Machine learning em saúde

Prof. Dr. Alexandre Chiavegatto Filho

LABDAPS

LABORATÓRIO DE BIG DATA E ANÁLISE PREDITIVA EM SAÚDE


INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY
The Economist
How to fix America's inner cities
The self-service economy
Time to open up Indonesia
Inside the anti-bribery business
Why humans cause heatwaves
Artificial Intelligence
The promise and the peril


FEATURING: MACHINE LEARNING ➤ DRIVERLESS CARS ➤ SUPERHUMANS
POPULAR SCIENCE
SPECIAL EDITION
THE NEW ARTIFICIAL INTELLIGENCE
THE FUTURE OF NANOBOTS
DETECTING CANCER


THE DECLINE OF INTERNATIONAL STUDIES
JULY/AUGUST 2015
FOREIGN AFFAIRS
Hi, Robot
Work and Life in the Age of Automation


What it takes to end an AIDS epidemic
Polar bears suffer through lean summers
Sperm produced in ovary of mutant fish
Science
$10
17 JULY 2015
AAAS
SPECIAL ISSUE
ARTIFICIAL INTELLIGENCE


The Perils of Special Counsels / This Is Your ISIS on Drugs
Newsweek
06.02.2017
THE DOCTOR WILL SEE YOU NOW
HOW AI IS GOING TO CURE OUR SICK HEALTH CARE SYSTEM

OS DOIS RITMOS DA LAVA-JATO
Curitiba: 107 condenados
Brasília: nenhum
MADONNA FALA A VEJA
"Sou uma rebelde e serei rebelde até o fim"
Abril
veja
DE MÃOS DADAS COM A INTELIGÊNCIA ARTIFICIAL
Longe dos cenários futuristicos da ficção científica, ela já faz parte do presente. Mas em que medida pode servir ao ser humano e, ao mesmo tempo, ameaçá-lo?
EXCLUSIVO Artigo de Yuval Noah Harari, autor de Homo Deus: uma Breve História do Amanhã

- Inteligência artificial não é hype criado pela mídia.
  - É consequência dos avanços científicos dos últimos anos.

- Por que têm ocorridos avanços exponenciais nos últimos 5 anos?
  - **Três fatores:**

    1 – Avanços em capacidade computacional (modelos de machine learning exigem muita memória).

    2 – Aumento da quantidade de dados (importante para melhorar performance).

    3 – Novos algoritmos para problemas mais complexos (deep learning).

Machine learning

- Inteligência artificial clássica: regras para a tomada de decisão ensinada por humanos.
  - Como identificar spam: via palavras-chave.
  - Como traduzir uma frase: dicionário e regras de gramática.
  - Como identificar caras humanas: ensinar o que é nariz, olho, boca etc.

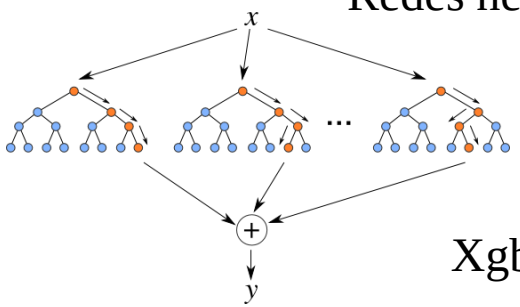
- Inteligência artificial com machine learning: máquinas aprendendo sozinhas!
  - Tomada de decisão via identificação de padrões complexos nos dados.

- É como uma criança aprende!

Machine learning

- Problemas **práticos** de predição (para a tomada de decisão).
  - Pouco interesse em *interpretar* os modelos.
  - Liberdade para modelar a complexidade do mundo real.

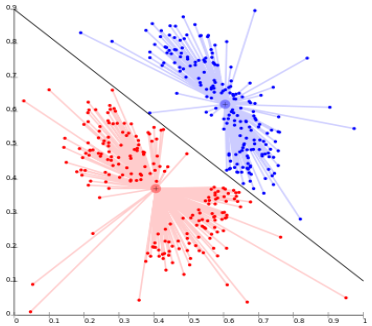
Redes neurais

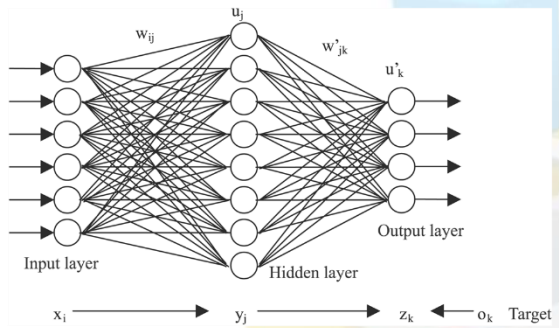


K-means

Xgboost

Florestas randômicas





Regressões penalizadas

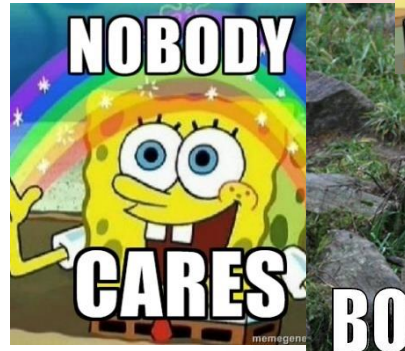
Machine learning

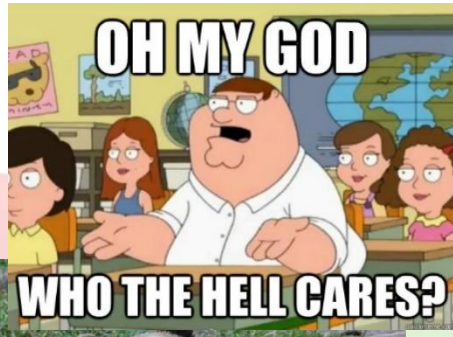
- Se machine learning não se importa muito com interpretação, então se importa de fato com o quê?

  - **Performance preditiva.**
    - Ou seja, acurácia das decisões.

