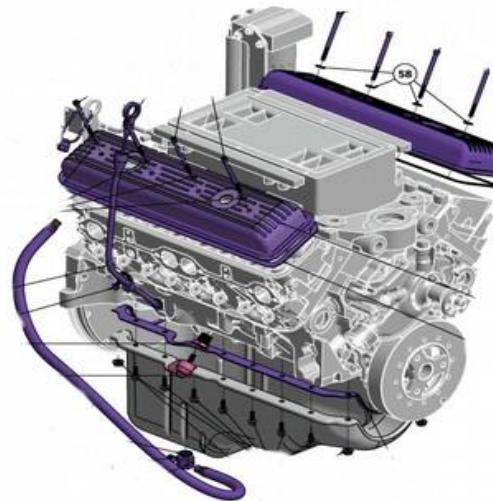


Aula 6.2.1. Melhores práticas (Parte 1)

- ❑ Melhores práticas de modelos preditivos.

Modelagem preditiva na indústria

- Na indústria, modelagem preditiva é, antes de tudo, um problema de engenharia: substituir peças de um motor sem fazê-lo parar.



Modelagem preditiva na indústria

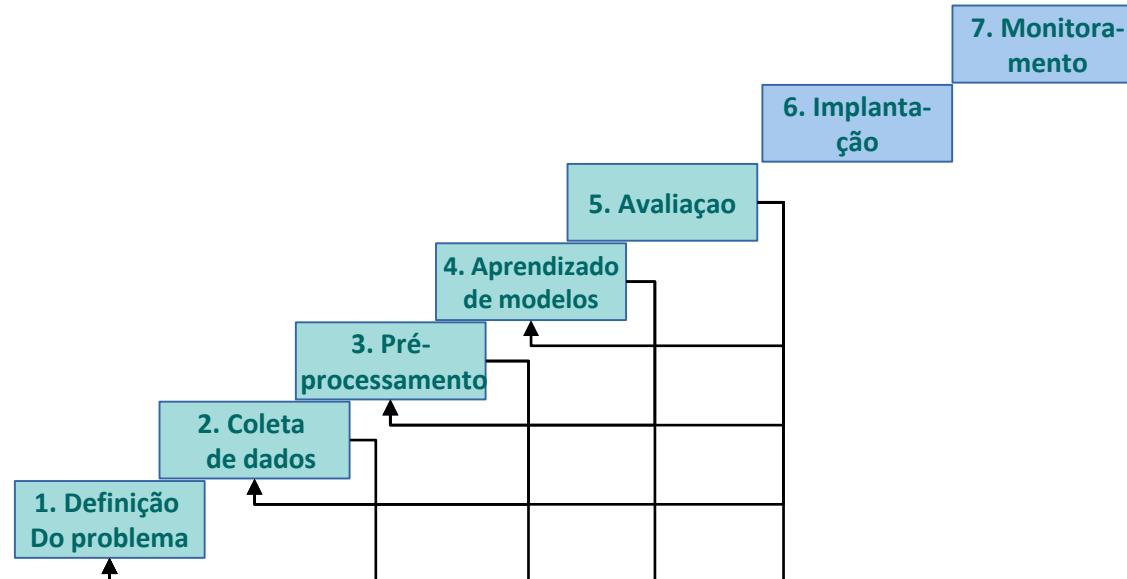
- Para fazer excelentes produtos, faça modelagem preditiva como o grande engenheiro que você é, não como o excelente especialista em modelagem preditiva que você não é!
- A maioria dos problemas que você enfrentará são, de fato, problemas de engenharia.
- A maioria dos ganhos provêm de excelentes features, não grandes algoritmos ML.
- Poucos cenários práticos requerem o projeto de novos algoritmos.

Make it simple! Make it work first!

