


Machine learning em saúde

Prof. Dr. Alexandre Chiavegatto Filho



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

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

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

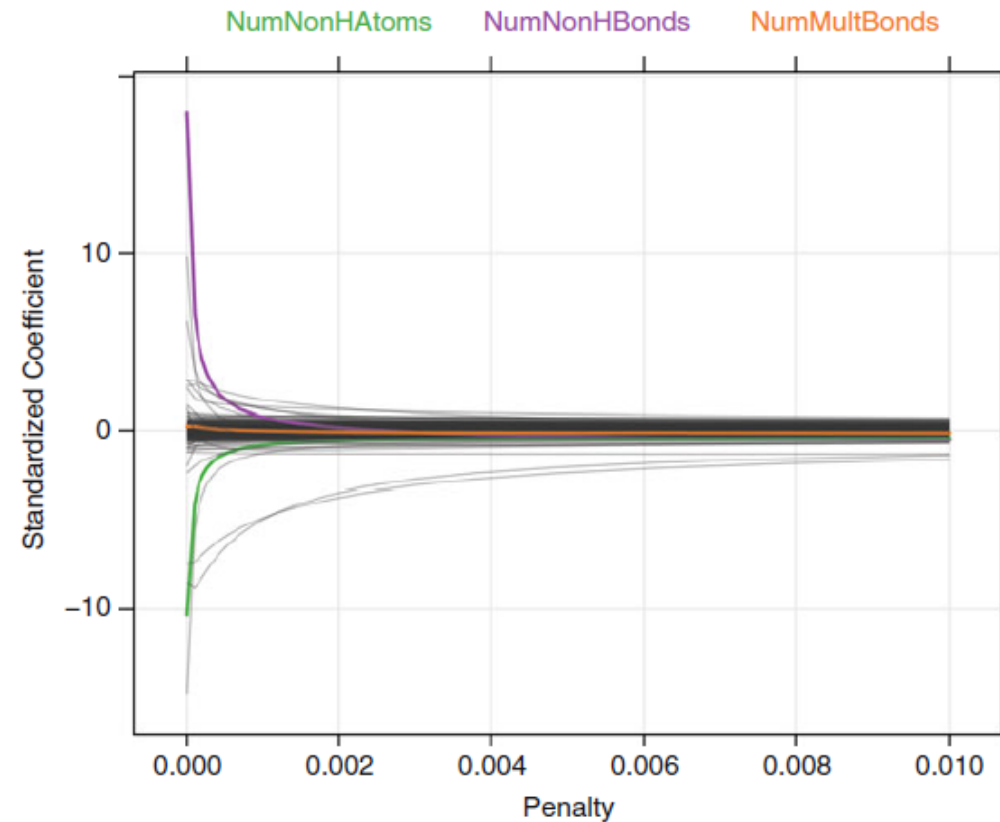
- A quantidade de regularização é controlada pelo parâmetro λ , quanto maior, maior a penalização.
- Ocorre um *encolhimento* dos parâmetros.
- Se $\lambda = 0$ não há penalização, regressão comum.
- Não tem encolhimento no β_0 , queremos diminuir os efeitos das variáveis individuais e não do intercepto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