- Inversão do interesse da reta de regressão:

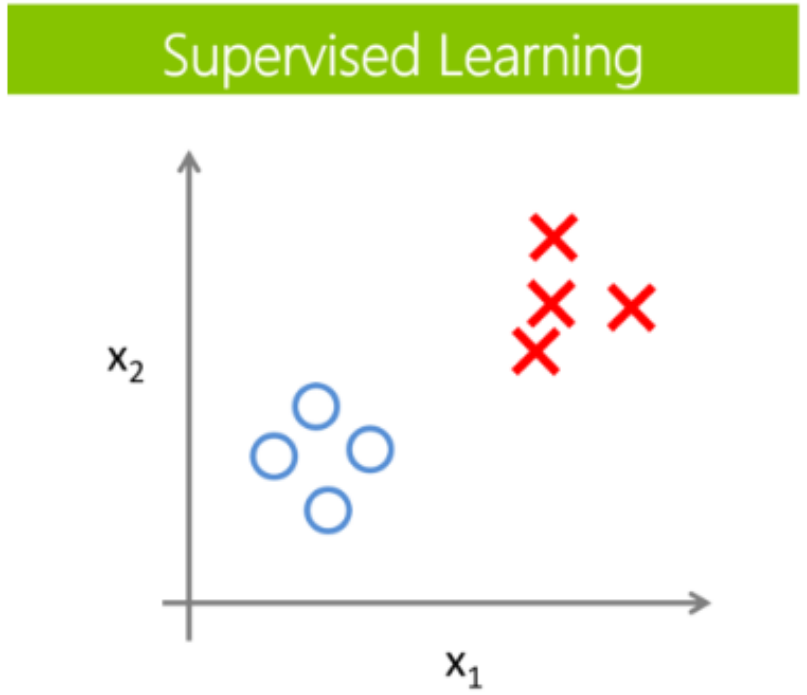
$$Y_{glicemia} = (\ldots\ldots\ldots)$$





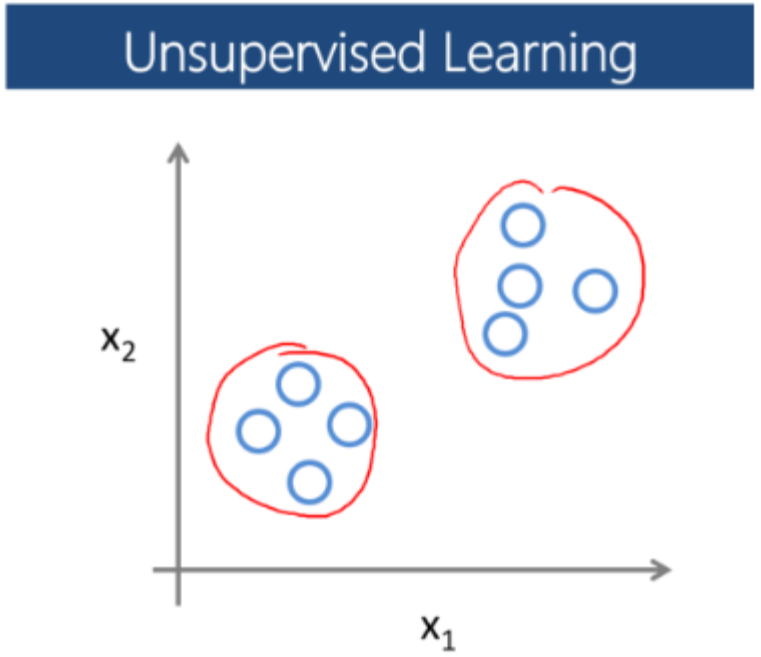
Quatro categorias de machine learning:

- Aprendizado supervisionado:
  - Quando os dados incluídos para treinar o algoritmo incluem a solução desejada, ou rótulo (“label”).
  - Resposta certa.



Quatro categorias de machine learning:

- Aprendizado não-supervisionado:
  - Não existe rótulo (“label”).
  - Algoritmo aprende sem uma resposta certa.
  - Mais comuns: clustering (agrupamentos) e redução de dimensão.



Quatro categorias de machine learning:

- Aprendizado semi-supervisionado:
  - Presença de alguns dados com rótulo e outros sem.
  - Identificação de fotos do Facebook: algoritmo identifica que a mesma pessoa está em várias fotos e só precisa de um rótulo.

- Aprendizado por reforço:
  - Interação com um ambiente dinâmico com feedbacks em termos de premiações e punições.

Predição com machine learning

- Desenvolver algoritmos que façam boas predições em saúde.

- Principais razões pelas quais algoritmos às vezes não apresentam boa performance preditiva:

  - Pré-processamento inadequado dos dados.
  - Validação inadequada dos algoritmos.
  - Extrapolação inadequada.
  - Sobreajuste (mais importante).

Pré-processamento dos dados

- A presença de outliers, correlações aleatórias e erros de medida podem prejudicar a performance preditiva dos modelos.

- Chave para a boa performance preditiva. É onde as competições do Kaggle são vencidas (deep learning + xgboost).

Técnicas de pré-processamento de dados:
- Seleção das variáveis.
- Padronização.
- Redução de dimensão.
- Colinearidade.
- Valores missing.
- One-hot encoding.

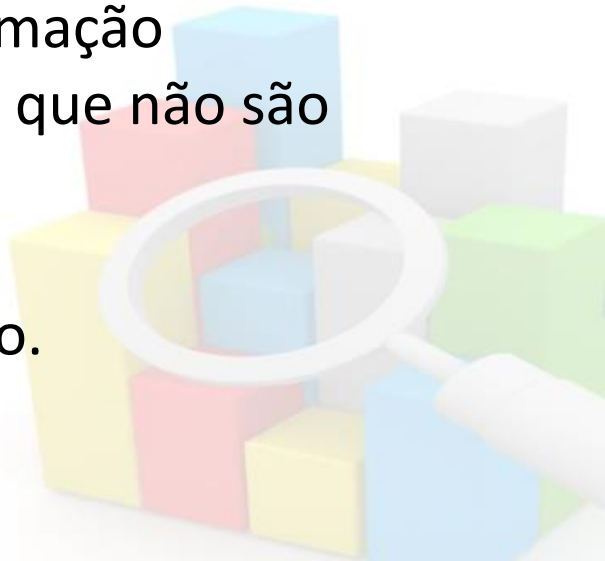
Pré-processamento dos dados

- Pré-selecionar variáveis que sejam preditoras plausíveis (bom senso do pesquisador).

