Introdução à Inteligência Artificial

Roteiro da aula:

- **Árvores de decisão vs modelos lineares**
- **Aprender com Conjunto de Classificadores**
- ***Bootstrapping***
- ***Bagging***
- **Florestas Randômicas**

Métodos Baseados em Árvores. Material baseado, com ilustrações, no Cap. 8 de "Introduction to Statistical Learning", James, Witten, Hastie & Tibshirani, Springer, 2017.

Árvores vs modelos lineares

- Qual modelo seria melhor?

Árvores vs modelos lineares

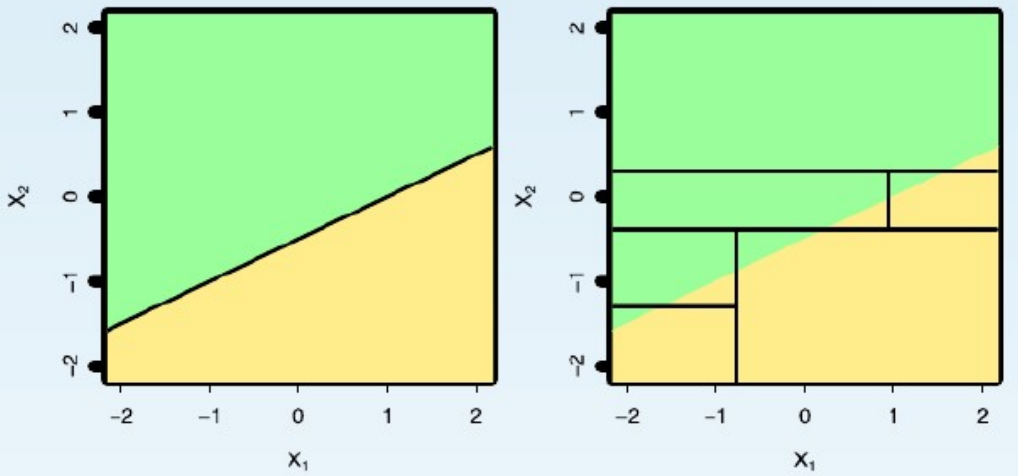
- Qual modelo seria melhor?
  - Se a relação entre os preditores e a resposta for linear, então os modelos clássicos lineares serão melhores que as árvores.

Árvores vs modelos lineares

- Qual modelo seria melhor?
  - Se a relação entre os preditores e a resposta for linear, então os modelos clássicos lineares serão melhores que as árvores.
  - Ao contrário, se a relação entre os preditores for não-linear, então árvores de decisão são melhores que os modelos clássicos lineares.

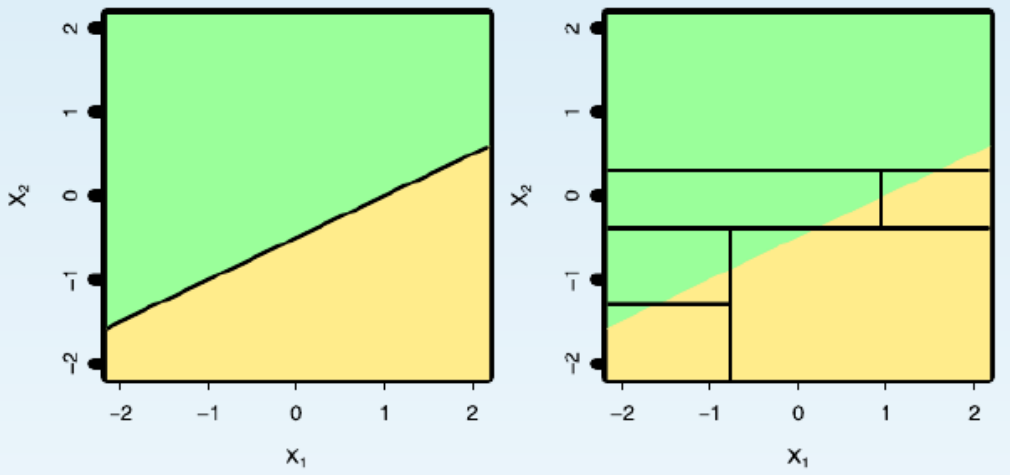
Árvores vs modelos lineares

- Linha sup: o verdadeiro contorno de decisão é linear
  - Esq: modelo linear (bom)
  - Dir: árv. de decisão

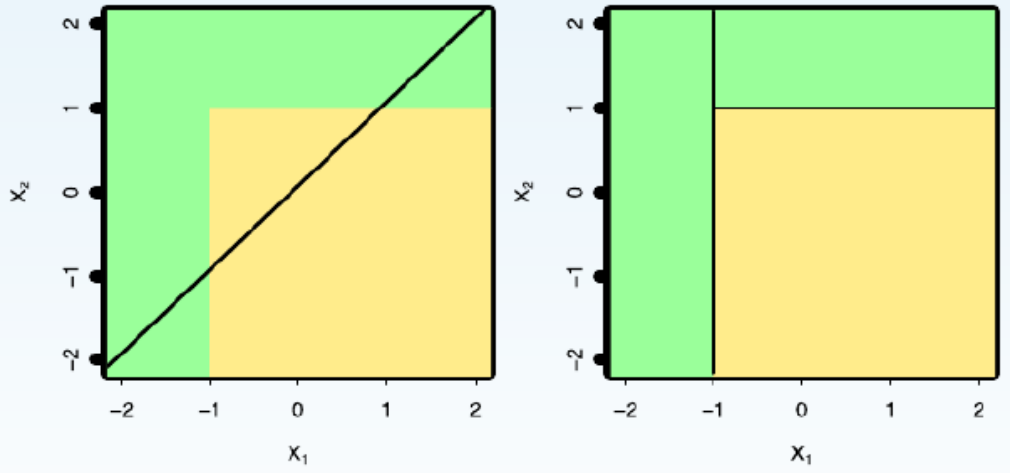


Árvores vs modelos lineares

- Linha sup: o verdadeiro contorno de decisão é linear
  - Esq: modelo linear (bom)
  - Dir: árv. de decisão



- Linha inf: o verdadeiro contorno de decisão é não-linear
  - Esq: modelo linear
  - Dir: árv. de decisão (bom)



Árvores vs modelos lineares

- Prós:
  - Árvores são fáceis de explicar
  - Árvores podem ser visualizadas graficamente e mais facilmente interpretadas
  - Funcionam bem em classificação e regressão

Árvores vs modelos lineares

- Prós:
  - Árvores são fáceis de explicar
  - Árvores podem ser visualizadas graficamente e mais facilmente interpretadas
  - Funcionam bem em classificação e regressão
- Contras:
  - Árvores não possuem as melhores acurácias de predição como em modelos mais complexos

Conjunto de métodos (ensemble)

- Uma única árvore de decisão nem sempre atinge um bom desempenho
- E se aprendermos múltiplas árvores?

Aprender em conjunto...

- Considerando um conjunto de classificadores $h_1, \ldots, h_L$
- A ideia-chave é construir um classificador $H(\boldsymbol{x})$

  que combine as predições individuais de $h_1, \ldots, h_L$

Aprender em conjunto...