Algoritmos de

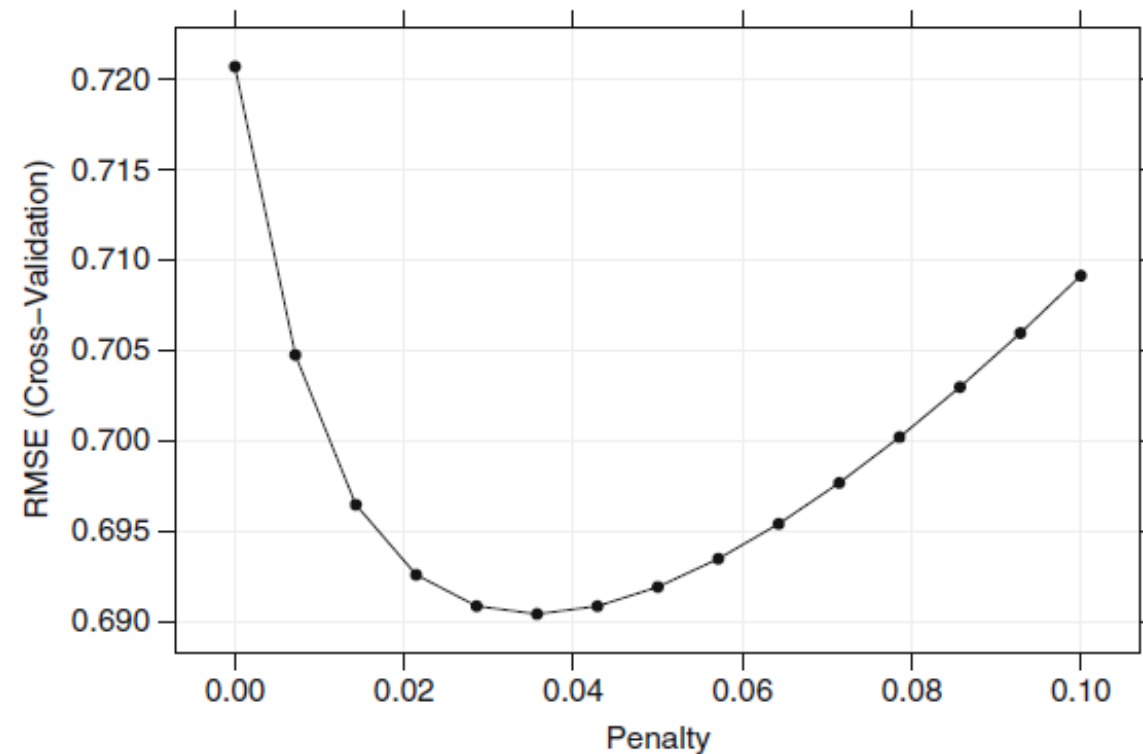
MACHINE LEARNING

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).

- Caso o número de variáveis seja maior que o número de observações, as regressões linear e logística não conseguem chegar a um resultado único.

Soluções

- Remover variáveis.
- Reduzir a dimensão via ACP.
- Incluir os novos componentes como variáveis da regressão linear ou logística (regressão de componentes principais).

- ACP é interessante para redução de dimensão, mas tem a limitação de ser um método não-supervisionado, ou seja, não leva em consideração o desfecho na sua decisão.
- Ou seja, a redução de dimensão que melhor resume os preditores não é necessariamente a melhor opção para explicar o desfecho.