- Cuidado com vazamento de informação (“data leakage”).

    - Acontece quando os dados de treino apresentam informação escondida que faz com que o modelo aprenda padrões que não são do seu interesse.

    - Uma variável preditora tem escondida o resultado certo.



Pré-processamento dos dados

- Exemplo: incluir o número identificador do paciente como variável preditora.
  - Problema: se pacientes de hospital especializado em câncer tiverem números semelhantes.
  - Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes.
  - Esse algoritmo aprendeu algo interessante para o sistema de saúde?

- Motivo pelo qual os dados e os algoritmos de machine learning precisam ser abertos.
  - Watson prediz bem: mas é informação útil ou vazamento?

Padronização

- A escala das variáveis pode afetar muito a qualidade das predições.

- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.

- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1.

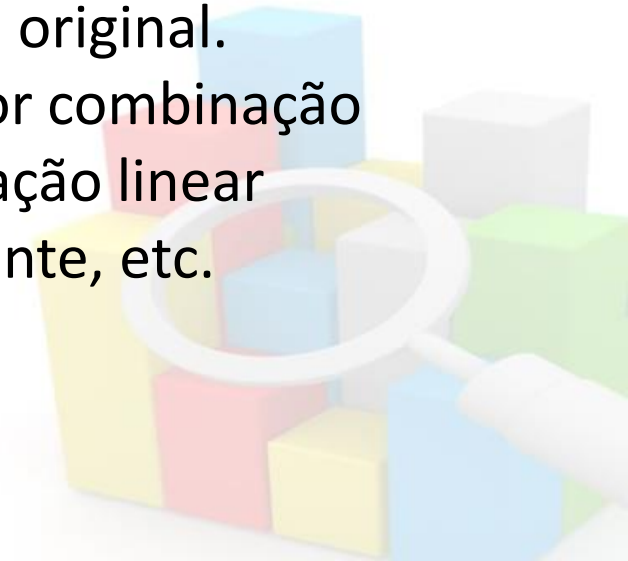
$$z_i = \frac{x_i - \mu}{\sigma}$$

- Ou seja, é feita a subtração da média e a divisão pelo desvio padrão dos valores da variável.

Redução de dimensão

- Quanto maior a dimensão dos dados (número de variáveis) maior o risco de sobreajuste do modelo.

- Análise de Componentes Principais:
  - Técnica de aprendizado não supervisionado.
  - O objetivo é encontrar combinações lineares das variáveis preditoras que incluam a maior quantidade possível da variância original.
  - O primeiro componente principal irá preservar a maior combinação linear possível dos dados, o segundo a maior combinação linear possível não correlacionada com o primeiro componente, etc.



Redução de dimensão

- Uma das razões pela qual a ACP é tão utilizada, é o fato de que cria componentes principais não correlacionados.
  - Na prática, alguns algoritmos conseguem melhor performance preditiva com variáveis com baixa correlação.

- Uma outra forma de diminuir a presença de variáveis com alta colinearidade é excluí-las.
  - Variáveis colineares trazem informação redundante (tempo perdido).
  - Além disso, aumentam a instabilidade dos modelos.
  - Estabelecer um limite de correlação com alguma outra variável (0,75 a 0,90).

Variáveis missing

- É importante entender por que valores de uma variável estão faltantes.

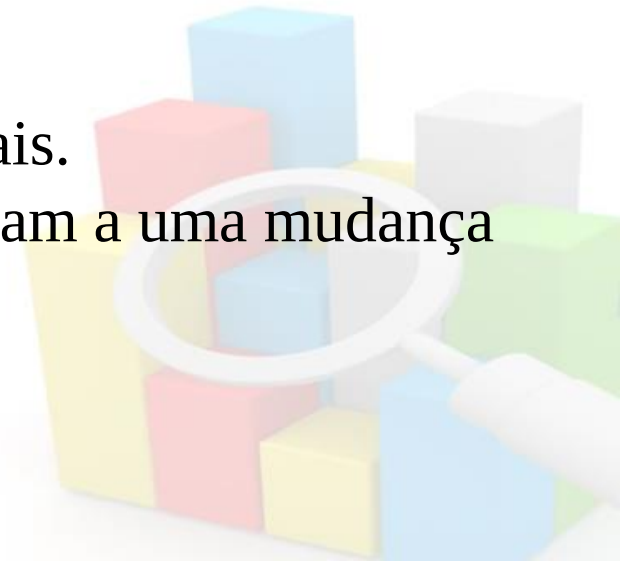
- Se for por um motivo sistemático: informação preditiva.
  - Grande diferença em relação a estudos de inferência, em que valores missing devem ser evitados.
  - Informação preditiva: não conseguiu responder a uma pergunta sobre o seu passado → pode ajudar na predição de problemas cognitivos graves no futuro.
  - Em variáveis categóricas adicionar uma categoria para missing.
  - Imputação com machine learning.

One-hot encoding

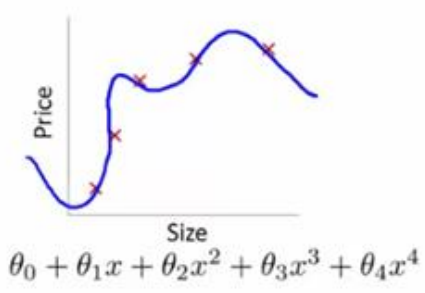
- Alguns algoritmos têm dificuldade em entender variáveis que têm mais do que uma categoria.
  - Acham que é uma variável contínua (0, 1, 2, 3...) → porém não têm significado contínuo.
- A solução é transformar todas as categorias em uma variável diferente de valores 0 e 1 (one-hot encoding).
  - Variável com n categorias → criadas n variáveis.
- Pode trazer problemas em alguns modelos, como na regressão linear.
  - Solução: criar dummies.
  - n-1 variáveis (deixar a mais frequente como categoria de referência).