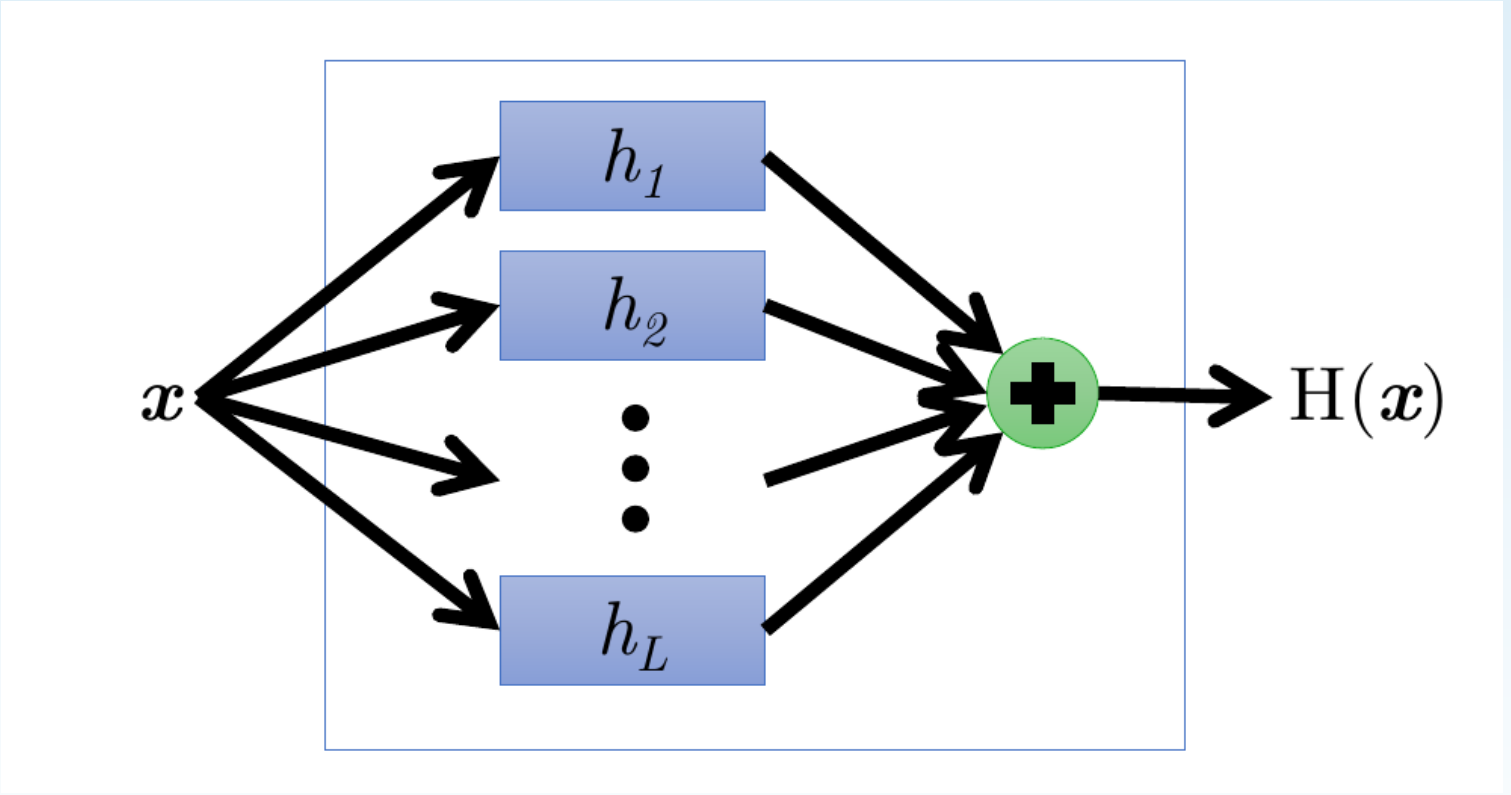
- **Aprendizagem de conjunto requer diversidade para melhor desempenho**

Aprender em conjunto...

- **Aprendizagem de conjunto requer diversidade para melhor desempenho**
  - **Classificadores individuais devem ter erros diferentes (i.e. não correlacionados)**

Combinando classificadores

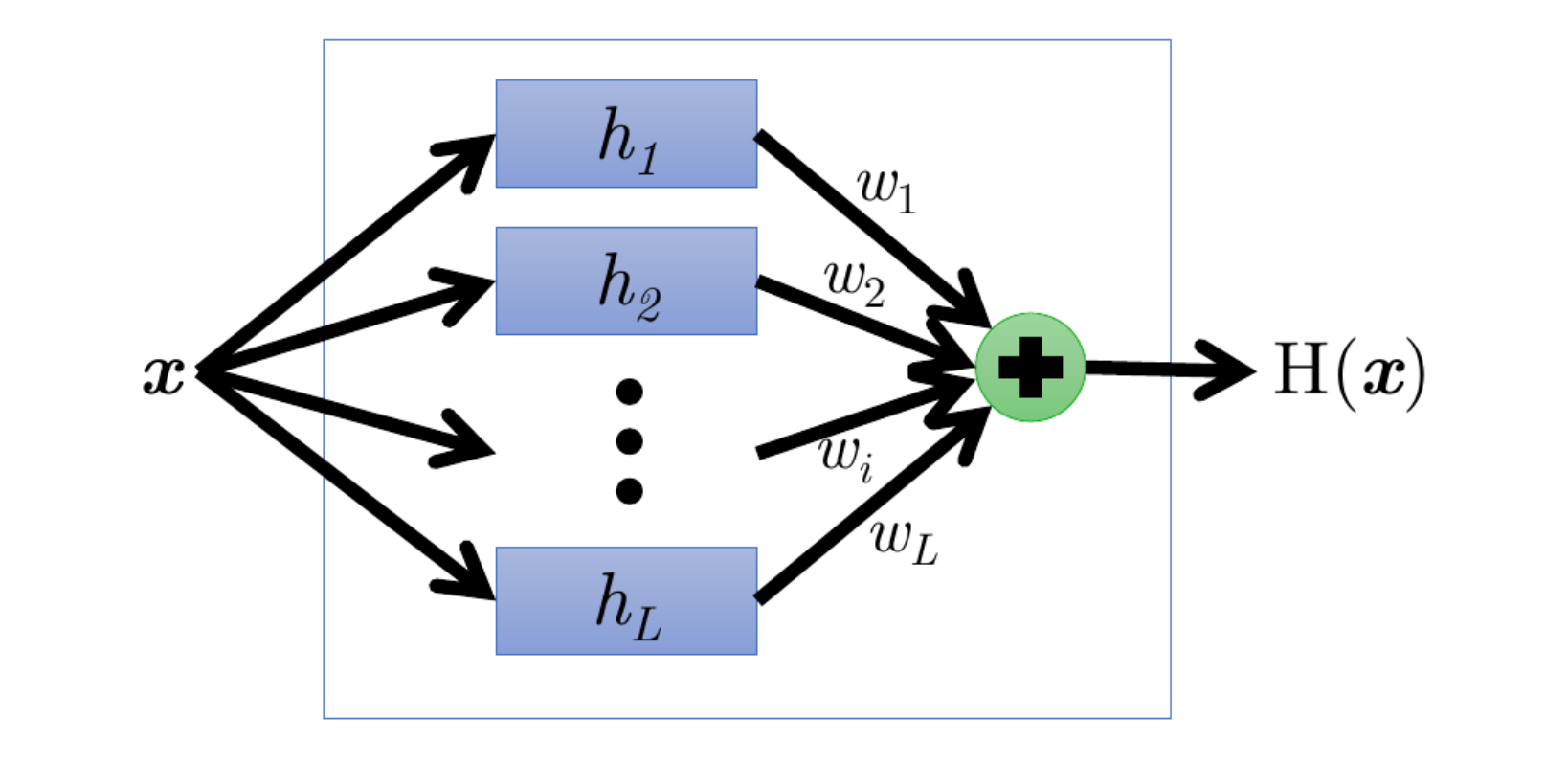
- **Uma forma de combinar seria por voto simples dos membros e uma decisão final**



image/credit: E.Eaton

Combinando classificadores

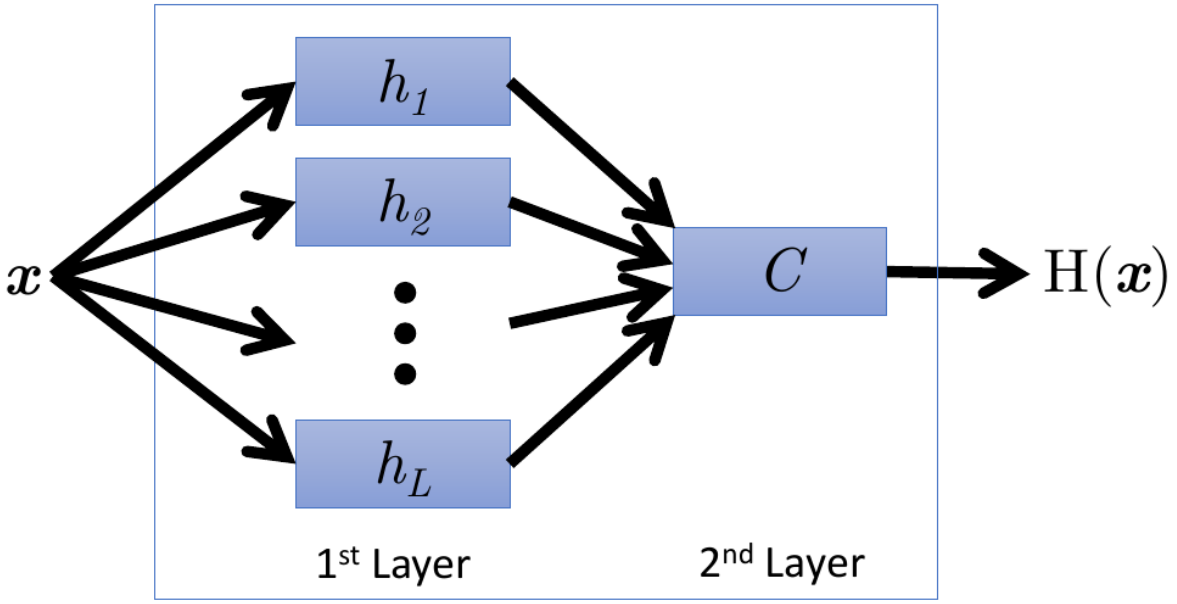
- **Uma forma de combinar seria pela média ponderada**