- Estratégia básica:
 - Certifique-se de que o seu pipeline é sólido de ponta a ponta.
 - Comece com um objetivo razoável.
 - Adicione recursos de senso comum de forma simples.
 - Certifique-se de que seu pipeline permaneça sólido.
- Essa abordagem gerará muito dinheiro e/ou fará muita gente feliz por um longo período de tempo.
- Divirja apenas quando não há mais soluções simples.
- A adição de complexidade retarda as versões futuras.

Documentação do ciclo de vida

- Qual deve ser o resultado final de cada etapa do ciclo?



Definição de objetivo

- Também conhecida como fase de entendimento de negócio.
- Esta fase deve terminar com um objetivo claramente estabelecido para o modelo implantado final.
- Deve-se consolidar um documento que descreva, de forma objetiva e clara, métricas de avaliação e metas específicas.

- A fase de coleta de dados conclui com a consolidação de uma documentação, que descreve:
 1. Dados coletados;
 2. Conclusões sobre como os dados devem ser explorados;
 3. Discussão sobre quais questões relacionadas a qualidade dos dados devem ser abordadas.

- A fase de preparação de dados conclui com a construção de documentação, que contém:
 1. A localização dos arquivos de entrada, usados na modelagem dos banco de dados;
 2. O código usado para limpar os dados e criar o banco de dados de modelagem;
 3. A localização do banco de dados de modelagem final;
 4. Uma descrição conceitual de como os aspectos técnicos relacionados ao pré-processamento e transformação foram abordados.

- A fase de construção do modelo conclui com um relatório, documentando:
 1. As variáveis preditoras e os coeficientes no modelo final;
 2. Os principais testes de significância estatística e visualizações, empregadas para variáveis de preditores individuais;
 3. Os testes do desempenho geral do modelo.

Avaliação

- Na conclusão da fase de avaliação do modelo, é elaborado um relatório que:
 1. Demonstra que os objetivos de modelagem foram cumpridos.
 2. Documentos de assinatura dos principais interessados.
 3. Resume o protótipo do impacto e da estratégia de implantação (deployment).

- Etapa que se preocupa com a implantação do modelo em ambiente de produção.
- Transformação de um protótipo em um produto final.
- Processo deve ser todo automatizado.
- Processo de aprovação rigorosamente definido.
- Necessário ter dados para testar o processo de implantação.
- Ao fim deste etapa, devemos ter um produto completamente operante.

- A fase de implantação deve terminar com um documento contendo:
 1. Etapas do processo de deployment;
 2. Especificações do ambiente;
 3. Script de implantação a ser seguido;
 4. Script de publicação de novas versões.

- Devemos suportar o monitoramento contínuo do modelo.
- O modelo tem o impacto pretendido no mercado?
 - Retenção.
 - Taxa de custo.
 - Livro de qualidade empresarial.
 - Índice geral de perda de livro de negócios.
- Monitoramento deve ser automático, dinâmico e intuitivo.

- A fase de implantação deve terminar com um documento, contendo:
 1. Descrição das ferramentas de monitoramento;
 2. Endereço e descrição de forma de acesso e interpretação do monitoramento;
 3. Descrição das metodologias e métricas empregadas.

- Melhores práticas de modelos preditivos.

■ Próxima aula

- ❑ Discussão sobre melhores práticas em modelagem preditiva.
- ❑ Relatos de experiência.
- ❑ Erros comuns.



Aula 6.2.2. Melhores práticas (Parte 2)

- ❑ Melhores práticas de modelos preditivos.

■ Primeiro, projete e implemente métricas

- Acompanhe o máximo possível no seu sistema atual.
- É mais fácil obter permissão do usuário do sistema para obter dados históricos no início.
- Projete seu modelo com a instrumentação de métricas em mente.
- Você notará o que muda, ou permanece igual ao longo do tempo.



Prefira machine learning a uma heurística complexa

- Uma heurística simples pode criar um novo produto. Porém, uma heurística complexa não é sustentável.
- Com dados e uma idéia básica do que você está tentando realizar, use ML.
- Como na maioria das tarefas de modelagem preditiva, você estará constantemente atualizando sua abordagem, seja através de uma heurística ou de ML, e ML é mais fácil de atualizar e manter.



