ML

Daniel O. Cajueiro

Departamento de Economia, Universidade de Brasília.

Brasília, 2018

ML

Big Picture

É uma área da ciência que aborda o desenvolvimento de algoritmos e técnicas que permitem computadores/máquinas aprenderem a realizar tarefas, fazer escolhas ou prever resultados.

ML

Good fellows

Nesses últimos anos tenho me beneficiado de trabalhar com pessoas superhabilidosas e que compartilharam suas idéias sobre ciência de dados comigo:

Estudantes (Atuais) ECO	**Professores**
Camila Pereira	Bernardo Mello (FIS)
Flávio Vitorino	Herbert Kimura (ADM)
Julia Scotti	Marina Rossi (ECO)
Murilo Diniz	Pedro Albuquerque (ADM)
Pedro Campelo	Rafael Terra (ECO)
Santiago Ravassi	Roberto Andrade (FIS-UFBA)
Saulo Benchimol	
Victor Candido	

ML

Áreas relacionadas

- Ciência da Computação: pois desenvolve algoritmos eficientes
- Inteligência Artificial: poisa outra desenvolve algoritmos inteligentes que tentam imitar o cérebro humano.
- Estatística: pois desenvolve técnicas para inferir a partir de amostras
- Econometria: pois técnicas similares são usadas na área de econometria de séries temporais e pesquisa recente mostra como cada uma das áreas podem contribuir com a outra.
- Matemática Aplicada: pois desenvolve técnicas de otimização numérica e prova convergência de resultados.

ML

Aplicações usuais

- Classificação de Padrões (ex: imagens, crédito bancário)
- Processamento de Linguagem Natural (ex: como usar notícias como fonte de informação)
- Reconhecimento de Objetos (ex: reconhecer pessoas ou objetos)
- Jogar jogos (xadrez, gamão)
- Previsão de variáveis relevantes (ex: inflação, retornos de ativos financeiros)
- Clustering

ML

Modelos de Aprendizagem

- Aprendizagem supervisionada (ex: regressão, Classificação)
- Aprendizagem não supervisionada (ex: clustering, estimação de densidade, principal component analysis, autoencoders)
- Aprendizagem por reforço (modelos de programação dinâmica, modelos baseados em simulações monte carlo)