image/credit: E.Eaton

Combinando classificadores

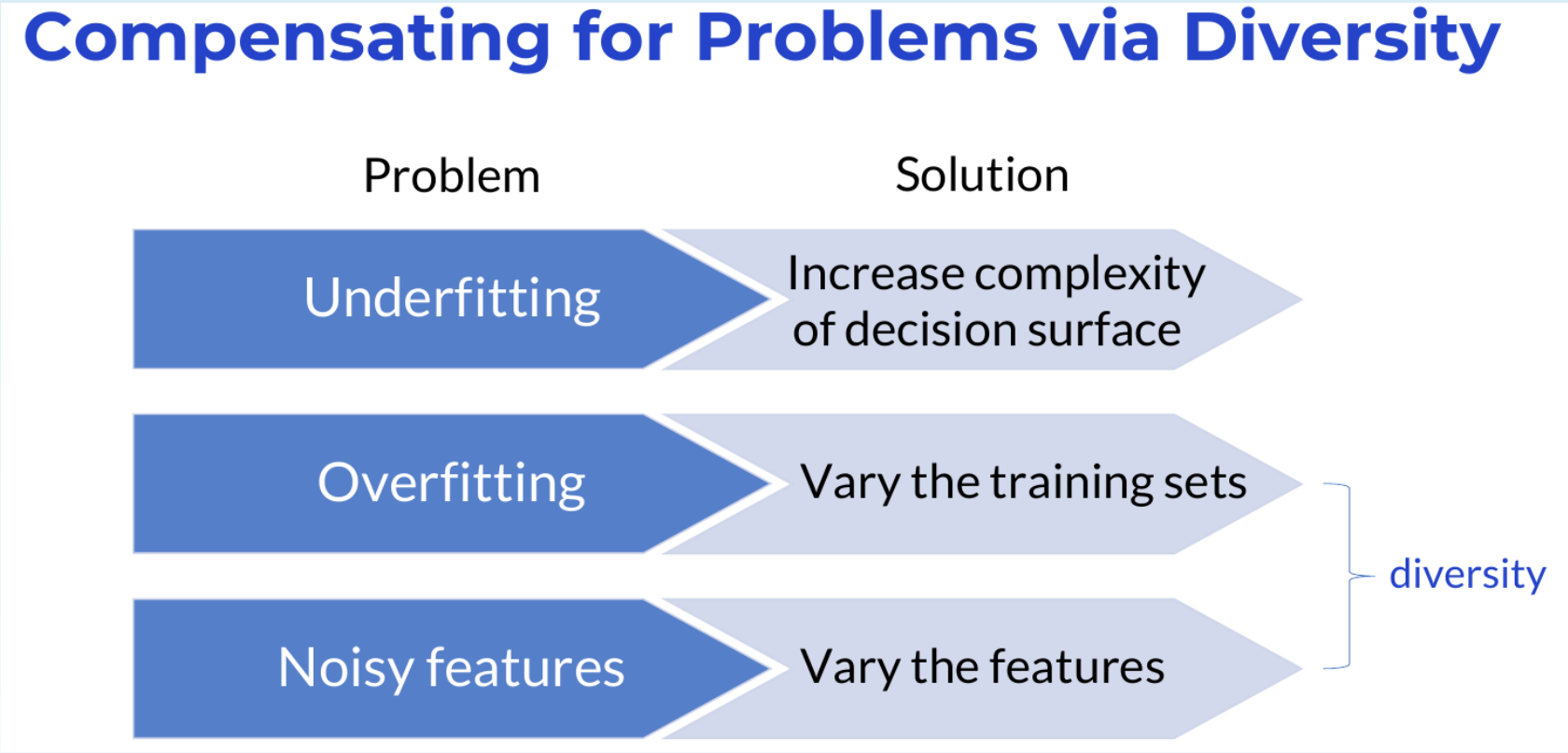
- **Uma forma de combinar seria em sequência (camadas) de classificadores**



- Predictions of 1st layer used as input to 2nd layer
- Train 2nd layer on validation set

image/credit: E.Eaton

Formas para induzir diversidade: *Bootstrapping and Bagging*



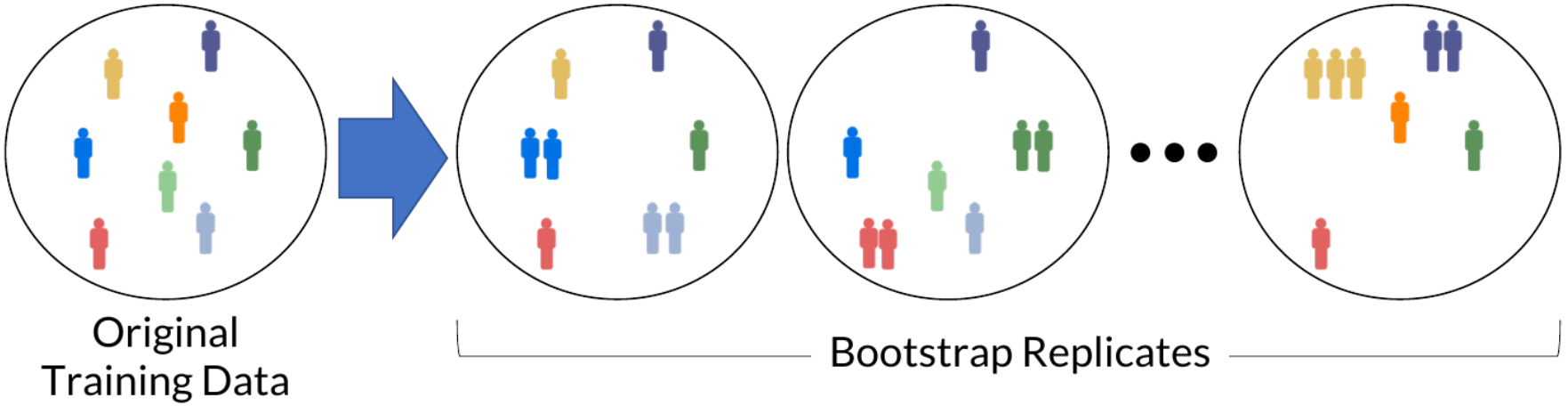
image/credit: E.Eaton

Boostrapping

Varying the Training Data

Bootstrap replication:

- Given n training instances, construct new training sets by sampling n instances with replacement
  - Excludes ~30% of the training instances in each of the replicates



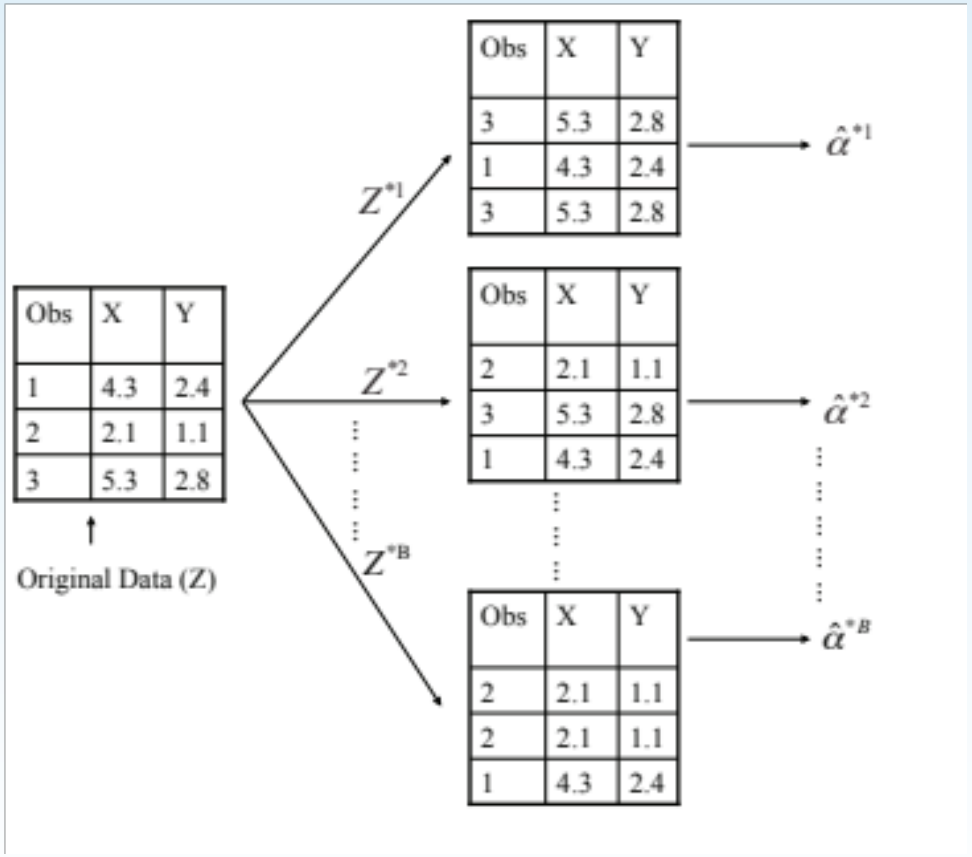
image/credit: E.Eaton

Boostrapping

- **Reamostragem do conjunto de dados observados (e de tamanho igual ao conjunto observado), cada desses é obtido por amostragem aleatória com reposição a partir do conjunto original.**