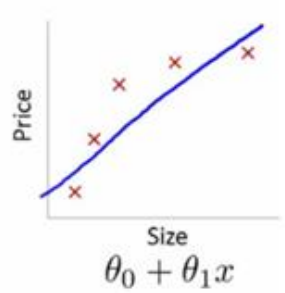
Sobreajuste

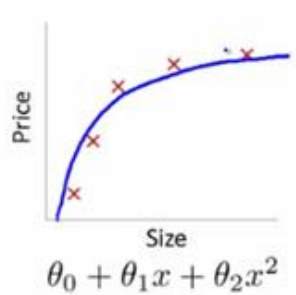
- Principal problema de machine learning.
  - Modelos muito complexos:
    - Funcionam perfeitamente para a amostra em questão, mas não muito bem para amostras futuras.
      - Dados influenciados por fatores aleatórios e erros de medida.
  - Tradeoff entre viés e variância:
    - Viés: erro gerado pelo uso de modelos para dados reais.
    - Variância: quando pequenas mudanças nos dados levam a uma mudança muito grande nos parâmetros.



Sobreajuste







- Tradeoff entre viés e variância:
  - Modelo com alta variância e pouco viés:
    - 2 variáveis: linha que passa exatamente por todos os pontos.
    - Se ajusta perfeitamente aos dados atuais, mas não aos futuros.
  - Modelo com baixa variância e alto viés:
    - 2 variáveis: linha reta para associação não-linear.
    - Modelo simples, com baixo poder preditivo.

Sobreajuste

- Como avaliar se o seu modelo está com sobreajuste?
  - Avaliar a performance preditiva do modelo em dados que não foram utilizados para definir o modelo.
    - Se a performance preditiva cair muito com os novos dados: o modelo tem sobreajuste.
    - É muito fácil ter boa predição nos dados que foram utilizados para definir o modelo: é só tornar o modelo muito complexo.

Sobreajuste

- Soluções:
  - Utilizar dados do período seguinte.
    - Exemplo: treinar o modelo em dados de 2016 e avaliar sua performance em dados de 2017.
    - Problema: na maioria das vezes, os dados são coletados num mesmo período.
  - Separar os dados aleatoriamente em treino e teste.
    - Dados de “treino” (70-80%) são usados para definir o modelo e dados de “teste” (20-30%) são usados para analisar a performance preditiva do modelo.

Sobreajuste

- O que significa “definir” o modelo?

  - Estabelecer os parâmetros (definidos automaticamente) e os hiperparâmetros (definidos pelo pesquisador).

  - Hiperparâmetros são em geral regularizadores: ou seja, tentam controlar a complexidade dos modelos (para evitar o sobreajuste).

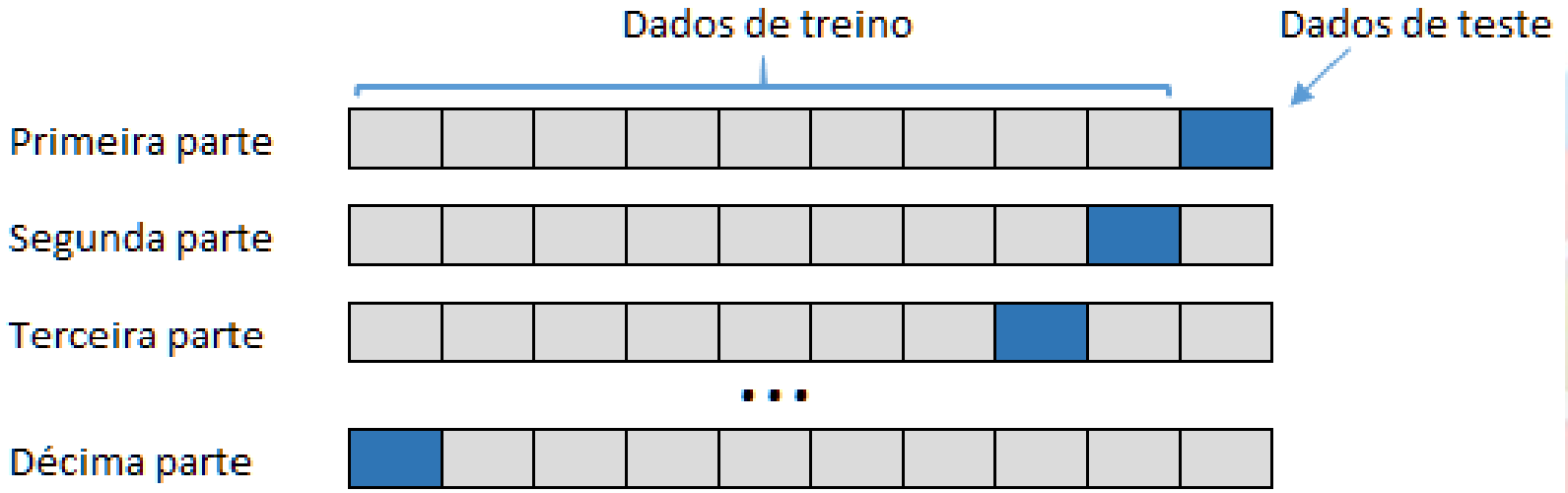
Sobreajuste

- Como selecionar os valores dos hiperparâmetros?

  - Pela análise da melhora da performance preditiva.
    - Problema: dados de teste só podem ser usados uma vez, para a seleção do melhor algoritmo.
    - Solução: selecionar os hiperparâmetros dos modelos nos dados de treino.
      - Problema: performance dos modelos deve sempre ser testada em dados que o algoritmo nunca viu.
      - Solução: validação cruzada.

Sobreajuste

- Validação cruzada de 10 partes (10-fold).
  - Dividir os dados em dez partes iguais e utilizar nove delas para treinar o algoritmo com um hiperparâmetro e a outra parte para testar a sua predição.



- Seleção do hiperparâmetro com melhor performance → definição do algoritmo com esse hiperparâmetro nos dados de treino.

- Muitos algoritmos têm mais de um hiperparâmetro: utilizar grid search.

- Testar todas as combinações possíveis de valores selecionados dos hiperparâmetros.
  - Se hiperparâmetros A e B tiverem valores selecionados de A = 1, 5, 10 e B = 50, 100, 150:
  - Testar (1;50), (1;100), (1;150), (5;50), (5;100)...