Modelos de Aprendizagem

Aprendizagem Supervisionada

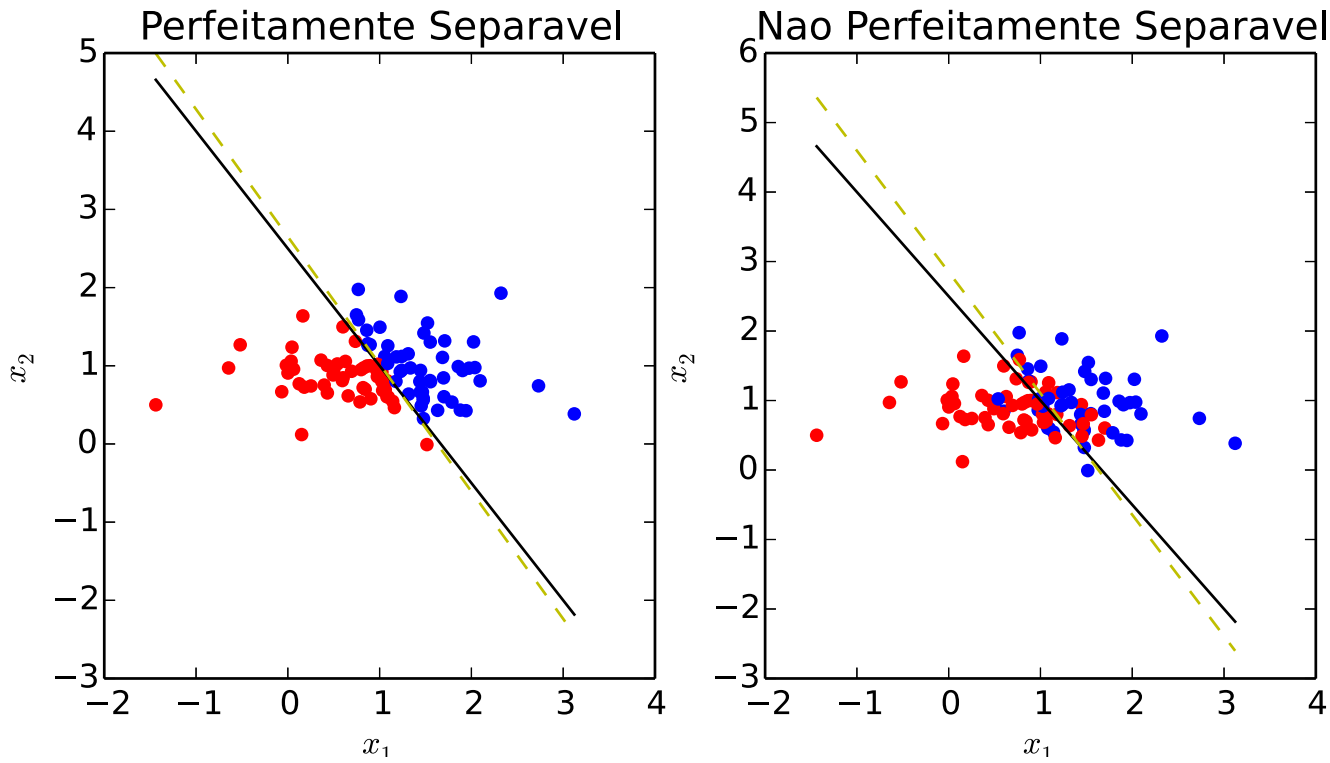
Fonte: PRorum.com

Algoritmos de aprendizagem supervisionada supõem a existência de um "Professor" que te ensina que tipo de comportamento você deve exibir em cada situação.

Exemplo

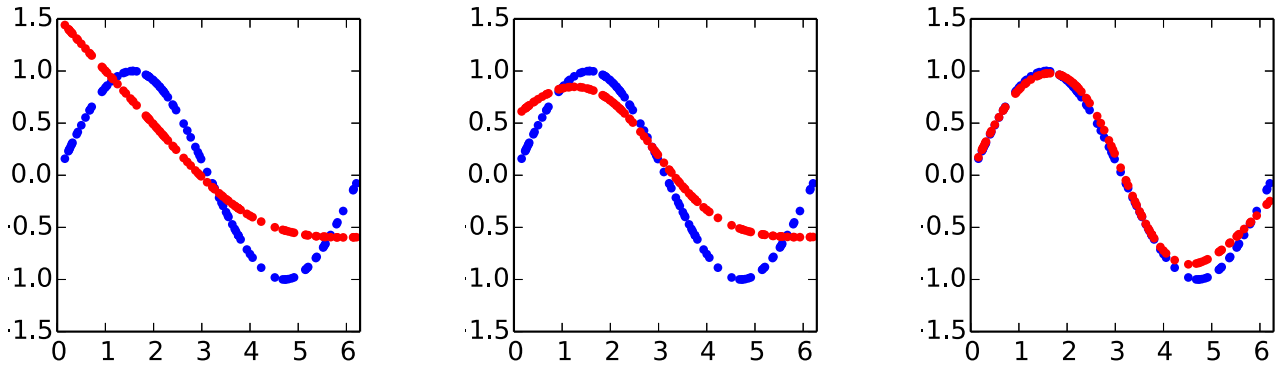
Imagine que você deseja classificar empresas saudáveis de não-saudáveis e para fazer isso você tem uma amostra que associa cada empresa saudável uma série de variáveis. Então, um algoritmo de aprendizagem supervisionado tentaria usar explicitamente essa informação para no futuro ser hábil para separar empresas saudáveis de não-saudáveis.

Um exemplo de classificação (Probit):



Fonte: Nosso curso de Métodos Computacionais

Um exemplo de regressão usando uma rede neural (perceptron multicamada):



Fonte: Nosso curso de Métodos Computacionais

Modelos de Aprendizagem

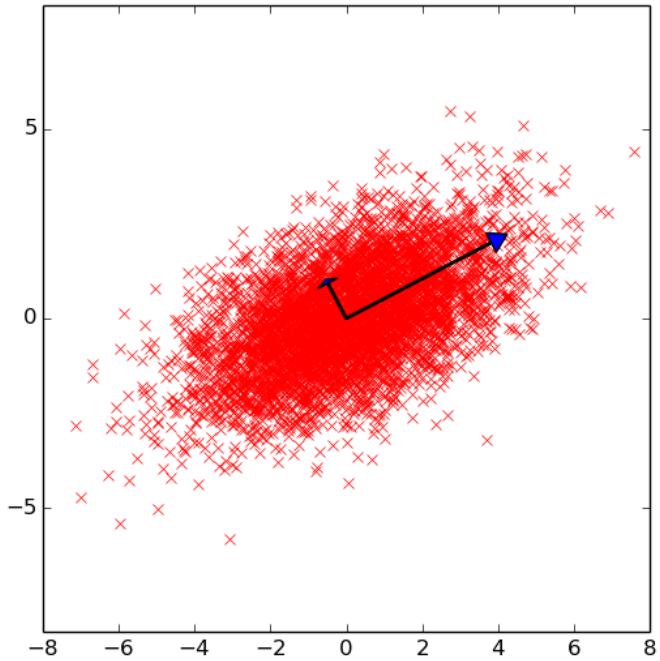
Aprendizagem Não Supervisionada

Eles normalmente associam o seu aprendizado a métricas que devem ser otimizadas. Muitas vezes, algoritmos não-supervisionados são usados, como um passo anterior a um algoritmo supervisionado, para buscar associações e similaridades entre entradas e saídas em massas de dados.

Modelos de Aprendizagem

Aprendizagem Não Supervisionada

Exemplo de Análise dos Componentes principais: Vetores principais de uma massa de dados gerada a partir de uma distribuição normal bivariada.