Boostrapping é simples

- **Reamostragem do conjunto de dados observados (e de tamanho igual ao conjunto observado), cada desses é obtido por amostragem aleatória com reposição a partir do conjunto original.**



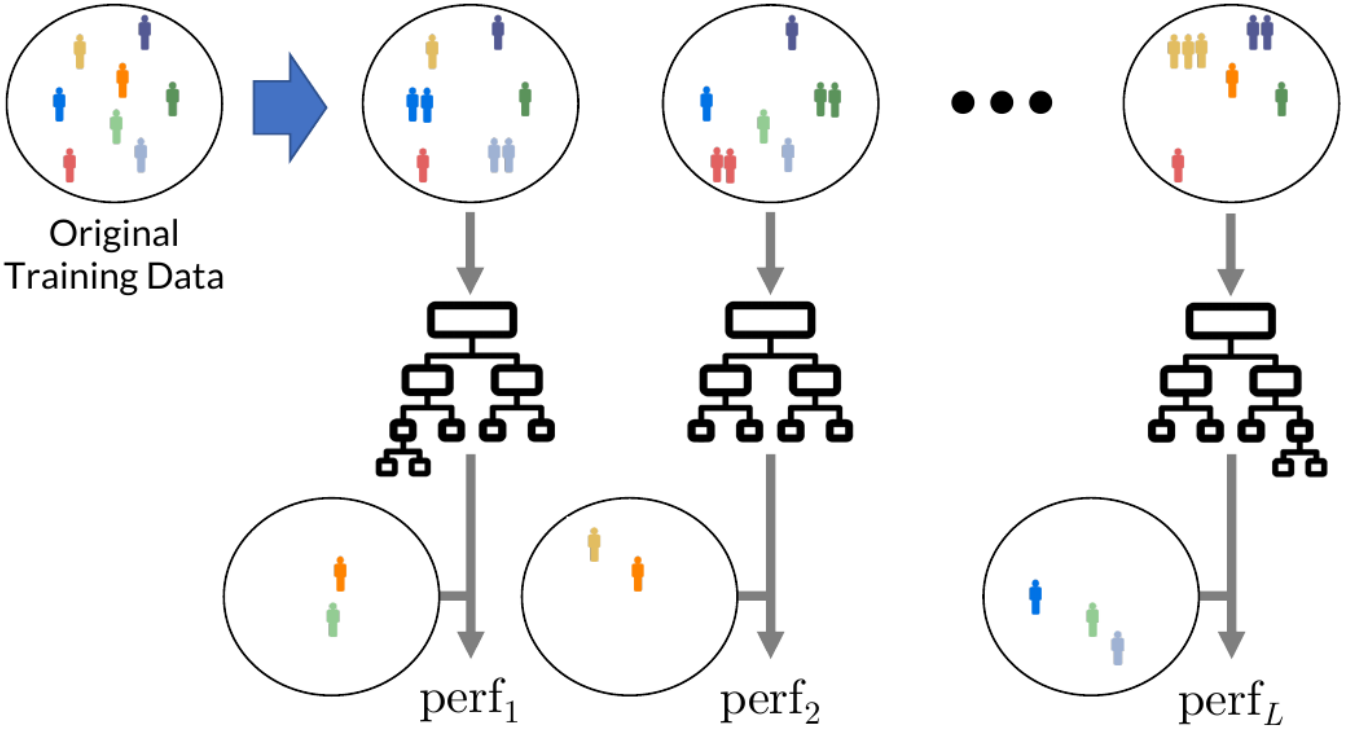
Bagging

Manipulating the Training Data: Bagging

1.) Create bootstrap replicates of training set

2.) Train a classifier for each replicate

3.) Estimate classifier performance using out-of-bootstrap data



4.) Create a weighted ensemble of the classifiers, with weights based on $\text{perf}_1, \ldots \text{perf}_L$

image/credit: E.Eaton

Bagging

- **Árvores de decisão podem sofrer de alta variância!**
  - **Se dividirmos os dados de treinamento aleatoriamente em 2 partes, e ajustar árvores de decisão em ambas, os resultados podem ser bem diferentes**

Bagging

- **Árvores de decisão podem sofrer de alta variância!**
  - **Se dividirmos os dados de treinamento aleatoriamente em 2 partes, e ajustar árvores de decisão em ambas, os resultados podem ser bem diferentes**

- **Gostaríamos de ter modelos com baixa variância**

Bagging

- **Árvores de decisão podem sofrer de alta variância!**
  - **Se dividirmos os dados de treinamento aleatoriamente em 2 partes, e ajustar árvores de decisão em ambas, os resultados podem ser bem diferentes**
- **Gostaríamos de ter modelos com baixa variância**
- **Para resolver esse problema, usa-se bagging (*b*ootstrap *agg*regat*ing*).**

O que é *bagging?*

- ***Bagging*** **é baseada em duas coisas:**
  - **Média: reduz variância!**
  - ***Bootstrapping*****: muitos conjuntos de dados de treinamento!**
- **Porque média reduz variância?**

O que é *bagging?*

- ***Bagging*** **é baseada em duas coisas:**
  - **Média: reduz variância!**
  - ***Bootstrapping*****: muitos conjuntos de dados de treinamento!**
- **Porque média reduz variância?**
  - **Média de um conjunto de observações reduz variância. Relembrar que dado um conjunto de n observações independentes Z_1, …, Z_n, cada com variância σ^2 ,**
    a variância da média $\overline{Z}$ das observações é σ^2/n

Como *bagging* funciona?

- Gerar B diferentes conjuntos de treinamento *bootstrapped*
- Treinar o método de aprendizagem em cada dos B conjuntos de treinamento, e obter a predição
- Para predição:
  - Regressão: média de todas as predições de todas as B árvores
  - Classificação: voto majoritário entre todas as B árvores

Bagging para Árvores de Classificação

- **Construir B árvores de regressão usando B conjuntos de treinamento *boostrapped***
- **Para predição, há duas abordagens:**

Bagging para Árvores de Classificação

- **Construir B árvores de regressão usando B conjuntos de treinamento *boostrapped***
- **Para predição, há duas abordagens:**
  1. **Gravar a classe que cada conjunto de dados *bootstrapped* prediz e proporcionar uma predição para a mais comum (voto majoritário).**

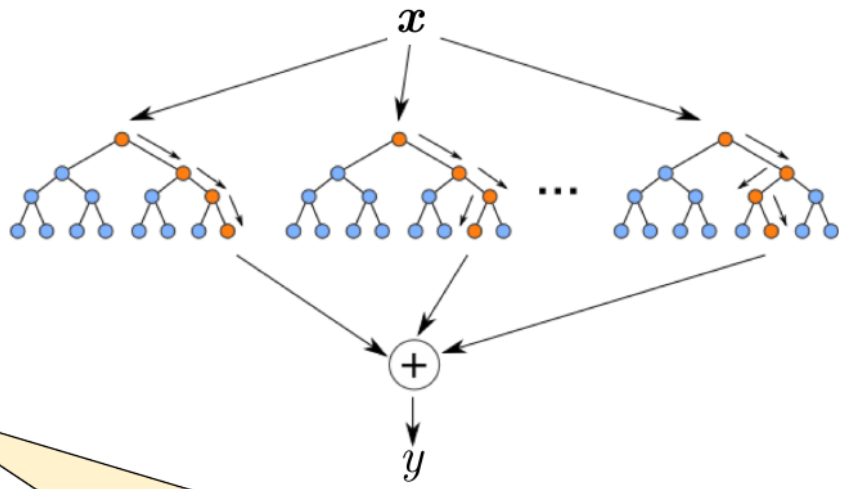
Bagging para Árvores de Classificação

- Construir B árvores de regressão usando B conjuntos de treinamento *boostrapped*
- Para predição, há duas abordagens:
  1. Gravar a classe que cada conjunto de dados bootstrapped prediz e proporcionar uma predição para a mais comum (voto majoritário).
  2. Se o classificador produz estimativas de probabilidade pode-se fazer a média das probabilidades e predizer a classe com maior probabilidade.
- Ambas opções funcionam bem.