- Fazer o mesmo para todos os algoritmos.

- Seleção do hiperparâmetro com melhor performance → definição do algoritmo com esse hiperparâmetro nos dados de treino.

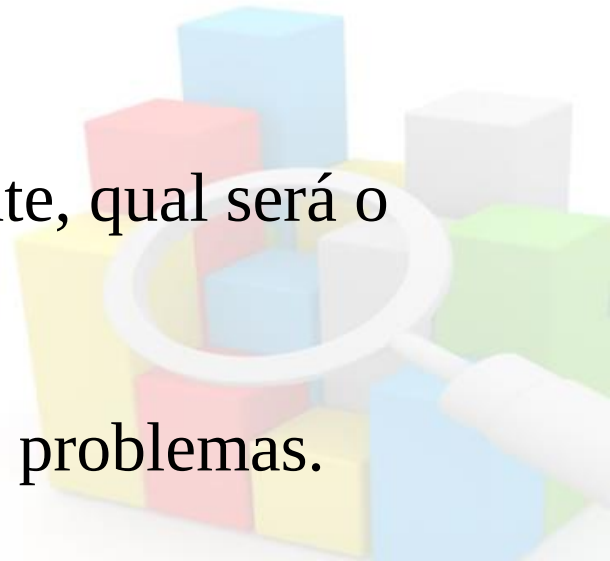
- Fazer o mesmo para todos os algoritmos.

- Teorema do “não há almoço grátis”:
  - Dado um conjunto infinito de dados, nenhum algoritmo é garantido a priori de ter melhor performance.
  - A única forma de saber qual vai ter melhor performance é testar todos.

  - Segredo: na prática, alguns costumam ganhar mais vezes (random forests e xgboost).

Tipos de modelos preditivos

- Divididos em dois grandes grupos:
  - Classificação.
    - Quando a variável a ser predita é qualitativa:
      - Ex: óbito em 5 anos, incidência de doença em 10 anos, etc.
  - Regressão.
    - Quando a variável a ser predita é quantitativa:
      - Ex: quantos meses de vida a pessoa tem pela frente, qual será o seu IMC no próximo ano, etc.
  - A maioria dos algoritmos pode ser utilizada para os dois problemas.



Medição de performance em problemas de regressão

- O mais comum é o uso da raiz quadrado do erro quadrático médio (RMSE, em inglês).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

  - Subtrair cada valor real do seu valor predito e elevá-lo ao quadrado. Somar todos e dividir pelo número de observações. Tirar a raiz quadrada para retomar o valor à sua escala original.
  - Outras possibilidades: R^2 e erro absoluto médio.

Medição de performance em problemas de classificação

Medição de performance em problemas de classificação

- Geralmente os modelos de classificação produzem dois resultados:
  - Probabilidade individual.
  - Categoria predita.

- Primeira possibilidade:
  - Acurácia: proporção de acertos.
  - Problema: algoritmos são malandros.
    - Se uma categoria ocorrer em 99% dos casos, o algoritmo vai predizer que todos os casos estão nessa categoria. Acurácia: 99%.

Medição de performance em problemas de classificação

- Acurácia:
  - Porém: isso não nos traz nenhuma informação.
  - Ex: identificar pacientes que possivelmente estão com câncer.
    - Esse algoritmo não nos diz nada.
    - Preferimos um algoritmo com **menor** acurácia.
    - Mas que acerte alguns/muitos casos de câncer.

Medição de performance em problemas de classificação

- Matriz de confusão:
  - Análise de concordância visual entre predição e realidade.

Predição	Realidade	
	Câncer	Sem câncer
Câncer	24	10
Sem câncer	36	130

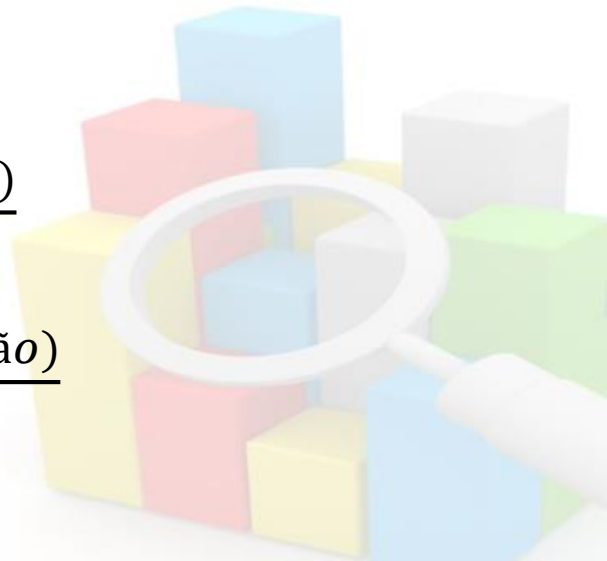
Medição de performance em problemas de classificação

- Matriz de confusão:

Predição	Realidade	
	Câncer	Sem câncer
Câncer	24	10
Sem câncer	36	130

$$\text{Sensibilidade} = \frac{Verdadeiros\ Positivos\ (predição)}{Positivos\ (realidade)}$$

$$\text{Especificidade} = \frac{Verdadeiros\ Negativos\ (predição)}{Negativos\ (realidade)}$$



Medição de performance em problemas de classificação

- Matriz de confusão:

Predição	Realidade	
	Câncer	Sem câncer
Câncer	24	10
Sem câncer	36	130

- Acurácia = (24+130) / 200 = 77%
- Sensibilidade = 24/(24+36) = 40%
- Especificidade = 130/(10+130) = 92,9%

Dificuldade de predizer câncer:

- Categorias desbalanceadas.
- Preditores ruins.
- Poucos casos.