Modelos de Aprendizagem

Aprendizagem Não Supervisionada

Exemplo de Análise dos Componentes principais: Representação de imagens usando autovetores.

Fonte: OpenCV documentation

Modelos de Aprendizagem

Aprendizagem por reforço

Considera o problema de mapear estados (situações) em ações de forma a maximizar um sinal numérico.

Exemplo

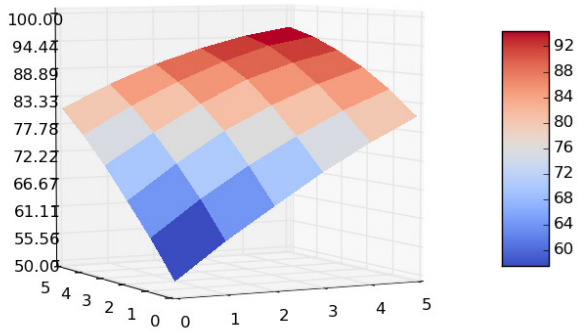
(Administração de Locadora de carros)

- Existem duas locadoras e um número de clientes chega em cada locadora para alugar carros.
- Se há um carro disponível, ela aluga este carro na locação e recebe 10 reais da matriz. Se ela não tem carro disponível, então o negócio é perdido.
- Pode-se mover carros de uma locação a outra no meio da noite ao custo de 1 real por carro movido.
- Supõe-se que o número de carros procurados e retornados em cada locadora são variáveis aleatórias do tipo poisson conhecidas (previamente estimadas).

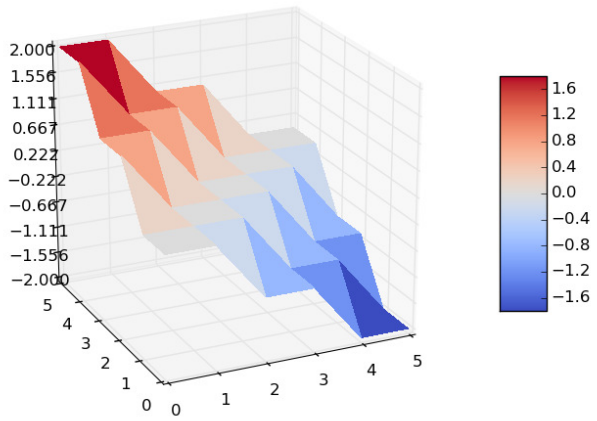
Aprendizagem por reforço

Exemplo

Value



Control

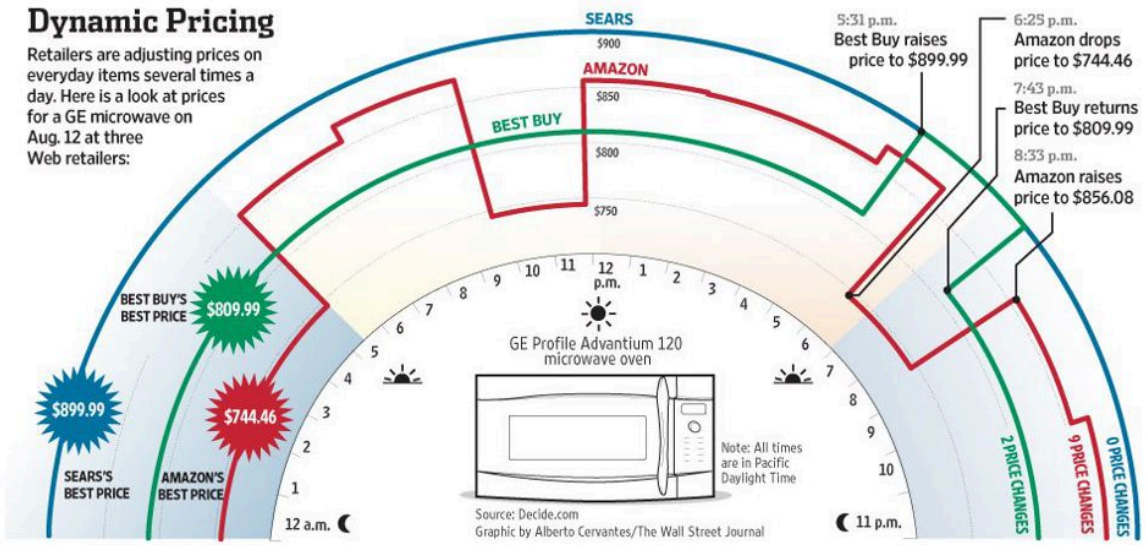


Fonte: Nosso curso de métodos computacionais

Aprendizagem por reforço

Exemplo

Exemplo de apreçamento dinâmico:



Fonte: http://www.customerforlife.com

Regras de ouro em ML

O objetivo é generalizar

- Não se espera que o mesmo exemplo apareça duas vezes
- De uma forma geral, quando maior a diversidade dos seus dados, melhor será o seu preditor
- O problema se torna mais difícil ainda em dimensões mais altas

Regras de ouro em ML

Apenas dados não são suficientes - precisamos de estrutura para representação

- Na maioria dos problemas existem um contínuo de exemplos. Logo, nunca você terá acesso a todos os exemplos.
- Faça hipóteses sobre os seus dados que ajude os métodos usados a resolverem o problema.
- Modelos diferentes possuem estruturas diferentes. Nem todos os métodos são adequados para todos os problemas.