Classificador “Florestas Randômicas”

Random Forest Classifier

- Manipulates both training data and the features to induce diversity
  - **Training data manipulation**: bagging to create an ensemble of decision trees
  - **Feature manipulation**: Each decision tree node focuses on a subset of features, chosen randomly at each node



Issue with only bootstrapping DTs

- If a few features are highly predictive, then they will be selected in many trees
- This will cause the ensemble members to become correlated

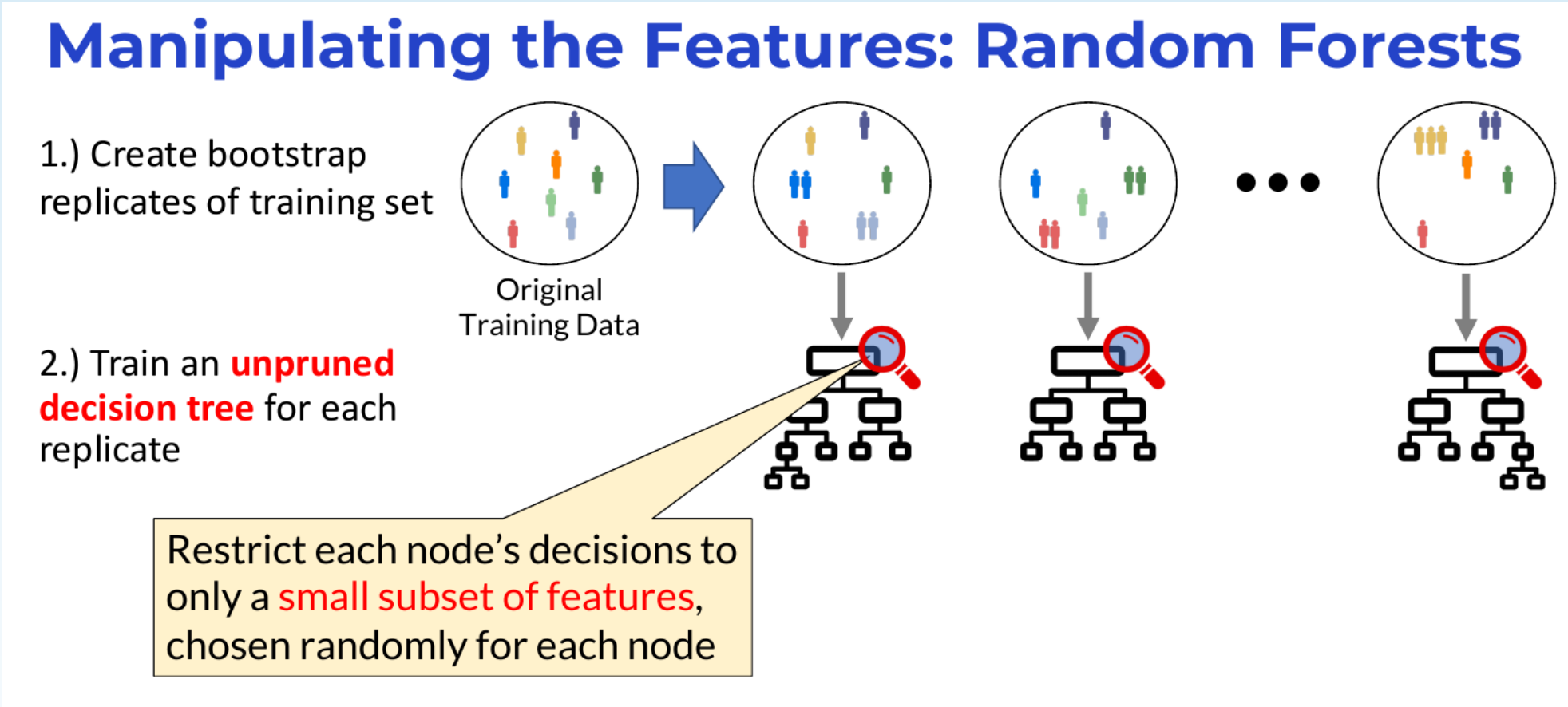
DT node creation procedure for a RF:

1.) At each node, choose $\sqrt{d}$ features randomly to consider

2.) Determine split among these features

image/credit: E.Eaton

Classificador “Florestas Randômicas”

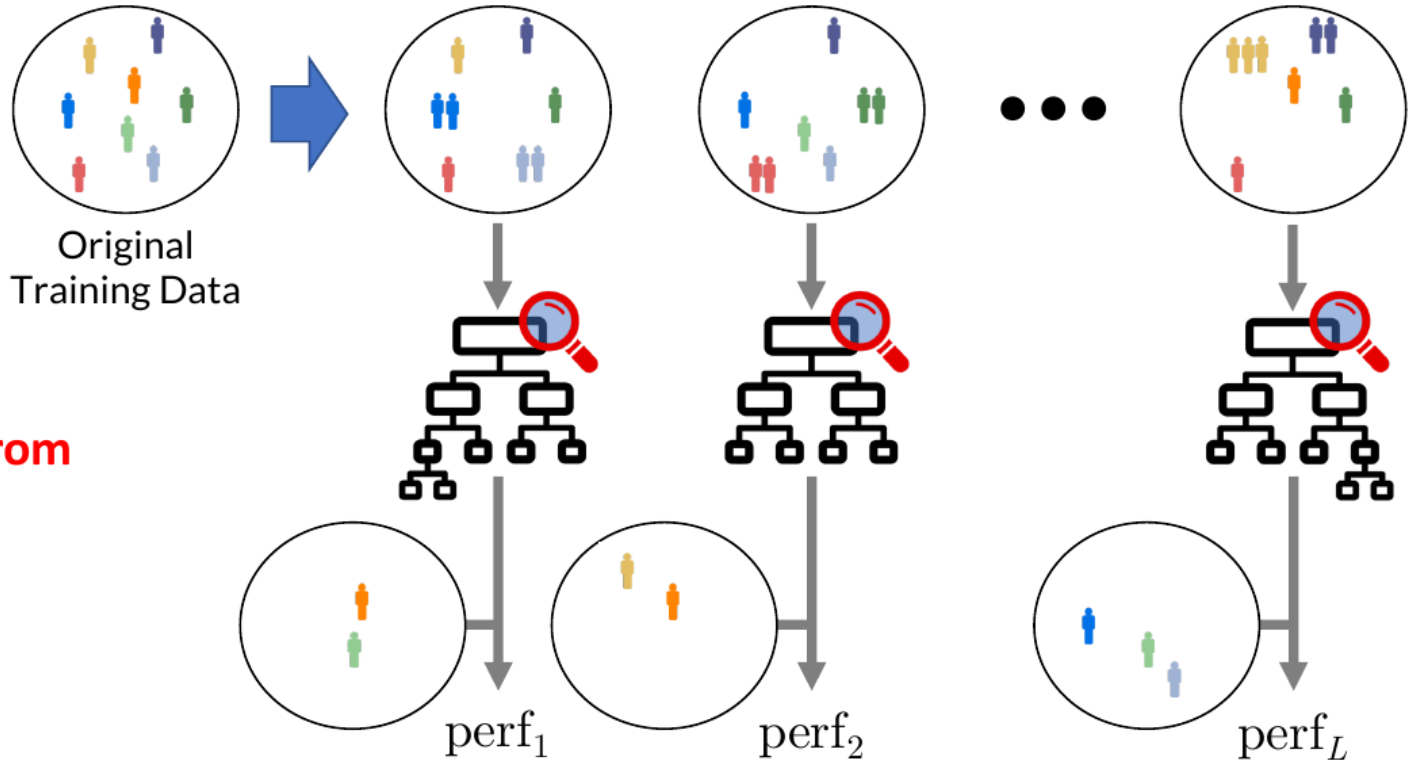


image/credit: E.Eaton

Classificador “Florestas Randômicas”

Manipulating the Features: Random Forests

1.) Create bootstrap replicates of training set



2.) Train an **unpruned decision tree** for each replicate; **splits chosen from random feature subsets**