Algoritmos de

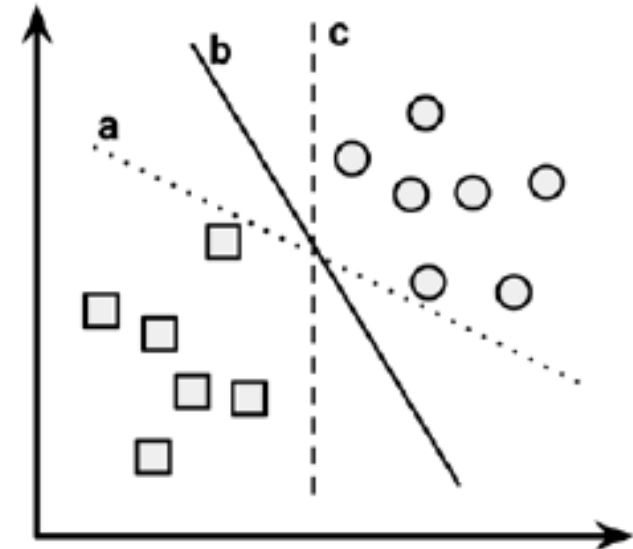
MACHINE LEARNING



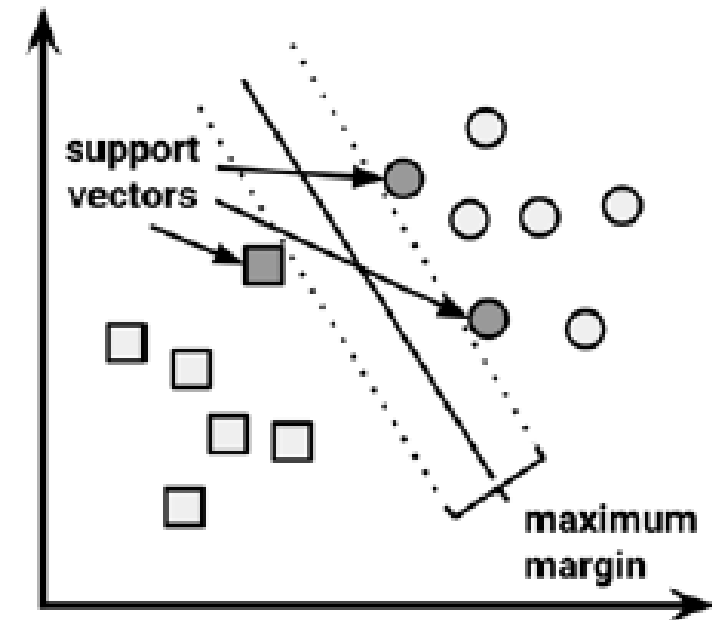
MÍNIMOS QUADRADOS PARCIAIS (PARTIAL LEAST SQUARES)

- PLS é uma alternativa supervisionada à ACP.
- Identifica componentes que maximizam a informação dos preditores, mas ao mesmo tempo garantindo alta correlação com o desfecho.
- Hiperparâmetro: número de componentes.

- Pode ser utilizado para problemas de classificação e regressão (mais intuitivo no caso de classificação, mas os mesmos passos podem ser aplicados para regressão).
- O objetivo é criar uma fronteira, chamado hiperplano, que divida o espaço amostral em áreas parecidas. Não é simples, mesmo em 2D.

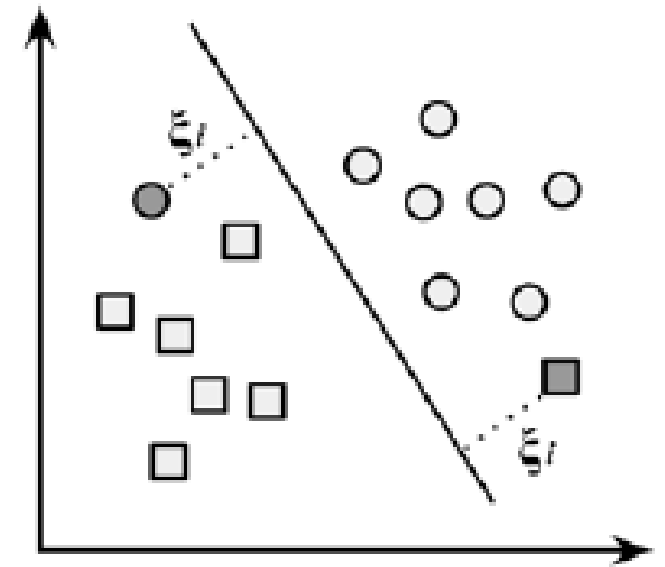


- Solução é identificar o hiperplano de margem máxima (HMM), que encontra a maior separação entre os dois grupos
- Para isso é utilizado o conceito de vetores de apoio (support vectors): os pontos de cada grupo mais próximos do HMM



Caso os grupos não sejam linearmente separáveis

- Utilização de margem fraca: permite que alguns pontos fiquem fora da margem, o que torna menos dependente de poucos pontos.
- Cada ponto fora da margem gera um custo e, em vez de encontrar a maximização da margem, o algoritmo procura minimizar o custo total.