Medição de performance em problemas de classificação

- Juntar sensibilidade e especificidade num mesmo resultado:

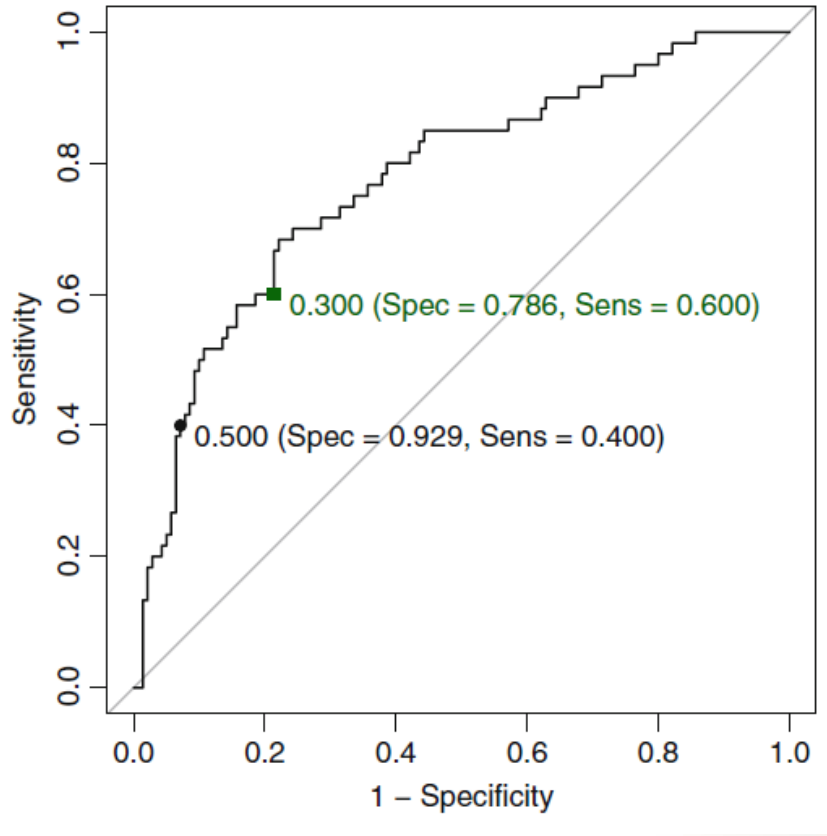
  - Curva ROC (Receiver Operator Characteristic).

  - No exemplo anterior, a sensibilidade foi baixa (40%) e a especificidade foi alta (92,9%).
  - Predição sobre câncer foi baseada em o algoritmo dar > 50% de probabilidade.
  - É possível melhorar a sensibilidade diminuindo o threshold?
  - Nesse exemplo, sim.
    - Threshold de 30% → sensibilidade de 60% e especificidade de 78,6%.

Medição de performance em problemas de classificação

- Curva ROC (Receiver Operator Characteristic).
  - Analisa diferenças na especificidade e sensibilidade de acordo com mudanças no threshold.

- Escolher o threshold mais interessante para a pesquisa.

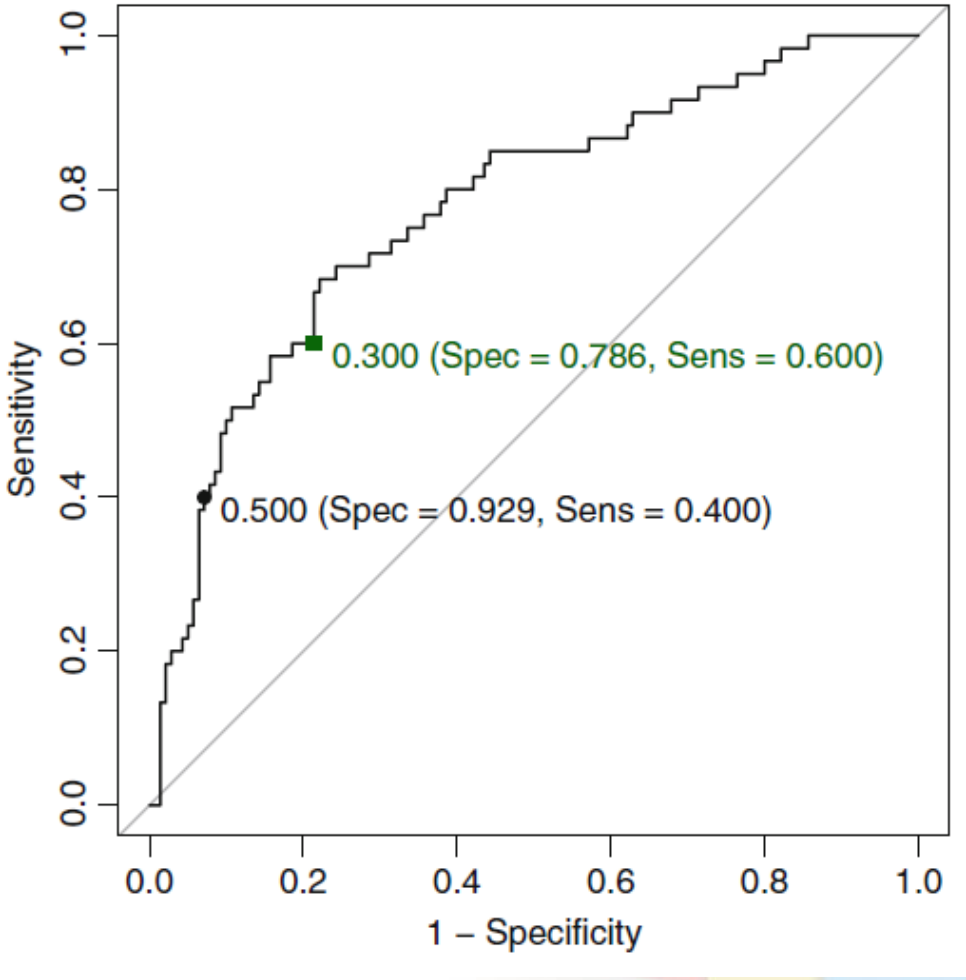


Medição de performance em problemas de classificação

- Ideal é uma curva mais à esquerda
e para cima possível.

- Linha diagonal 45º → modelo ineficiente.

- Valor único: área abaixo da curva (AAC).
  - Perfeito: 1,0
  - Ineficiente: 0,5
  - Exemplo: 0,78



Medição de performance em problemas de classificação

- Para alguns desfechos de saúde é fundamental pensar em termos de sensibilidade e especificidade.

- Por exemplo:

  - Teste de HIV/AIDS é importante diminuir falsos negativos (falsos positivos são um problema menor porque teste será refeito).

