

Machine learning em saúde

Prof. Dr. Alexandre Chiavegatto Filho



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.

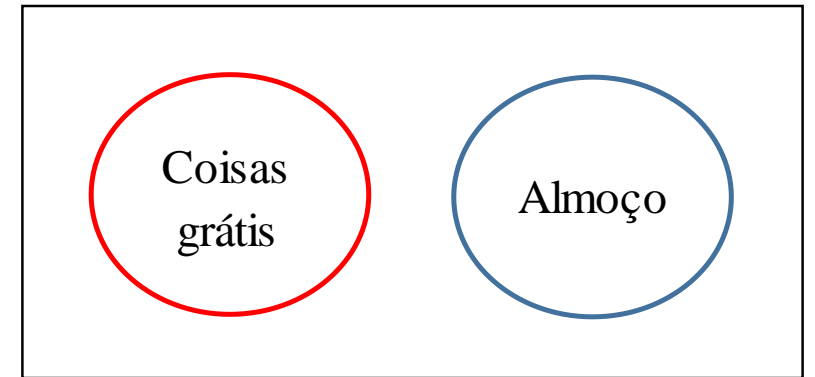
- Seleção dos hiperparâmetros com melhor performance → definição do algoritmo com esse hiperparâmetro nos dados de treino.

Teorema do “não há almoço grátis”:

- Em infinitos conjuntos diferentes 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

- *com dados reais, alguns costumam ganhar mais vezes (random forests e xgboost).*



Data-driven advice for applying machine learning to bioinformatics

Randal S. Olson, William La Cava, Zairah Mustahsan, Akshay Varik, Jason H. Moore. arXiv, 2018.

Análise da performance preditiva
(acurácia de VC) de 13 algoritmos
em 165 bancos de dados
diferentes.

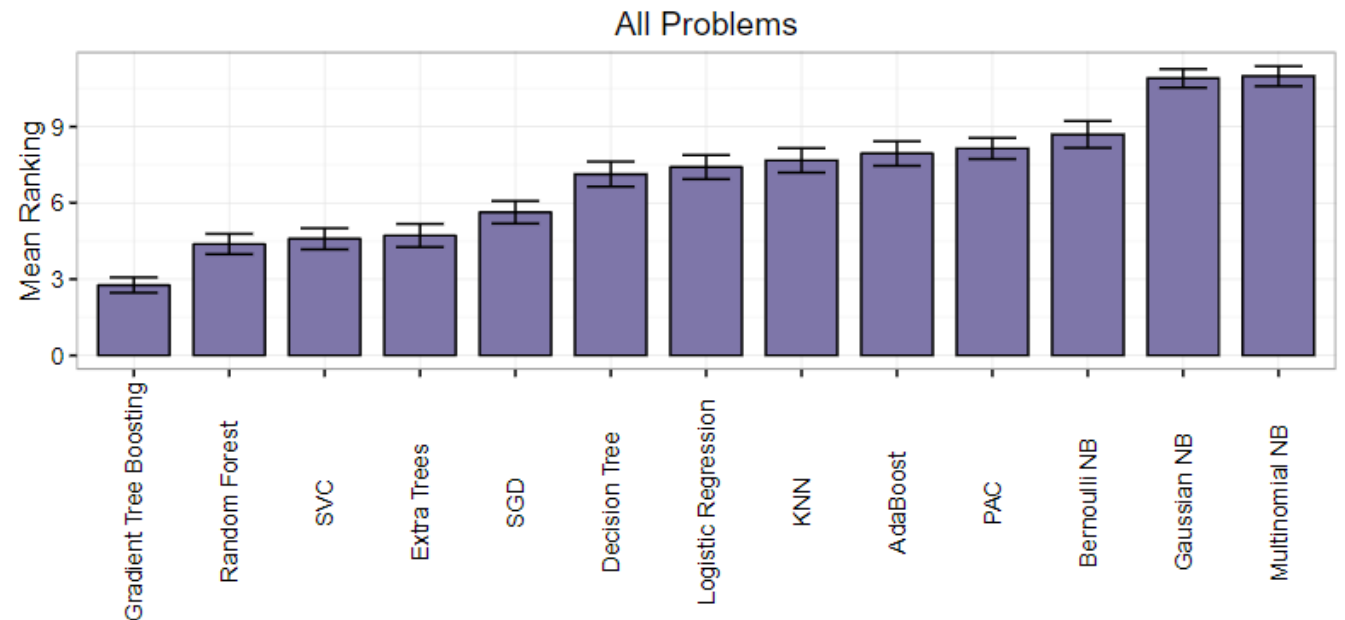


Fig. 1. Average ranking of the ML algorithms over all datasets. Error bars indicate the 95% confidence interval.

Data-driven advice for applying machine learning to bioinformatics

Randal S. Olson, William La Cava, Zairah Mustahsan, Akshay Varik, Jason H. Moore. arXiv, 2018.

- Ganhos de performance ao selecionar hiperparâmetros via validação cruzada de 10-folds (vs. usar o hiperparâmetro default do scikit-learn): média 3-4%, em alguns casos 50%.

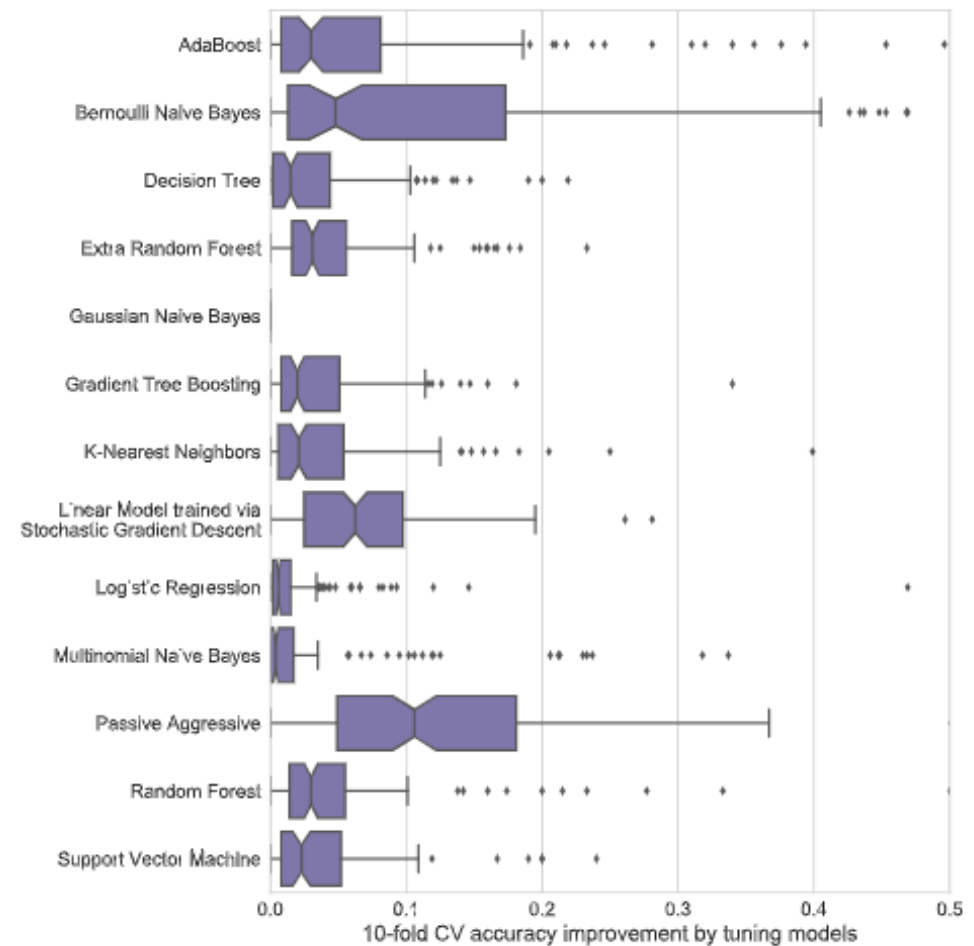


Fig 3. Improvement in 10-fold CV accuracy by tuning each ML algorithm's parameters instead of using the default parameters from scikit-learn

Data-driven advice for applying machine learning to bioinformatics

Randal S. Olson, William La Cava, Zairah Mustahsan, Akshay Varik, Jason H. Moore. arXiv, 2018.

- Ganhos de performance nos 165 bancos de selecionar hiperparâmetros com VC 10-folds + seleção do melhor algoritmo (vs. performance média de todos algoritmos sem tuning): média de 20%, em alguns casos 60%.

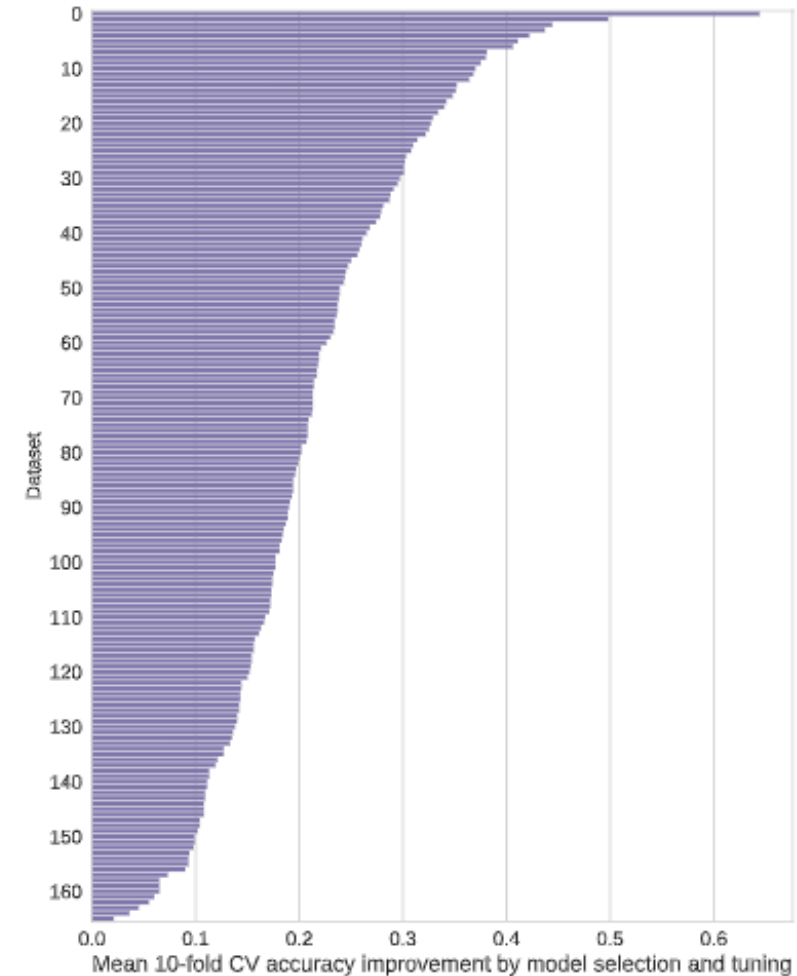


Fig 4 Improvement in 10-fold CV accuracy by model selection and tuning, relative to the average performance on each dataset

TIPOS DE MODELOS PREDITIVOS

Dois grandes grupos

1

Classificação

- Quando a variável a ser predita é categórica:
 - Ex: óbito em 5 anos, incidência de doença em 10 anos, etc.

TIPOS DE MODELOS PREDITIVOS

Dois grandes grupos

2

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)

- 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.

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

MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO



**KEEP
CALM**

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%.

Porém: isso não nos traz nenhuma informação.

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

Ex: Identificar pacientes que possivelmente estão com câncer em amostra que só 1% tem câncer.

Algoritmo: “ninguém tem câncer”! Acurácia = 99%

- 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

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

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

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 = $130/(10+130) = 92,9\%$

MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

Juntar sensibilidade e especificidade num mesmo resultado

Curva ROC

(Receiver Operator Characteristic).

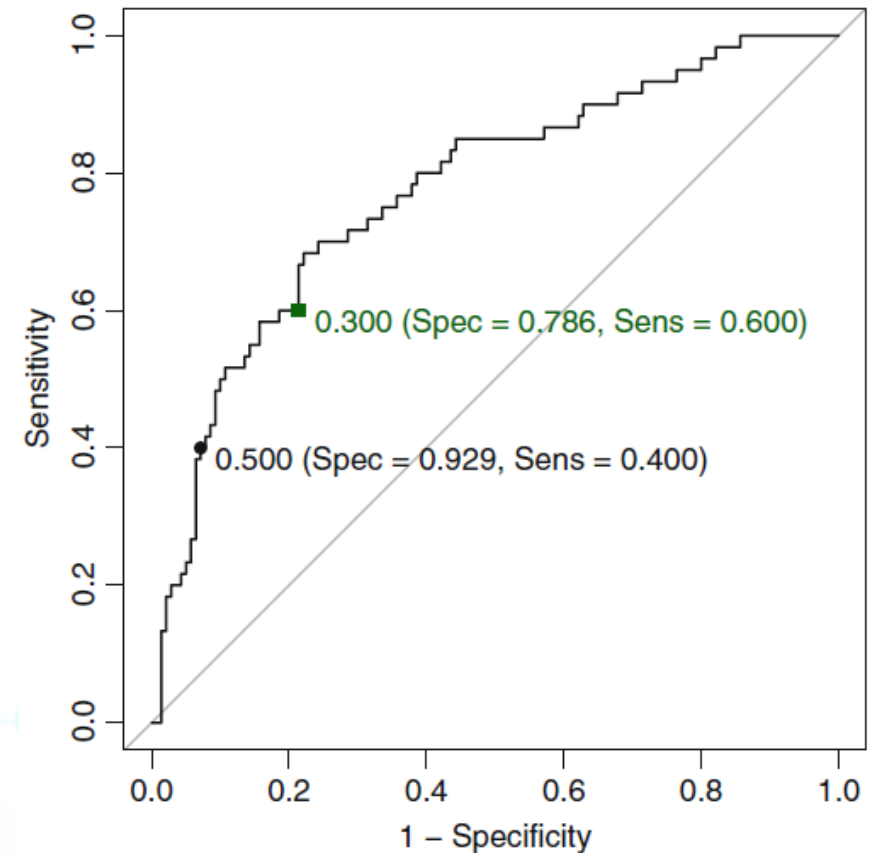
- No exemplo anterior, a sensibilidade foi baixa (40%) e a especificidade foi alta (92,9%).
- Predição sobre câncer foi baseada em o algoritmo dar $> 50\%$ de probabilidade.
- É possível melhorar a sensibilidade diminuindo o threshold?
- Nesse exemplo, sim.
- Threshold de 30% \rightarrow sensibilidade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

MEDIÇÃO DE PERFORMANCE EM PROBLEMAS DE CLASSIFICAÇÃO

- Ideal é uma curva mais à esquerda e para cima possível.
- Linha diagonal 45° → 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.

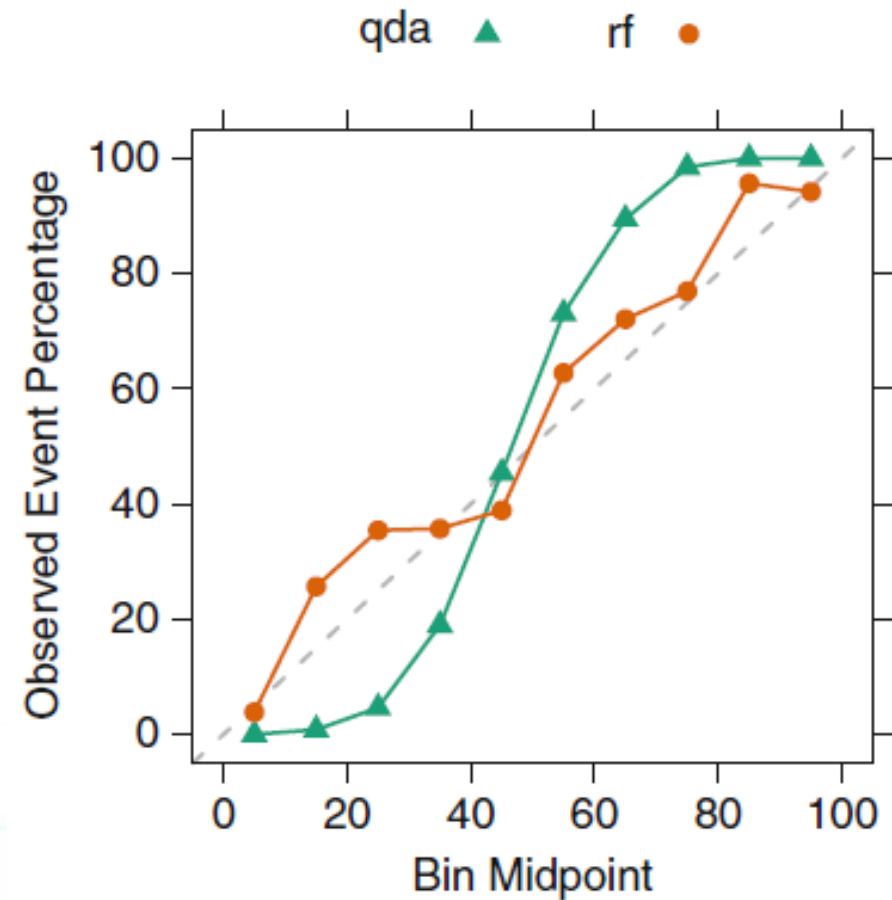
Gráfico de calibração:

- 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?




MOTIVOS PARA UMA PREDIÇÃO RUIM

Se tudo foi feito corretamente e a performance preditiva mesmo assim não foi boa:

- Talvez o evento seja aleatório (algum exemplo?).
- Poucos dados (variáveis ou observações) para o algoritmo aprender os padrões.
- As variáveis não são boas preditoras do desfecho.

MACHINE LEARNING




Não é jogar qualquer coisa no algoritmo e ver no que dá.

É possível fazer isso, mas qualidade preditiva será ruim.

- 1 - Trata-se de um problema de inferência ou predição?
- 2 - Qual o tipo de problema (supervisionado, não supervisionado, semi-supervisionado ou por reforço)?
- 3 - Classificação ou regressão?
- 4 - Selecionar variáveis preditoras plausíveis.
- 5 - Existe vazamento de dados?
- 6 - Padronização de variáveis contínuas.

MACHINE LEARNING



7 – Excluir variáveis altamente correlacionadas?

8 – Utilizar alguma técnica de redução de dimensão?

9 – O que fazer com valores *missing*?


10 – One-hot encoding (ou dummies) para preditores categóricos?

11 – Divisão da amostra em treino e teste (70-30, 80-20 90-10 ou validação)?

12 – Quais valores de hiperparâmetros testar (olhar literatura)?

13 – Validação cruzada? De quantos folds?

MACHINE LEARNING



14 – Quando houver mais de um hiperparâmetro (grid search)?


15 – Quais algoritmos testar (“não há almoço grátis”)?

16 – Para amostras desbalanceadas (classificação), usar técnica de reamostragem (up, down ou SMOTE)?

17 – Como medir performance de problema de regressão (RMSE, R^2 e erro absoluto médio)?

18 – Como medir performance de problema de classificação (acurácia, sensibilidade, especificidade, área abaixo da curva ROC e gráfico de calibração)?

MACHINE LEARNING



Algoritmos

Regressões penalizadas.

Mínimos quadrados parciais.

Support vector machines.

Redes neurais.

Árvores de decisão.

Random forests.

Gradient boosted trees.

Deep learning.

Algoritmos de

MACHINE LEARNING

REGRESSÕES PENALIZADAS

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.

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

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

Algoritmos de

MACHINE LEARNING

REGRESSÕES PENALIZADAS

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

- 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.

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

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

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

- 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

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

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

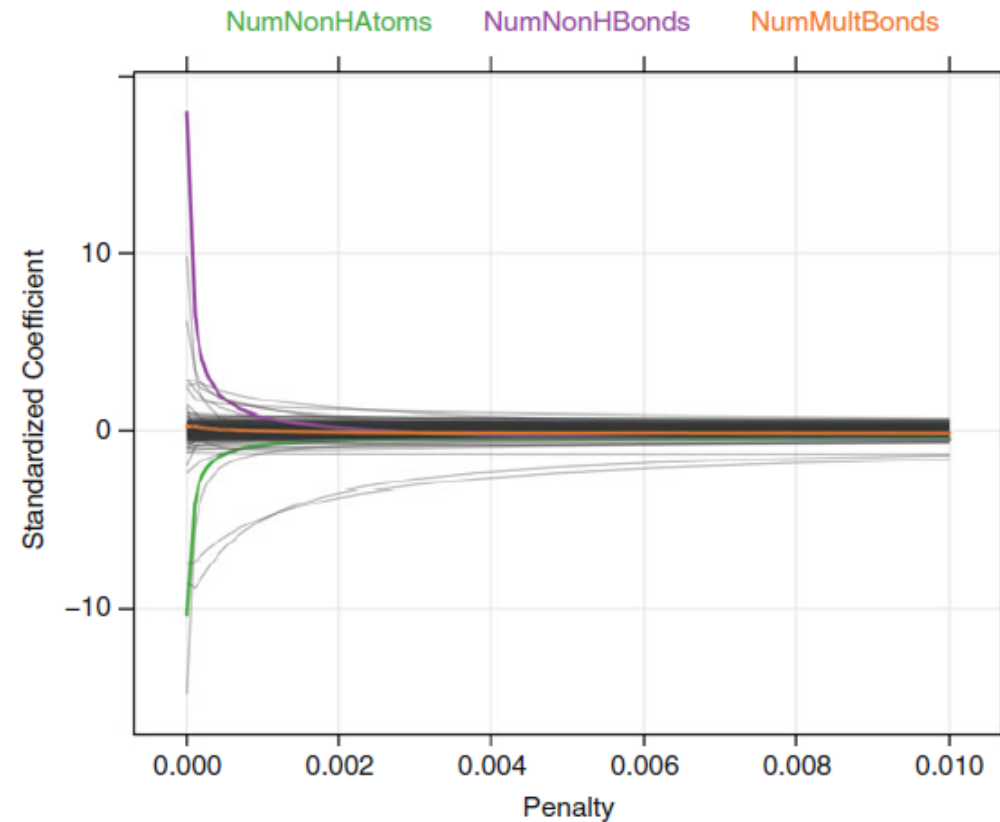
- A quantidade de regularização é controlada pelo parâmetro λ , quanto maior, maior a penalização.
- Ocorre um *encolhimento* dos parâmetros.
- Se $\lambda = 0$ não há penalização, regressão comum.
- Não tem encolhimento no β_0 , queremos diminuir os efeitos das variáveis individuais e não do intercepto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

Algoritmos de

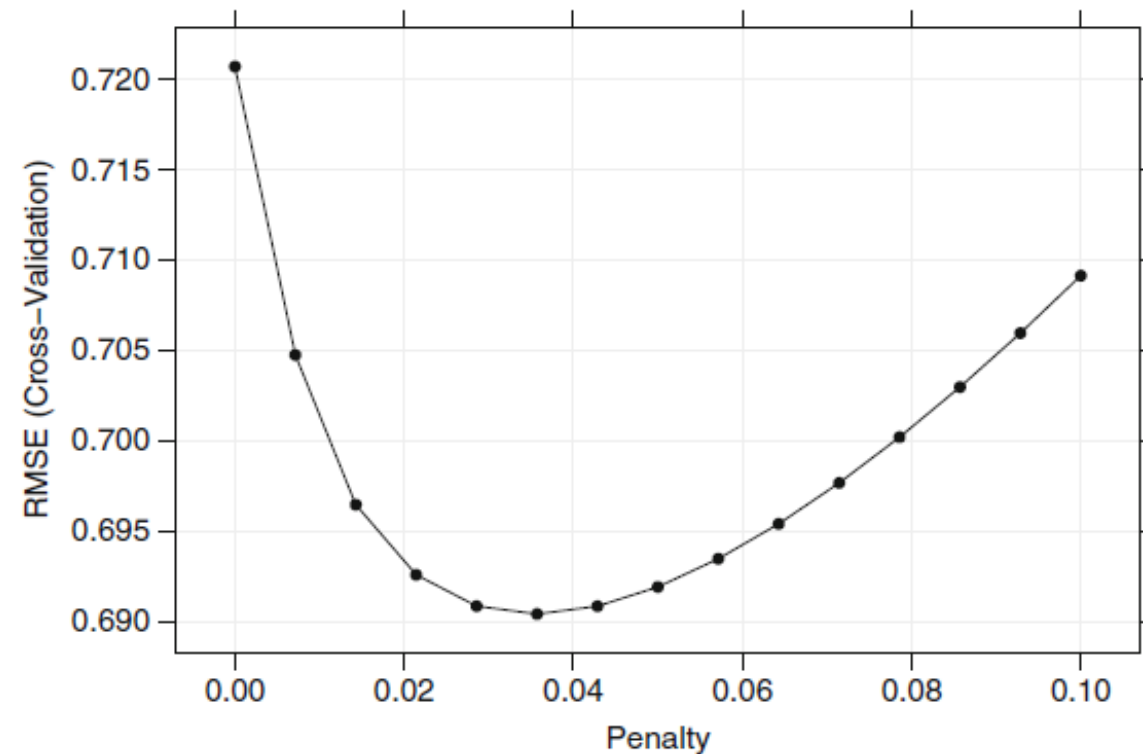
MACHINE LEARNING

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.



- 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.

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

- 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.

Algoritmos de

MACHINE LEARNING

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).