Regras de ouro em ML

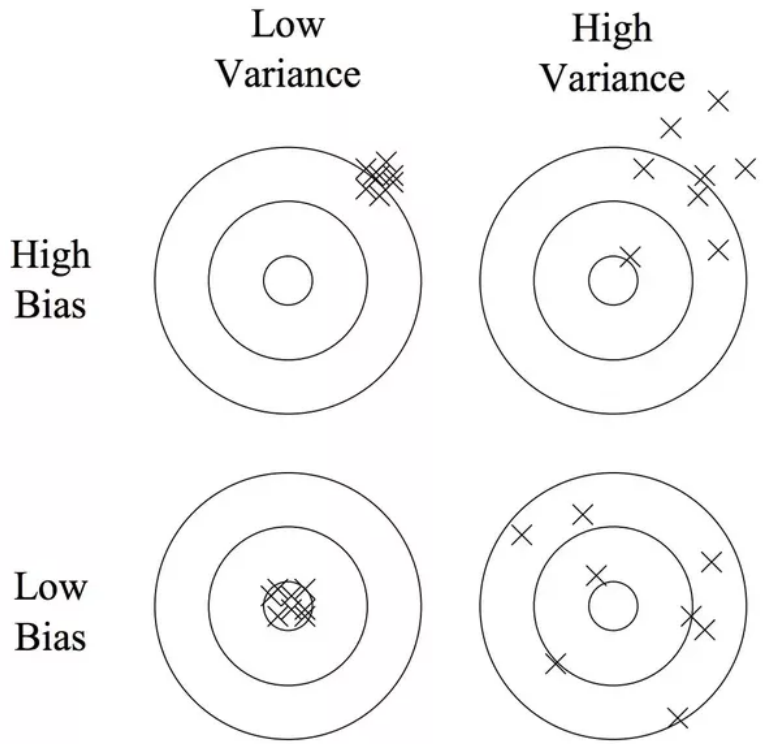
Tome cuidado com overfitting

- Não precisamos chegar a mínimos globais (pois isso pode significar overfitting)
- Considere cross-validation
- Considere regularização
- Considere usar o método gradiente estocástico para a otimização da função objetivo
- Considere dropout

Regras de ouro em ML

Existe um tradeoff entre viés e variância

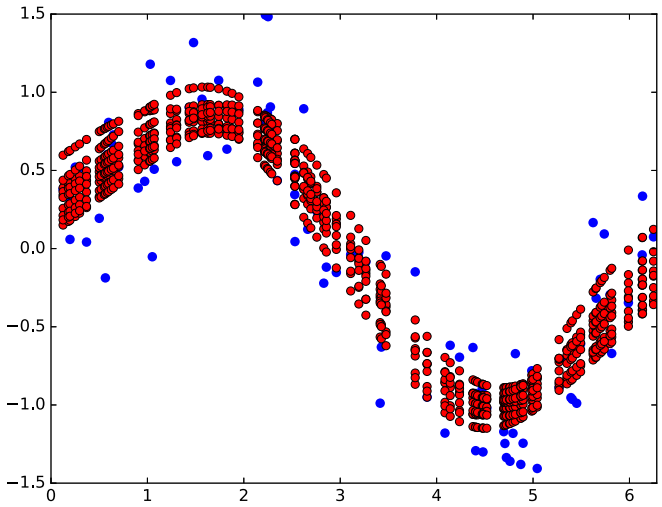
Viés e variância em diferentes modelos:

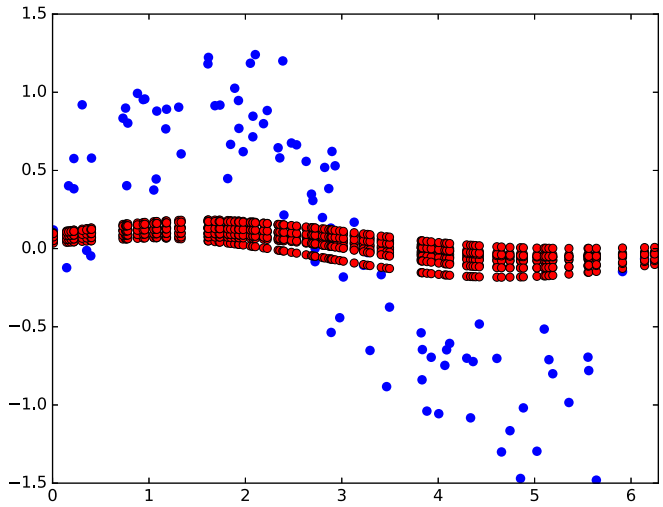


Regras de ouro em ML

Existe um tradeoff entre viés e variância

- Modelos muito flexíveis $\Rightarrow$ Alta variância e baixo viés
- Modelos pouco flexíveis $\Rightarrow$ baixa variância e alto viés