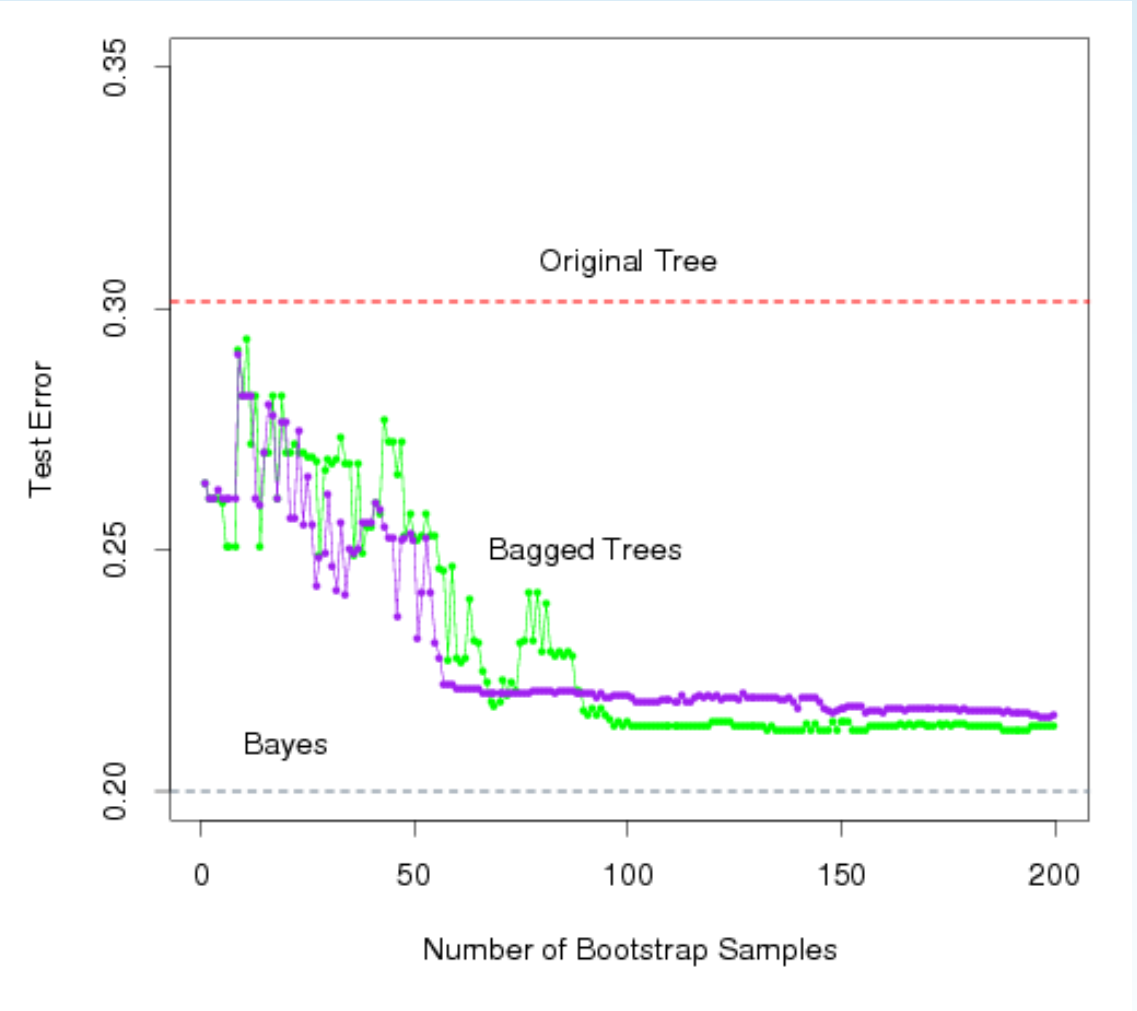
3.) Estimate classifier performance using out-of-bootstrap data

4.) Create a weighted ensemble of the classifiers, with weights based on $\text{perf}_1, \ldots \text{perf}_L$

image/credit: E.Eaton

Uma comparação de taxas de erro

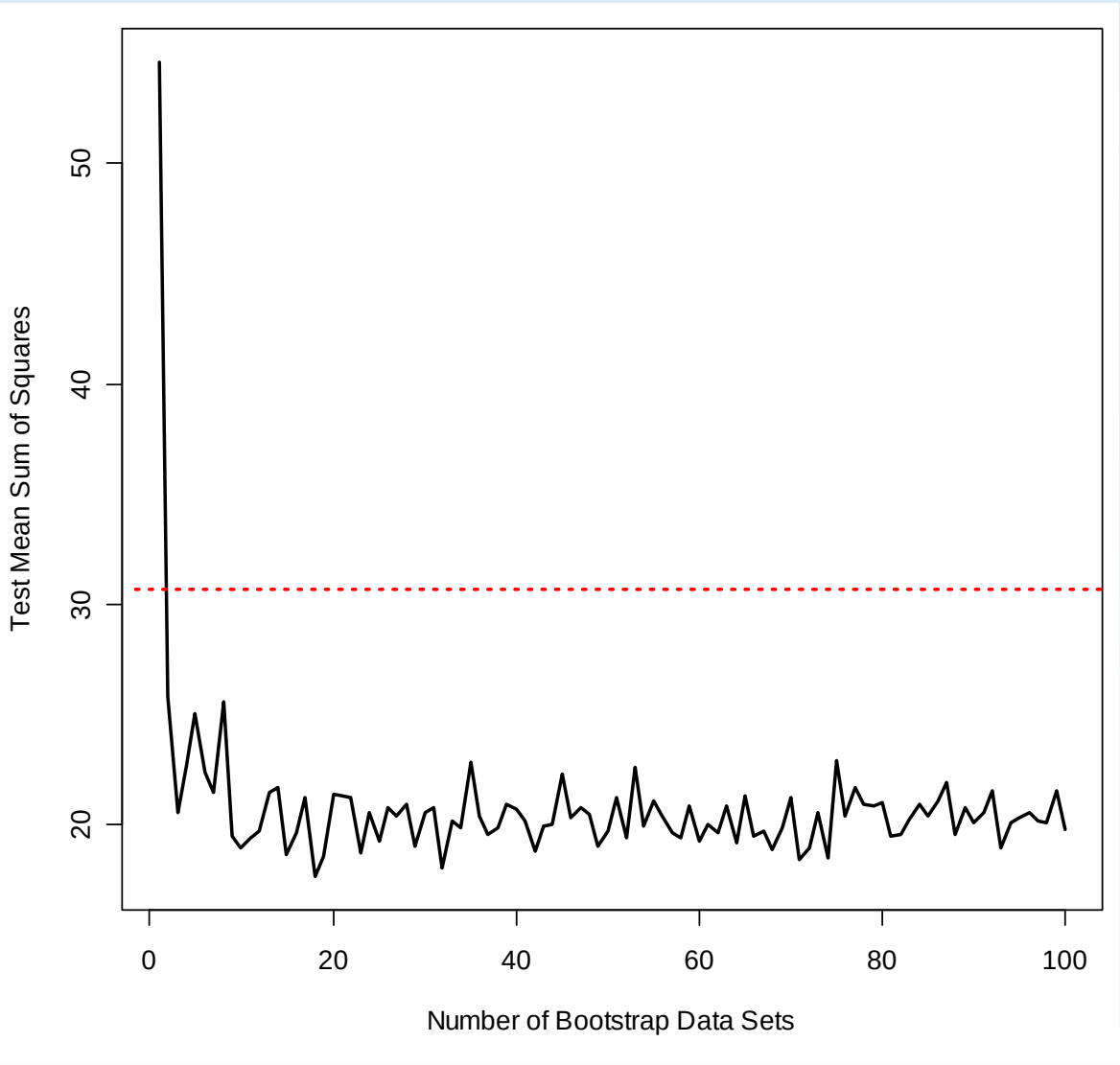
- A linha verde representa uma abordagem de voto majoritário
- A linha roxa corresponde à média das estimativas de probabilidade.
- Ambos são melhores que uma única árvore (vermelho pontilhado) e ficam próximas à taxa de erro de Bayes (cinza pontilhado).



image/credit: James et. al. 2017

Exemplo: *Housing data*

- **A linha vermelha representa a média de erro de uma única árvore.**
- **A linha preta corresponde a taxa de erro de *bagging***



image/credit: James et. al. 2017

Estimação de erro (*Out-of-Bag*)

- **Como *bootstrapping* envolve a seleção aleatória de subconjuntos de observações para construir um conjunto de treinamento, então a parte restante não selecionada pode ser dados de teste.**
- **Na média, cada árvore *bagged* faz uso de 2/3 das observações, então acaba-se tendo 1/3 das observações para teste.**

Medidas de importância da variável

- ***Bagging*** **tipicamente melhora a acurácia de predição de uma única árvore, mas com interpretação mais difícil!**