  - Indicação de cuidados paliativos: importante diminuir falsos positivos (não indicar seu início quando o tratamento aumentará a sobrevida).

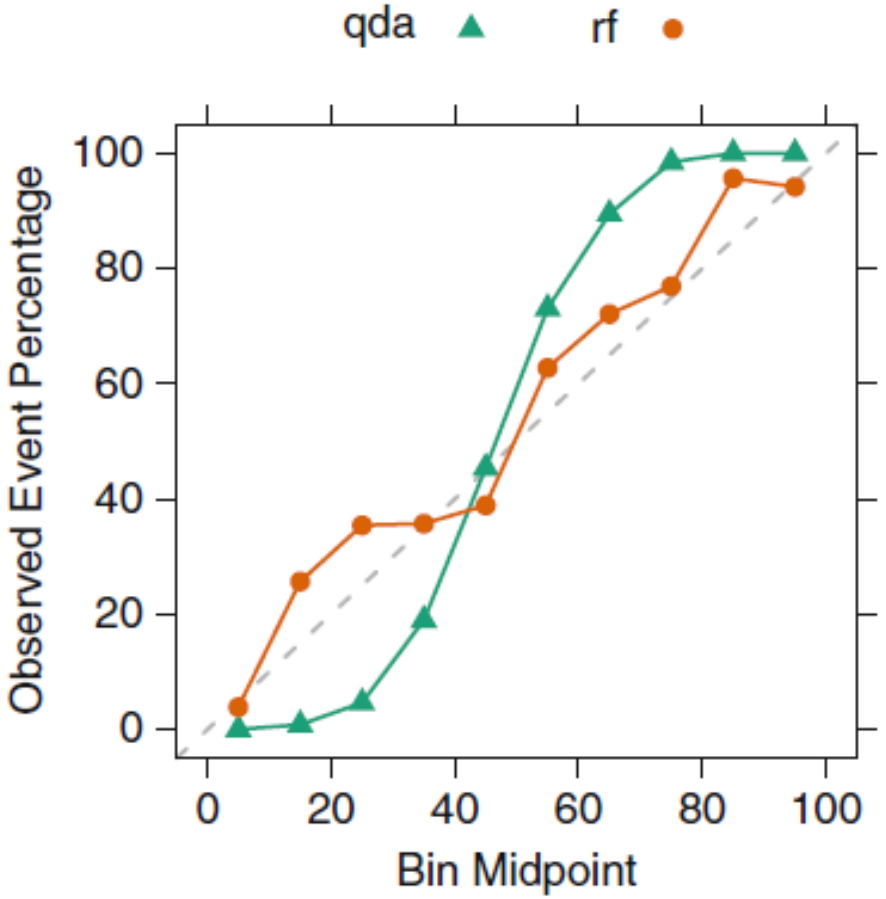
Medição de performance em problemas de classificação

- Solução para identificar onde a predição está errando.
  - Gráfico de calibração:
  - Separar observações segundo grupos de probabilidade predita.
    - Ex: [0 – 10%], ...]90 – 100%]
    - Em cada grupo identificar quantos de fato apresentaram o evento.

Medição de performance em problemas de classificação

- Gráfico de calibração (quadratic discriminant analysis e random forests).

Qual o melhor?



Importância de variáveis preditoras

- A análise da importância preditora das variáveis depende do algoritmo.
  - Regressão linear: interpretação simples pelos parâmetros.
  - Outros algoritmos: interpretação mais complexa.

  - Solução mais comum:
    - Análise da mudança do erro de predição ao permutar valores da variável.
      - Variável é importante para predição se erro aumenta. Se modelo não utiliza essa variável o erro não muda.

- Regressões

- Tanto a regressão linear (para desfecho contínuo) quanto a regressão logística (para desfecho categórico) são também utilizadas em machine learning.

- São mais comuns em estudos de inferência, mas também geram uma predição.
- Predição de índice glicêmico.

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i}$$

$$Y_{glicemia} = 1{,}23 + 1{,}35 * X_{idade} + 0{,}91 * X_{dieta}$$

- Regressões

- Modelos facilmente interpretáveis.

- Problema: em geral esses modelos têm sobreajuste quando há muitas variáveis preditoras, principalmente se forem colineares (baixo viés e alta variância).

- Solução: adicionar hiperparâmetros regularizadores.
  - Penalização contra a complexidade dos modelos.
  - Forçar o aumento do viés.
  - Pode ajudar a diminuir a variância e o erro de teste.

- Regressões penalizadas

  - Adicionar uma penalização se os parâmetros (β) ficarem muito altos.

  - Na regressão o objetivo é encontrar os β que minimizem a soma dos erros quadráticos.

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Regressões penalizadas

- É possível controlar (regularizar) o tamanho dos coeficientes pela adição de uma penalização à formula anterior.

$$\mathrm{SSE}_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

- Nesse caso é uma penalização L_2 , ou seja, quadrática nos parâmetros.
- A consequência é que agora estamos tentando minimizar o erro e o tamanho dos parâmetros.