Mantenha o primeiro modelo simples e obtenha a infraestrutura correta

- Um primeiro corte sobre o que é "bom" e "mau", significa para o seu sistema que:
 - Escolher recursos simples torna mais fácil garantir que as features sejam utilizadas pelo algoritmo de aprendizagem corretamente.

KEEP
CALM
AND USE
**OCCAM'S
RAZOR**

■ Teste a infraestrutura independentemente do modelo

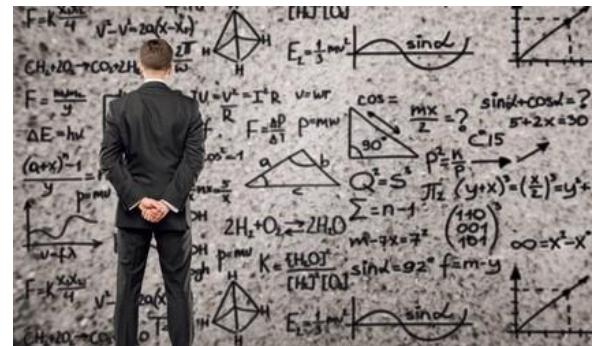
IGTI

- Teste a obtenção de dados no algoritmo.
- O modelo em seu ambiente de treinamento, dá o mesmo resultado que o modelo no seu ambiente de produção?



Comece com um modelo interpretável

- Facilite a depuração.
- Sempre tente um modelo linear.
- Obtenha informações sobre a qualidade de cada recurso.
- Tente entender porquê o modelo funciona ou não.



Planeje a publicação de várias versões

- Espere lançar mais de um modelo.
- Analise o impacto da complexidade em futuros lançamentos.
- Verifique o quanto fácil é:
 - Adicionar, remover ou recombinar features.
 - Crie uma nova cópia do pipeline.
 - Verificar a correção do pipeline.
 - Tenha duas ou três cópias em paralelo.

Explore features simples

- Com toneladas de dados, é mais fácil aprender milhões de features simples, do que poucas complexas.
- Mantenha grupos de features em que cada feature se aplica a uma fração muito pequena de seus dados, mas a cobertura geral é superior a 90%.
- Use a regularização para eliminar as features que se aplicam a poucos exemplos.

- Por que você não deve avaliar os resultados do modelo:
 - Você está muito perto do código e pode procurar um aspecto particular dos dados.
 - Seu tempo é muito valioso.
- Como avaliar:
 - Pague leigos para responder perguntas em uma plataforma de crowdsourcing.
 - Execute um experimento em tempo real, com usuários reais.

Não perca tempo com novas features, se objetivos conflitantes existem

- À medida em que suas medidas se estendem, sua equipe começará a analisar questões que estão fora do escopo dos objetivos de sua ML atual.
- Você precisa alterar o objetivo, ou os objetivos, do seu produto



Decisões de publicação são um proxy para objetivos de longo prazo

- As únicas decisões de publicação fáceis, são quando todas as métricas melhoram (ou, pelo menos, não pioram).
- As decisões de publicação dependem de múltiplos critérios, apenas alguns dos quais podem ser otimizados diretamente.
- As métricas mensuráveis nos testes A/B, em si, são apenas um proxy para objetivos a longo prazo: satisfazer os usuários, aumentar os usuários, satisfazer os parceiros e lucrar.
- Nenhuma métrica cobre a preocupação final da equipe: “como este produto estará em cinco anos?”

Quando nada muda, procure fontes de informação qualitativamente novas

- Quando é difícil ver melhorias significativas nas métricas, embora você esteja adicionando novas features:
 - É hora de começar a construir a infraestrutura para recursos radicalmente diferentes.
 - Talvez seja a hora de usar o deep learning.
 - Comece a ajustar suas expectativas quanto ao retorno esperado no investimento.
 - Pese o benefício de se adicionar novas features, contra o custo de uma complexidade maior.

- Melhores práticas de modelos preditivos.



Muito obrigado!