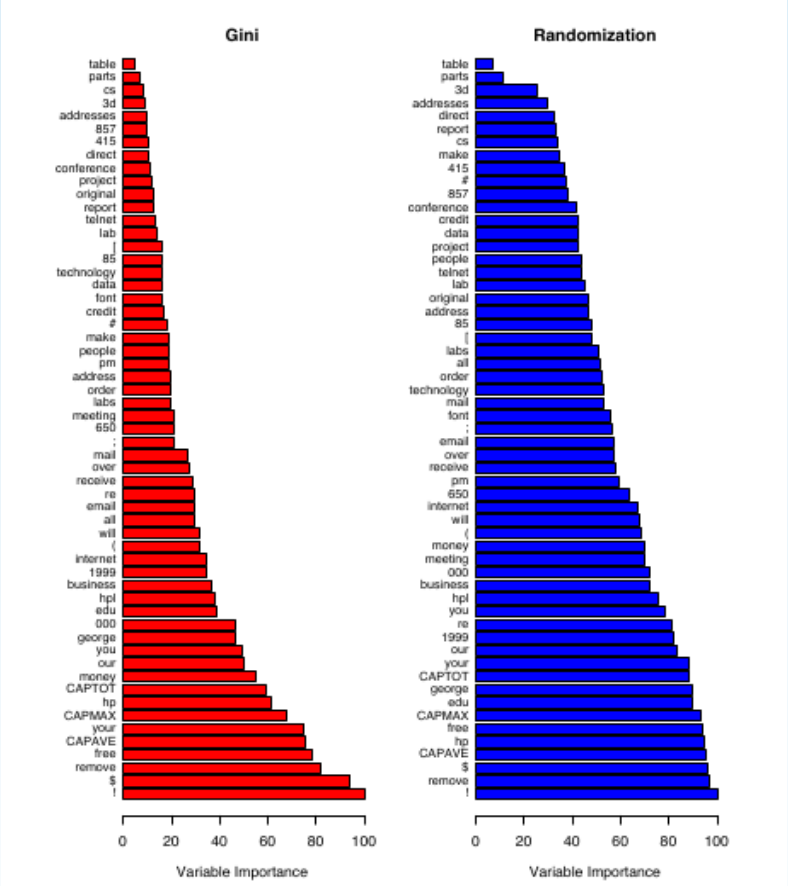
- **Mas, podemos obter um resumo da importância de cada preditor usando Gráficos de Influência Relativa**

Gráficos de influência relativa

- **Como decidir quais variáveis são mais úteis em predizer a resposta?**
  - **Computando gráficos de influência relativa.**
  - **Esses gráficos fornecem um *score* para cada variável.**
  - **Esses *scores* representam o decréscimo no MSE quando dividindo uma variável em particular**
  - **Um número próximo a zero indica que a variável não é importante e poderia ser descartada.**
  - ***Scores* mais altos indicam variáveis de maior influência.**

Influência relativa

- **As medidas de influência relativa podem ser diferentes**



Influência relativa (ex.)

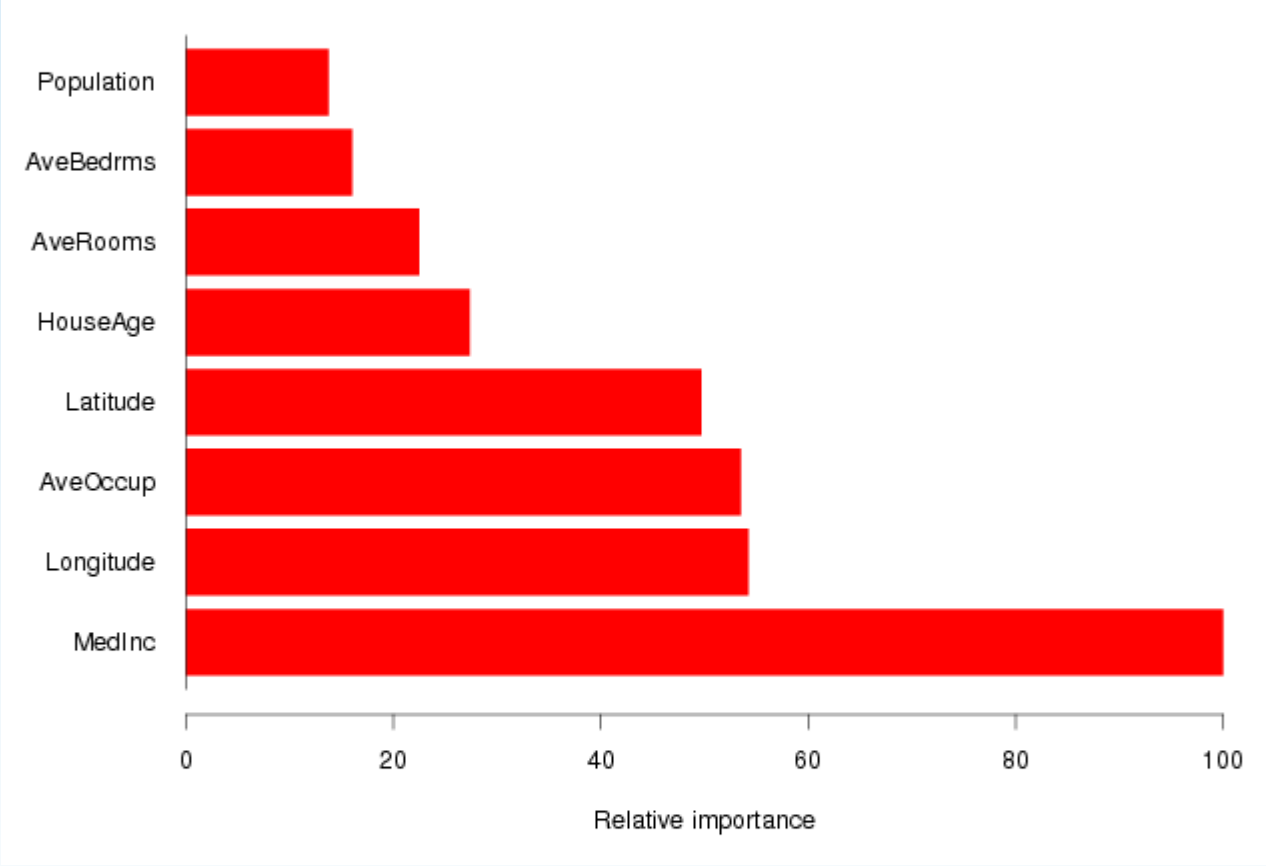
- **Dicas de código na biblioteca scikit-learn e variações das medidas de influência/importância dos atributos em árvores randômicas**

https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Exemplo: *Housing data*

- **Renda mediana é a variável mais importante.**

- **Longitude, Latitude e ocupação média são as próximas mais importantes.**



image/credit: James et. al. 2017

Florestas Randômicas

- Queremos variância menor
  - E se consideramos somente um subconjunto dos preditores a cada divisão?
  - Árvores geradas ainda serão correlacionadas, a menos que

Florestas Randômicas

- Queremos variância menor
  - E se consideramos somente um subconjunto dos preditores a cada divisão?
  - Árvores geradas ainda serão correlacionadas, a menos que

    selecionarmos o subconjunto randomicamente

Florestas Randômicas

- **Método de aprendizagem eficiente**
- **Baseado na ideia de *bagging*, mas proporciona uma melhora pois descorrelaciona as árvores**
- **Como funciona?**

Florestas Randômicas

- **Método de aprendizagem eficiente**
- **Baseado na ideia de *bagging*, mas proporciona uma melhora pois descorrelaciona as árvores**
- **Como funciona?**
  - **Construir um número de árvores de decisão em amostras de treinamento *bootstrapped*, mas ao construir essas árvores, cada vez que uma divisão em uma árvore for considerada, uma amostra randômica de *m* preditores é escolhida como candidatos de divisão do conjunto total de *p* preditores (Usualmente $m \approx \sqrt{p}$)**

Florestas Randômicas

Porque considerar uma amostra randômica de m preditores ao invés de todos os p preditores para divisão?

Florestas Randômicas

Porque considerar uma amostra randômica de m preditores ao invés de todos os p preditores para divisão?

- **Supor que temos um preditor muito forte no conjunto de dados juntamente com um número de outros preditores. moderadamente fortes, então no conjunto de árvores *bagged*, a maioria ou todos usarão o preditor mais forte para a primeira divisão!**

Florestas Randômicas

Porque considerar uma amostra randômica de m preditores ao invés de todos os p preditores para divisão?