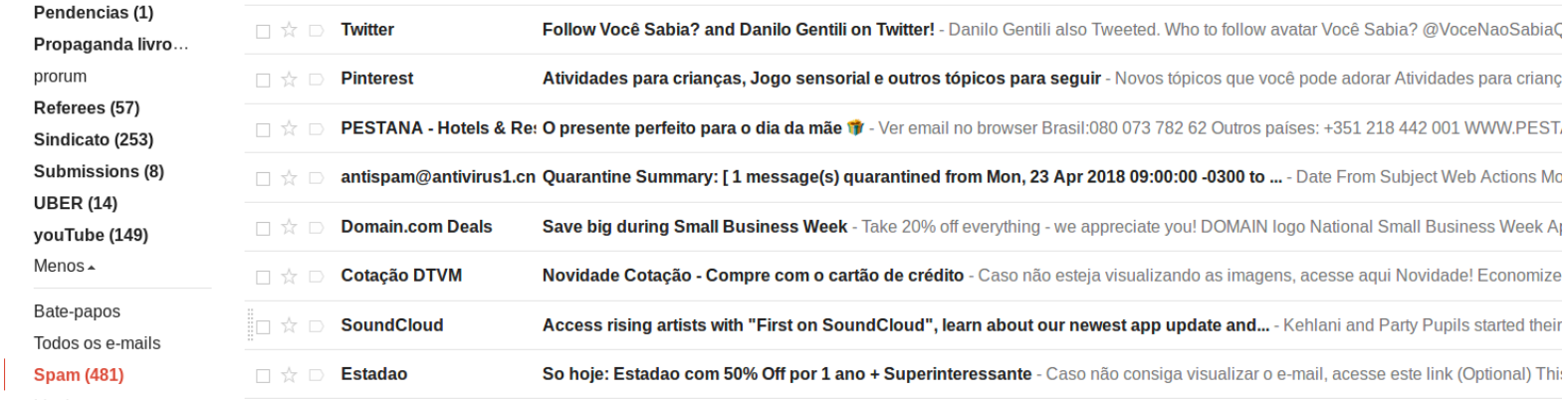




Exemplo usando uma rede neural de função de base onde estamos variando o parâmetro de regularização

Regras de ouro em ML

Escolha cuidadosamente quais características de suas bases de dados você deseja usar



- São comuns problemas em que o número de características de cada exemplo (variáveis independentes) é maior que 100. Quais delas usar?
- Você pode eliminar características inúteis
- Você pode automaticamente fazer a escolha das características (ex. Lasso)
- Você pode usar técnicas como PCA para buscar uma nova representação para essas variáveis (aprendizagem não supervisionada)

Regras de ouro em ML

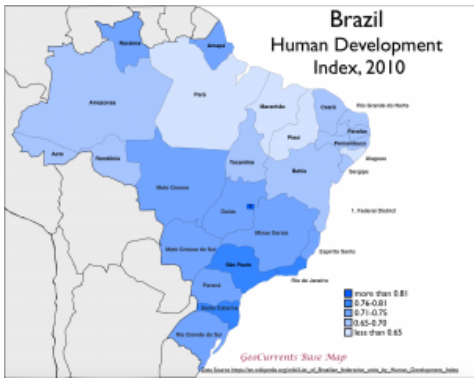
Tome cuidado com a escalabilidade de sua solução

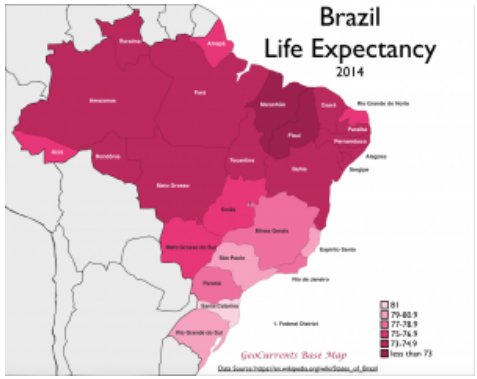
- Em muitos problemas atuais, existe uma infinidade de dados disponíveis na internet. O maior problema é tempo para processá-los.
- Não necessariamente os algoritmos mais inteligentes são os melhores. Alguns algoritmos possuem complexidade computacional muito alta que pode ser inviável para a sua base de dados
- Heurísticas são muito bem vindas
- Faça implementações inteligentes e eficientes (não apenas que funcionam)

Regras de ouro em ML

Considere muitos modelos e não apenas um - aprenda com todos eles

Um mapa não é um território!