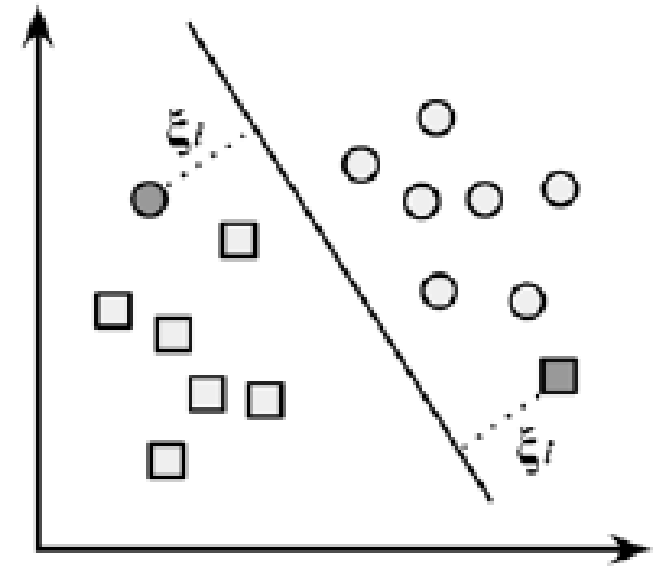


Algoritmos de

MACHINE LEARNING

SUPPORT VECTOR MACHINE

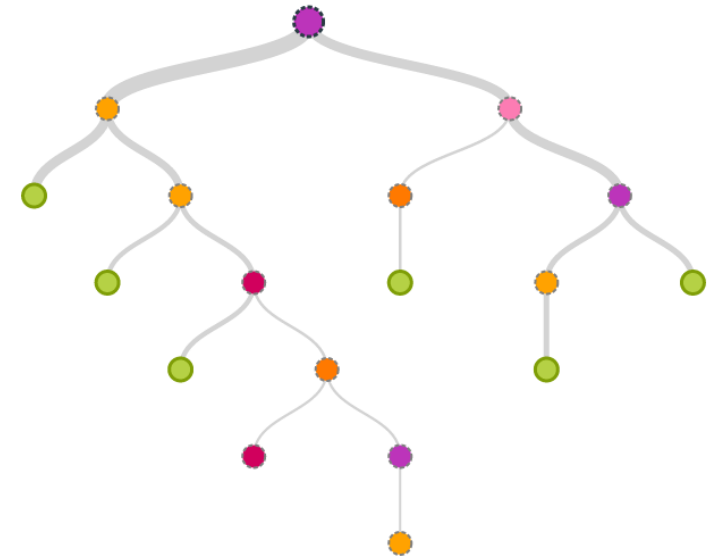
- Hiperparâmetro \rightarrow "c" (penalização pela classificação errada), quanto maior, maior a penalização (aumenta variância).
- O problema pode ser estendido a margens não-lineares (uso de kernels).



Algoritmos de árvore têm como objetivo separar as observações em grupos cada vez menores e mais homogêneos em relação ao desfecho de interesse.

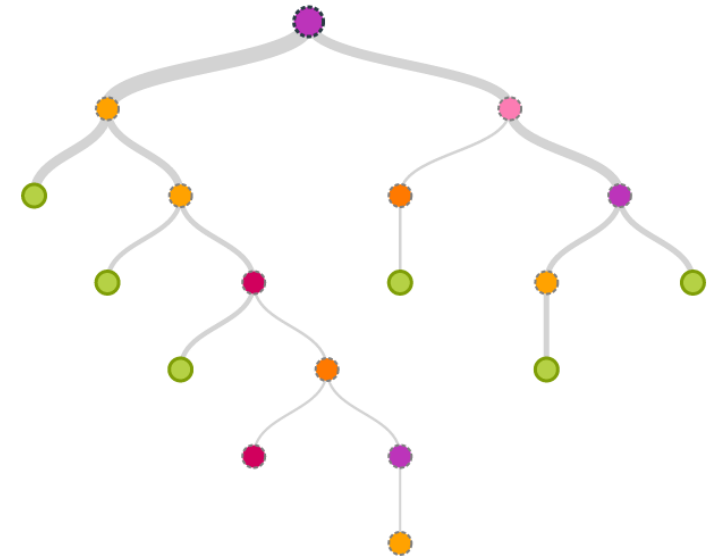
Decisões da árvore:

- Qual preditor e qual valor dividir os dados.
- Profundidade e complexidade da árvore.
- Resultado de cada folha.



Pontos positivos:

- Facilmente interpretáveis.
- Fáceis de implementar.
- Lidam bem com todo tipo de preditor (assimétrico, esparsos, contínuo).
- Realiza seleção de variáveis (não utiliza algumas)