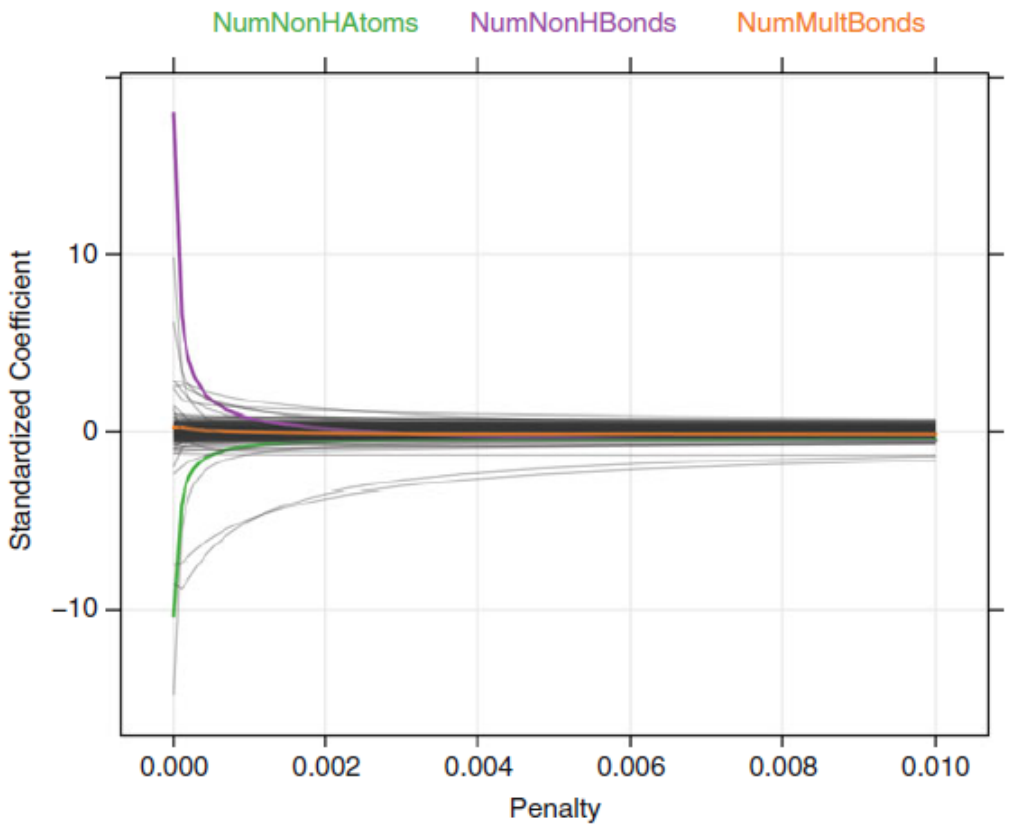
- Regressões penalizadas

$$\mathrm{SSE}_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

- A quantidade de regularização é controlada pelo parâmetro λ, quanto maior, maior a penalização.
- Ocorre um *encolhimento* dos parâmetros.
- Se λ = 0 não há penalização, regressão comum.
- Não tem encolhimento no β_0, queremos diminuir os efeitos das variáveis individuais e não do intercepto (média quando todas as variáveis são 0).

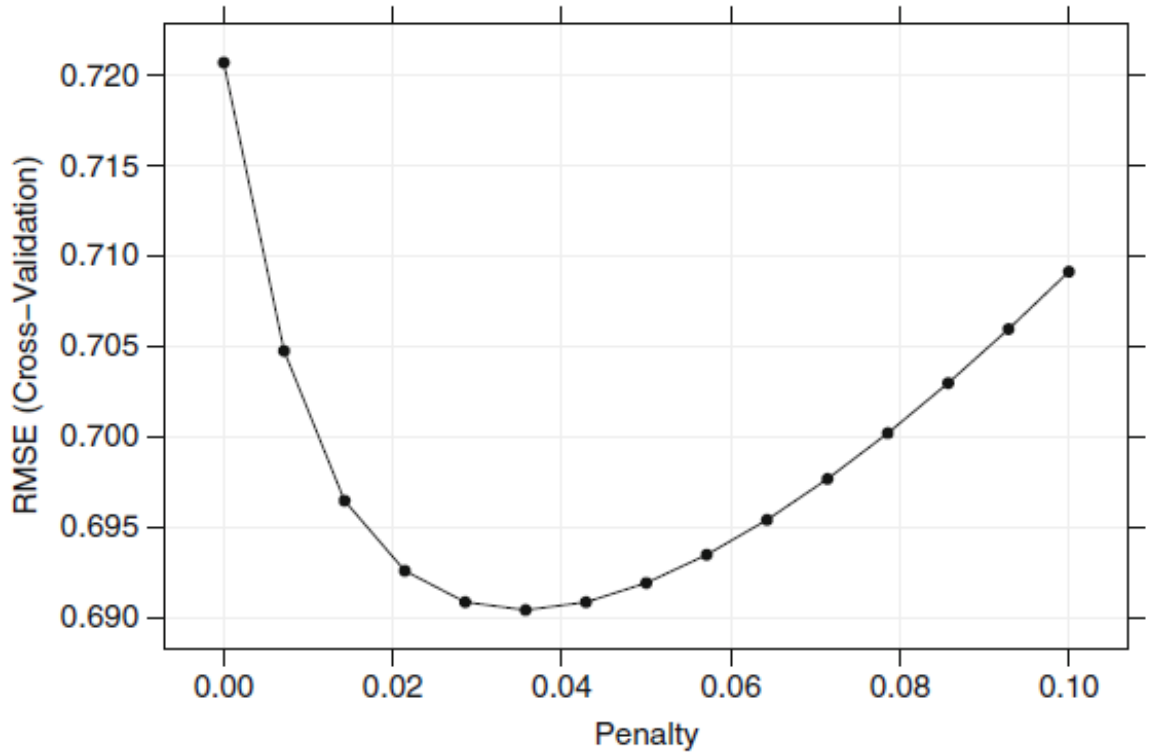
- Regressões penalizadas

  - A penalização L_2 é também conhecida como penalização ridge.



- Regressões penalizadas

  - O valor do valor de λ (penalização) é escolhido por validação cruzada. No caso: 0,036.



- Regressões penalizadas

- A penalização ridge encolhe os parâmetros, mas não reduz nenhum a 0, ou seja, não faz seleção de variáveis.

- Para isso, existe a penalização lasso (L_1).
- Lasso faz regularização e seleção de variáveis preditoras, pela penalização dos valores absolutos dos parâmetros.

$$\mathrm{SSE}_{L_1} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j|.$$

- Regressões penalizadas

- O valor do valor de λ (penalização) também é escolhido por validação cruzada. No caso: 0,15.

- Resultado para o exemplo:
  - RMSE (regressão linear): 0,71.
  - RMSE (regressão ridge): 0,69.
  - RMSE (regressão lasso): 0,67.

- Existe a possibilidade também de juntar as duas penalizações: rede elástica.

- Regressões penalizadas

  - Regressão logística.

    - Utilizada quando o desfecho a ser predito é categórico (problema de classificação).

    - Também é possível incluir penalizações de ridge, lasso e redes elásticas.

    - Mesmo sistema (só que os parâmetros da regressão logística são estabelecidos pelos métodos de máxima verossimilhança).