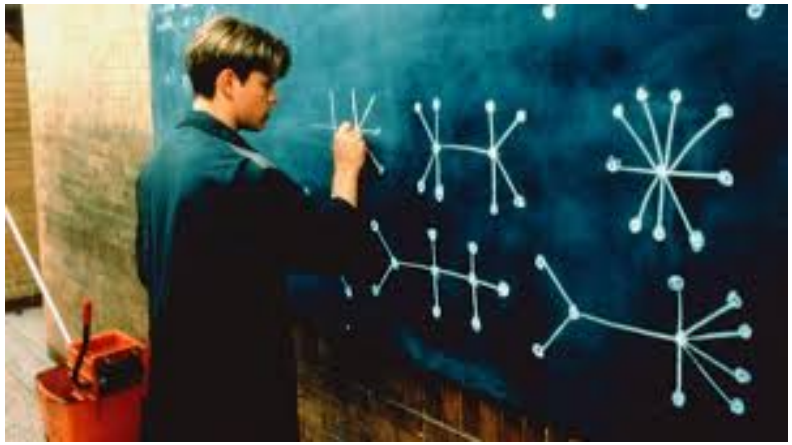




- Os melhores modelos dependem da Aplicação específica [Contraste com Econometria: escolhe-se um modelo tendo como base princípios econômicos]
- Bagging: Gere subamostras de seu conjunto de treino, aprenda um modelo para cada subamostra e depois combine os resultados
- Boosting: Usa-se pesos para combinar os modelos
- Stack: Outputs de vários modelos se tornam entradas de um outro modelo que escolhe a melhor forma de combina-los

Regras de ouro em ML

Resultados matemáticos relacionados com aproximação não necessariamente implicam em aprendizagem



– Será que isso tem alguma utilidade?

- "O espaço de polinômios é denso no espaço de funções contínuas"
- O erro de previsão é assintoticamente normal

ML X Estatística

ML:

- Objetivo: Representação do conhecimento de forma que possa ser aprendido por algoritmos e ser usado para a previsão ou tomada de decisão em ambiente sujeito a incerteza
- Área: Subárea da Ciência da Computação
- Escopo: Inferência Estatística, Aprendizagem por Reforço, Inteligência Artificial (algoritmos de busca, sistemas especialistas), Deep Learning (super pouco motivado pelo que se estuda em estatística)
- Ponto de vista: Preocupa-se com algoritmos escaláveis que possam ser validados empiricamente e minimizem o erro de previsão. Complexidade de algoritmos, pois normalmente lida com problemas com um número grande de dimensões. Não se preocupa se resultados assintóticos são válidos ou não.
- Linguagens de programação: Python
- Divulgação dos resultados: Conferências [e Journal of Machine Learning Research, IEEE Transactions]
- Idade: Campo recente da pesquisa que pôde realmente começar a funcionar quando os computadores se tornaram mais potentes e foi possível acessar uma quantidade maior de dados.
- Como praticantes de ML veem estatística?: Campo que traz uma caixa de ferramentas (Estatística Básica, Estatística Multivariada, Estatística Bayesiana, Statistical Learning - SVM e Kernel methods) que inclui vários modelos industriais. Se preocupam muito com propriedades assintóticas de estimadores.

Estatística:

- Objetivo: Construção de modelos que possam ser usados para testar hipóteses para a população usando apenas parte dela.
- Área: Subárea da Matemática (Teoria da Probabilidade)
- Escopo: Inferência Estatística, Planejamento de Experimentos e Técnicas de Amostragem
- Ponto de vista: Preocupa-se com propriedades assintóticas de estimadores e processos geradores de dados. Preocupa-se com a validade de resultados assintóticos.
- Linguagens de programação: R, SAS
- Divulgação dos resultados: Jornais [Journal of American Statistical Association, Communications in Statistics, Biometrika].
- Idade: Campo clássico da matemática que foi evoluindo com o tempo e com a potência dos computadores. Mais especificamente a popularização das áreas conhecidas como estatística bayesiana e statistical learning é bem recente.
- Como estatísticos veem ML?: Algo muito legal, que pode ser facilmente trabalhado sem muito esforço usando R (visão adequada se forem considerados apenas os modelos industriais). Aceitam muitas heurísticas sem provas.

ML X Econometria

O que cada uma delas faz?

ML:

- Objetivo: Minimizar erro de previsão (não está interessado em causalidade).
- Seleção de variáveis e modelos: Usa procedimentos como validação cruzada.
- Hipóteses: Independência entre as observações e mesmas distribuições para os conjuntos de treino e teste.
- Interpretabilidade dos resultados: Não se preocupa (na prática gostaria de entender melhor).
- Viés nos estimadores: Não é um problema desde que o erro de previsão esteja bom.
- Overfitting: É um desastre.
- Moedas de troca: (1) Paga com viés para NÃO receber overfitting. (2) Paga com (possível) ausência de causalidade, para receber previsibilidade.
- Número de amostras: Pode ser ou não menor que o número de variáveis.
- Uso dos dados: Usa parte dos dados para estimar e parte dos dados para testar.