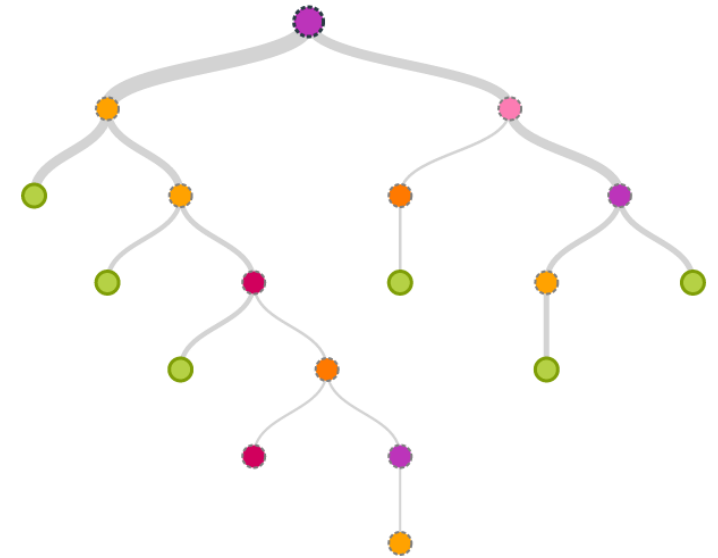


MACHINE LEARNING

ÁRVORES DE DECISÃO

Pontos negativos

- Alta variância / instabilidade.
- Baixa performance preditiva.



Muitas técnicas para criar árvores de regressão.

- Mais comum é CART (Classification and Regression Tree).
- Começa com todo o banco de dados e procura entre todos os valores de todos os preditores a divisão em dois grupos em que:
- Em regressão: minimiza a soma dos erros quadráticos dos dois grupos.

$$\text{SSE} = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2$$

Algoritmos de

MACHINE LEARNING

ÁRVORES DE DECISÃO

Em classificação:

- Contém maior pureza em cada grupo:
 - Maior proporção de categoria igual em um dos lados.

A divisão continua até que seja atingido o critério estabelecido

- Profundidade limite da árvore (parâmetro de complexidade).
- Número mínimo de amostras em um nó.
- Não consegue mais diminuir impureza.

Em problemas de regressão, produz um número limitado de valores preditos, correspondentes ao número de folhas.

- Predição da folha é, em geral, o valor médio das observações dessa folha.

Hiperparâmetros: profundidade máxima da árvore, número mínimo de observações antes da divisão, número máximo de folhas...

- rpart2 usa maxdepth (profundidade máxima da árvore), enquanto rpart usa cp (parâmetro de complexidade – tem de melhorar ajuste em pelo menos o valor de cp para nova divisão).

Maior problema das árvores de decisão simples são a sua forte tendência ao sobreajuste (sensíveis a pequenas variações nos dados).

- A solução mais comum é o uso de bagging (bootstrap aggregation).

Bagging: selecionar amostras aleatórias dos dados de treino.

Cada observação tem a mesma probabilidade de ser sorteada (com reposição).

- Mesmo tamanho dos dados originais, então cada observação tem 63,2% de probabilidade de ser sorteada em cada rodada.

Algoritmos de

MACHINE LEARNING

ÁRVORES DE DECISÃO

- Cada amostra de bootstrap é utilizada para definir a sua árvore.
- Árvores são definidas sem regularização (complexas).
- No caso de regressão, a predição final de uma observação é a sua média de todas as árvores.
- No caso de classificação, a predição final de uma observação é a sua categoria mais predita em todas as árvores (probabilidade é a proporção de votos).
- É considerada uma técnica de ensemble por agregar os resultados de vários algoritmos (de árvore).

Algoritmos de

MACHINE LEARNING

ÁRVORES DE DECISÃO

- O problema do uso de bagging é o custo computacional (porém como as árvores são independentes é fácil paralelizar a análise).
- Pode envolver centenas ou milhares de árvores.
- Uma outra limitação importante é que a facilidade de interpretação, um dos principais pontos fortes das árvores de decisão, foi perdida.

Algoritmos de

MACHINE LEARNING

ÁRVORES DE DECISÃO

Apesar do ganho de variância ao não utilizar determinadas observações em algumas árvores, todos os preditores são analisados em todos os nós:

- Alguns preditores podem ter efeito desproporcional e indesejado nos resultados (utilizados muitas vezes em todas as árvores).
- Maior risco de sobreajuste e semelhança entre as árvores.

- Redução da correlação entre as árvores por meio do uso de um subgrupo aleatório das variáveis preditoras.
- Bagging + seleção de preditores em cada nós = random forests.
- O regularizador das random forests é o número de preditores analisados em cada nó (em geral, a raiz quadrada do total de preditores).