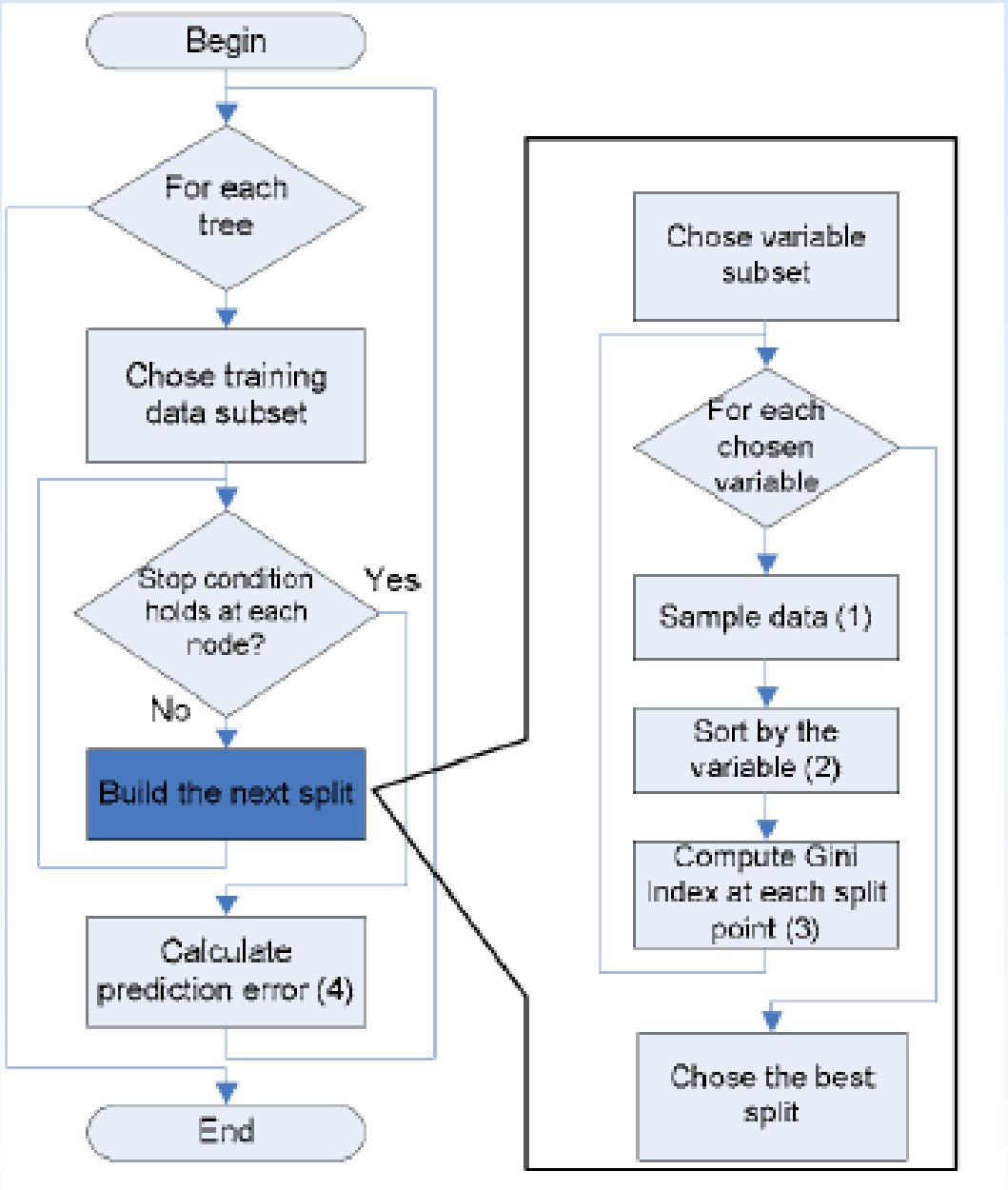
- **Supor que temos um preditor muito forte no conjunto de dados juntamente com um número de outros preditores. moderadamente fortes, então no conjunto de árvores *bagged*, a maioria ou todos usarão o preditor mais forte para a primeira divisão!**
- **Todas as árvores *bagged* parecerão similares. Logo todas as predições dessas árvores serão altamente correlacionadas**

Florestas Randômicas

Porque considerar uma amostra randômica de m preditores ao invés de todos os p preditores para divisão?

- Supor que temos um preditor muito forte no conjunto de dados juntamente com um número de outros preditores. moderadamente fortes, então no conjunto de árvores *bagged*, a maioria ou todos usarão o preditor mais forte para a primeira divisão!
- Todas as árvores *bagged* parecerão similares. Logo todas as predições dessas árvores serão altamente correlacionadas
- Fazendo a média de quantidades altamente correlacionadas não atinge uma alta redução na variância, e logo florestas randômicas "descorrelacionam" as árvores *bagged* levando a maior redução na variância.

Florestas Randômicas



- Algoritmo original publicado em 2001, Breiman L, Random Forests. *Machine Learning*, 45 (1), pp 5-32.

Exemplo: Florestas Randômicas

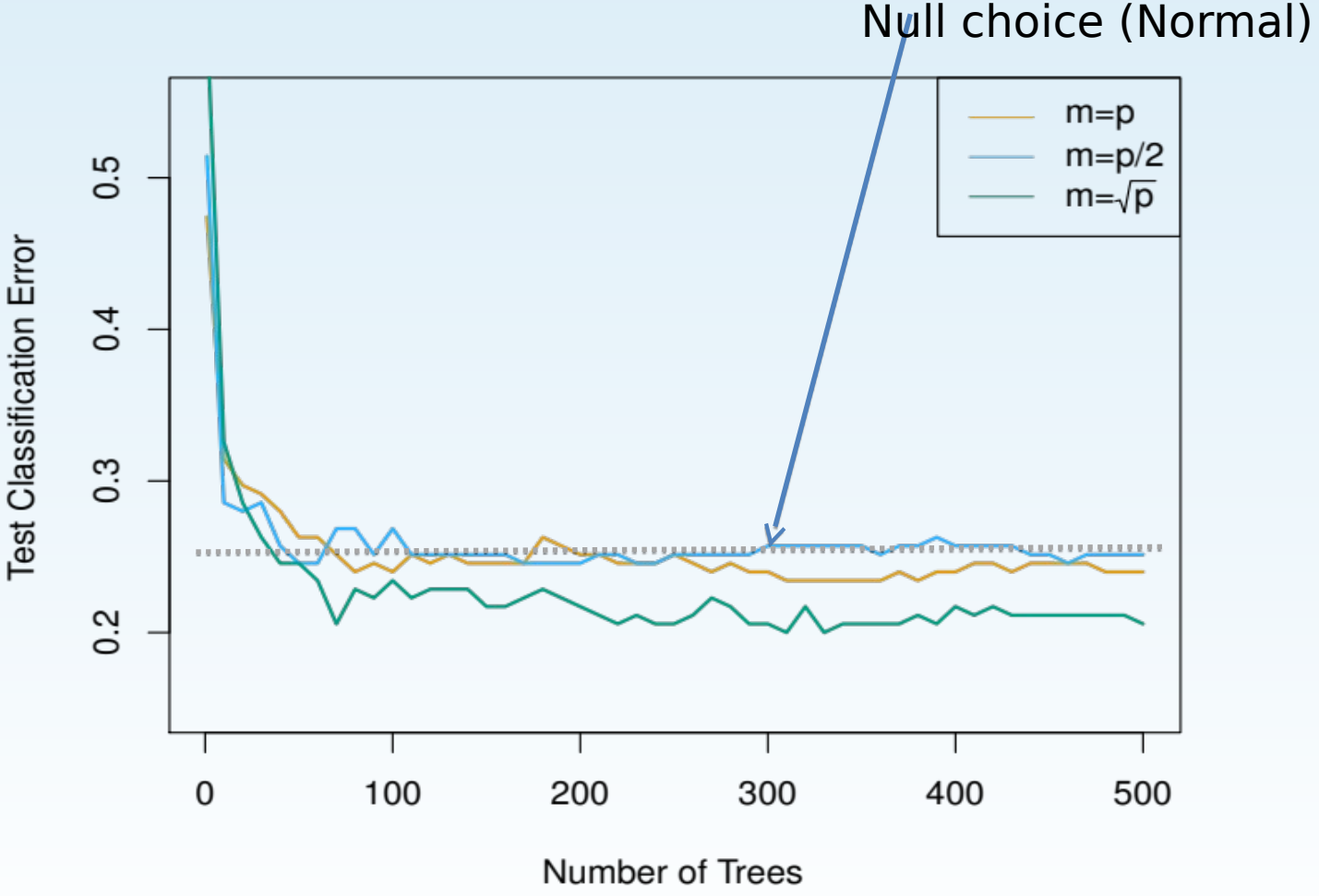
- 4.718 genes medidos em amostras de 349 pacientes
- Cads das amostras dos pacientes tem um rótulo qualitativo com 15 níveis diferentes: normal, ou 1 a 14 diferentes tipos de câncer.

- Usar florestas randômicas para predizer tipo de câncer baseando-se em 500 genes que possuem as maiores variâncias no conjunto de treinamento.

Florestas Randômicas

Floresta randômica com valores diferentes de "m"

- Notar que quando florestas randômicas são construídas usando m = p, então isso torna-se simplesmente *bagging*.



image/credit: James et. al. 2017

Exercícios/Leitura

- **Capítulo 19 do livro "Russell & Norvig, Artificial Intelligence: a modern approach", 4th ed, Pearson, 2020.**
- **Ler o Capítulo 8 do livro "*James, Witten, Hastie & Tibshirani, Introduction to Statistical Learning with applications in R, Springer, 2017*."**

Referências Bibliográficas

- Alpaydin, E. *Introduction to Machine Learning*. MIT Press, 2010.
- Bishop, C. *Pattern Recognition and Machine Learning.* Springer, 2006.
- James, G.; Witten, D.; Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with applications in R,* Springer, 2017.
- Mitchell, T. *Machine Learning.* McGraw Hill, 1997.