Econometria:

- Objetivo: Estimar de forma confiável os parâmetros do modelo (causalidade é fundamental).
- Seleção de variáveis e modelos: Teoria Econômica (na prática não é bem assim: Vamos torturar esses dados até eles nos contarem o que queremos ouvir!).
- Hipóteses: Independência entre as observações, exogeneidade do erro.
- Interpretabilidade dos resultados: É um dos aspectos principais.
- Viés: É fundamental que os modelos sejam não-viesados.
- Overfitting: Não é um problema.
- Moedas de troca: (1) Paga com overfitting para NÃO receber viés. (2) Paga com perda de perda de previsibilidade, para receber causalidade.
- Número de amostras: Normalmente é muito maior que o número de variáveis (preferência por modelos parsimoniosos).
- Uso dos dados: Usa todos os dados para estimação.

Como ML pode contribuir para econometria?

Após 5 semanas de alta, mercado prevê inflação menor para 2017 e vê PIB maior

Expectativa dos analistas para o IPCA de 2017 passou de 3,51% para 3,45%. Para o PIB, previsão de alta subiu de 0,34% para 0,39% neste ano; mercado também prevê queda maior do juro em 2017.

Previsão de séries temporais: Técnicas de ML podem ser usadas para prever variáveis aleatórias

- Tema clássico! Out of sample forecast usando ML!
- Existem diversos exemplos relevantes sobre o assunto:
  - Previsão de inflação
  - Previsão da estrutura a termo da taxa de juros
  - Previsão de atividade econômica
  - Previsão de retornos de ativos financeiros (Velha discussão entre eficiência estatística e eficiência financeira)
- Nowcasting = Now + Forecasting
  - Acessar estatísticas relevantes sem atraso (exemplo PIB)
  - Estimativa de algo hoje que só terá acesso no futuro

Como ML pode contribuir para econometria?

Previsão de variáveis relevantes para políticas públicas (ou privadas...)

- Técnicas de ML podem ser usadas para prever um resultado que é essencial para a tomada de decisão na alocação de recursos.
- Vários exemplos interessantíssimos:
  - Um indivíduo deve sair ou não da prisão?
  - Um indivíduo deve ou não receber um empréstimo?
  - Um indivíduo deve ou não ser indicado para receber um rim?
  - Um indivíduo deve ou não ser revistado? Qual a chance dele possuir uma arma ou ser um terrorista?

Como ML pode contribuir para econometria?

Fonte: Rachel Thomas (Twitter)

Previsão de variáveis relevantes para políticas públicas (ou privadas...)

- Preocupações:
  - Possíveis preconceitos disponíveis nos dados podem ser transferidos para as máquinas?
  - Humanos respondem a incentivos e podem adaptar seus comportamentos a depender da alocação de recursos.

Como ML pode contribuir para econometria?

Seleção de variáveis em modelos econométricos

- Idéia clássica: Enquanto economistas usam dados para VALIDAR teorias, cientistas da área de Aprendizagem de Máquinas usam dados para GERAR teorias: *The wire: (...) detectives Mcnulty and Moreland revisit the scene of an old, unsolved homicide. (...) (They) let the data speak by themselves. (...) The idea that a detective would investigate a crime scene only to confirm a predetermined hunch about the identity of the murderer would seem abhorrent to many (...)*

Como ML pode contribuir para econometria?

Seleção de variáveis em modelos econométricos

- Economistas defendem: "Teoria deve justificar a inclusão ou não de uma variável no modelo"
- Economistas torturam dados implementando um procedimento bizarro e difícil de replicar:
  - Rodam um monte de regressões (não necessariamente as melhores e não dizem quais)
  - Ficam com aquelas variáveis importantes ou que deram coeficientes significantes.
  - Esse procedimento pode inclusive gerar um viés por causa da omissão de variáveis

Como ML pode contribuir para econometria?

Seleção de variáveis em modelos econométricos

- Idéia básica: Ao invés de tortura de dados, pode-se usar técnicas de ML para fazer a escolha de variáveis.
- Problema: ML funciona para previsão e não para estimação.
- Solução em dois estágios: ML para seleção de variáveis e OLS para o segundo estágio.
- Consequência: Replicabilidade de resultados.
- Necessita de procedimentos especiais para construir os intervalos de confiança (separação da amostra em duas partes).
- Pode ser usado para estimar o efeito médio de um programa ou tratamento

Como ML pode contribuir para econometria?

Análide de robustez de coeficientes

- Idéia básica: ML permite considerar vários possíveis modelos para a escolha de um parâmetro
- Variações dos modelos originais podem ser criadas interando os parâmetros de interesse com as outras covariáveis
- Um medida de robustez pode ser o valor do desvio padrão do coeficiente de interesse em todos os possíveis modelos.

Como ML pode contribuir para econometria?

Completamento de matrizes de dados

- Idéia básica: Em uma matriz de dados cujos alguns dados são faltantes, o objetivo é prever o melhor conjunto de dados que preenchem a matriz.
- No caso de problemas de avaliação de política pública, se considerarmos as unidades tratadas, obviamente não temos a situação dessas unidades não tratadas (contrafactual)

Como ML pode contribuir para econometria?