Algoritmos de

MACHINE LEARNING

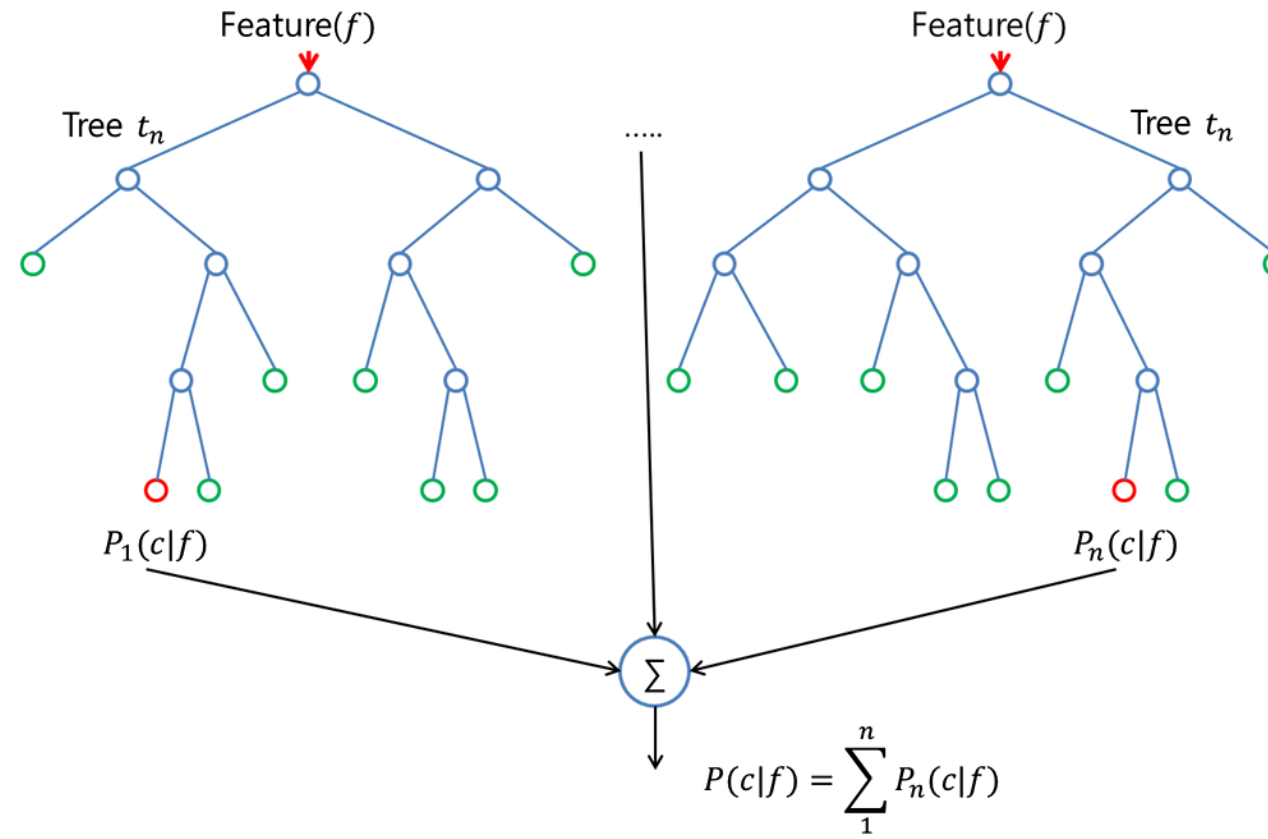
RANDON FORESTS

- Menor custo computacional que bagging (menos variáveis a serem examinadas em cada nó).
- Hiperparâmetro: mtry (número de variáveis consideradas em cada nó).

Algoritmos de

MACHINE LEARNING

RANDON FORESTS



Algoritmos de

MACHINE LEARNING

GRADIENT BOOSTED TREES

- Combinar predições de um conjunto de modelos fracos (taxa de erro ligeiramente melhor que uma classificação aleatória) para construir um comitê poderoso, responsável pela predição final.
- Treinar modelos sequenciais, cada um tentando corrigir o anterior por meio do ajuste dos resíduos.

Algoritmos de

MACHINE LEARNING

GRADIENT BOOSTED TREES

- Atualizar o valor predito de cada observação adicionando o valor predito anterior.
- Muitos hiperparâmetros: 7 no xgbTree (melhoria do ajuste para nova divisão, profundidade máxima, peso das variáveis...)