Identificação dos efeito de um tratamento ou programa em subgrupos

- Embora métodos usuais de avaliação de política pública sejam capazes de inferir os efeitos médios de um programa ou tratamento em uma população, essa população pode ser bastante heterogênea.
- Muitas vezes é particularmente interessante descobrir em quais subgrupos da amostra os tratamentos são realmente efetivos.
- A solução desse problema nessa nova literatura fornece a divisão em subgrupos utilizando uma árvore de decisão que pretende minimizar o erro quadrático do efeito médio do tratamento.
- Isso pode ser útil para ao escolher de novos elegíveis para o programa, escolher apenas aqueles que são fortemente afetados.

Como ML pode contribuir para econometria?

02/05/2018 às 08h26

Eletrobras fecha acordo para encerrar ação coletiva nos EUA

Por **Marcelle Gutierrez | Valor**

SÃO PAULO - A Eletrobras assinou memorando de entendimentos para entrar em acordo com relação a ação coletiva (class action) de investidores no Tribunal Distrital dos Estados Unidos para o Distrito Sul de Nova York (SDNY), conforme comunicado divulgado nesta quarta-feira (2).

O acordo tem como objetivo encerrar todas as ações em curso iniciadas por investidores que adquiriram ações ordinárias e preferenciais da Eletrobras representadas por American Depository Shares (ADS) e terminará com o pagamento de US$ 14,75 milhões.

Segundo a Eletrobras, no documento, o acordo elimina o risco de um julgamento adverso durante a instrução do processo, o que poderia afetar a empresa e sua situação financeira.

A corte americana deve aprovar o acordo, após uma revisão preliminar, e os investidores podem se opor e não aderir. Se a aprovação preliminar for concedida, os membros da classe da ação coletiva serão notificados sobre os termos do acordo e seus direitos.

Tratamento de dados não-estruturados para a criação de novas variáveis que podem ser usadas em modelos econométricos

- Dados de mídia
  - Análise textual de jornais, revistas e outras mídias sociais.
- Imagens de satélite.
  - Medidas de crescimento econômico [inclusive validação de estatísticas disponibilizadas pelos países].
  - Medidas de características de cidades.
  - Estimativa do estoque de petróleo no mundo.

Como ML pode contribuir para econometria?

Criação de novas variáveis utilizando aprendizagem não-supervisionada

- Criação de variáveis dummies que representam grupos similares
- Criação de variáveis dependentes que representem grupos.

Como ML pode contribuir para econometria?

Tornar experimentos online mais eficientes

- Multi-armed bandit problem: Um número fixo de recursos deve ser alocado entre escolhas alternativas e parcialmente conhecidas de forma que maximizem o ganho esperado. A informação sobre cada uma dessas escolhas alternativas aumenta proporcionalmente ao tempo em que se aloca recursos em cada uma delas.
- Como experimentos podem ser projetados para associar indivíduos a tratamentos usando dados de indivíduos anteriores para associar os novos indivíduos a tratamentos mais adequados, equalizando "exploration" (exploração) e "exploitation" (extração)?

Como ML pode contribuir para econometria?

Contribuições técnicas para a área de econometria estrutural

- Técnicas de aproximação da função valor.
- Técnicas para acelerar a convergência.

Como ML pode contribuir para econometria?

Criação de uma interface entre economia e ciência da computação

- Mudanças em cursos de graduação de economia.
- Laboratórios de economistas com perfil multidisciplinar.
- Professores precisarão sair da zona de conforto para atender as demandas das ciências de dados.
- Replicabilidade de resultados empíricos.

Referências

- The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World - Pedro Domingos
- Pattern Recognition and Machine Learning - Christopher Bishop
- Deep learning - Ian Goodfellow, Yoshua Bengio and Aaron Courville
- The elements of statistical learning - Hastie, Tibshirani e Friedman
- Modern multivariate statistical techniques - Alan Julian Izenman
- The discipline of machine learning - T. M. Mitchel
- A few useful things to know about machine learning - P. Domingos
- Learning deep architectures for AI - Y. Bengio

Referências

- In defence of forensic social science - Amir Goldberg [Big data and Society, 2015]
- Sociology in the era of big data: the ascent of forensic social science - D. A. McFarland e K. Lewis [American Sociology, 2015]
- Economic reason and artificial intelligence - D. C. Parkes and M. P. Wellman [Sience 349, p.267, 2015]

Referências

- Big Data: New Tricks for Econometrics - H. R. Varian
- The Impact of Machine Learning on Economics - Susan Athey
- The State of Applied Econometrics: Causality and Policy Evaluation Susan Athey e Guido W. Imbens.
- Beyond Prediction: Using Big Data for Policy Problems - Susan Athey
- "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics - Victor Chernozhukov, A. Belloni and C. Hansen
- Prediction Policy Problems - Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer