



UNIVERSIDADE  
FEDERAL DO CEARÁ



## Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2020

# Agenda

- ① Comitês de modelos
- ② Bagging
- ③ Boosting
- ④ Tópicos adicionais
- ⑤ Referências

# Comitês de modelos

- **Questão:** Por que combinar predições de diferentes modelos de aprendizagem?

# Comitês de modelos

- **Questão:** Por que combinar predições de diferentes modelos de aprendizagem?
- **Intuição:** Diferentes modelos podem apresentar erros distintos.

# Comitês de modelos

- **Questão:** Por que combinar predições de diferentes modelos de aprendizagem?
- **Intuição:** Diferentes modelos podem apresentar erros distintos.
- **Dilema viés-variância** (mais detalhes no próximo slide)
  - Viés alto em geral está relacionado a *underfitting* (modelo muito simples).
  - Variância alta em geral está relacionada a *overfitting* (modelo muito complexo).

# Comitês de modelos

- **Questão:** Por que combinar predições de diferentes modelos de aprendizagem?
- **Intuição:** Diferentes modelos podem apresentar erros distintos.
- **Dilema viés-variância** (mais detalhes no próximo slide)
  - Viés alto em geral está relacionado a *underfitting* (modelo muito simples).
  - Variância alta em geral está relacionada a *overfitting* (modelo muito complexo).
- O uso de múltiplos modelos visa **controlar o viés ou a variância**, buscando melhor **capacidade de generalização**.

# Comitês de modelos

## Dilema viés-variância

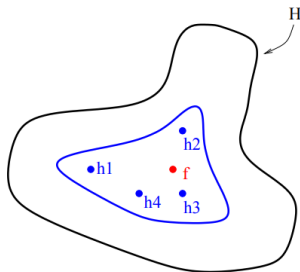
- Por exemplo, considere um estimador  $\hat{\theta}$  e  $\bar{\theta} = \mathbb{E}[\hat{\theta}]$  sua esperança para dados variantes.
- O erro quadrático médio obtido em relação ao parâmetro real  $\theta^*$  pode ser escrito como:

$$\begin{aligned}\mathbb{E}[(\hat{\theta} - \theta^*)^2] &= \mathbb{E} \left[ \left[ (\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^*) \right]^2 \right] \\ &= \mathbb{E}[(\hat{\theta} - \bar{\theta})^2] + 2(\bar{\theta} - \theta^*)\mathbb{E}[\hat{\theta} - \bar{\theta}] + (\bar{\theta} - \theta^*)^2 \\ &= \underbrace{\mathbb{E}[(\hat{\theta} - \bar{\theta})^2]}_{\text{variância de } \hat{\theta}} + \underbrace{(\bar{\theta} - \theta^*)^2}_{\text{viés de } \hat{\theta}}.\end{aligned}$$

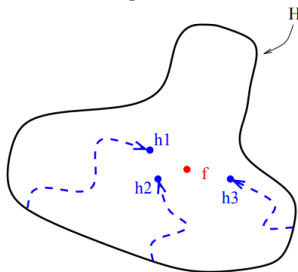
- Indica que um estimador com viés pode reduzir a sua variância no contexto de minimização do erro quadrático médio.

# Motivação para o uso de comitês (*ensembles*)

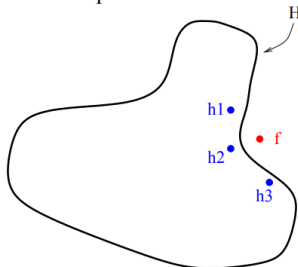
Statistical



Computational



Representational





# Comitês de modelos

- Condições para que um comitê de classificadores seja melhor que suas componentes individuais:
  - **Acurácia**: Os modelos individuais devem ser ao menos melhores que uma predição aleatória.
  - **Diversidade**: Os modelos individuais devem apresentar erros distintos.

# Comitês de modelos

## Comitês de modelos por votação

- Combina as saídas de múltiplos modelos obtidos por diferentes estratégias de aprendizagem.
- **Classificação**: retorna a média das probabilidades preditas ou a classe mais presente entre as predições individuais.
- **Regressão**: retorna a média das predições individuais.
- Pode-se ainda **ponderar as predições** de cada modelo individual.
  - Ponderação pelo desempenho em um conjunto de validação.

# Comitês de classificadores por votação majoritária

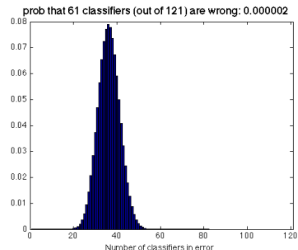
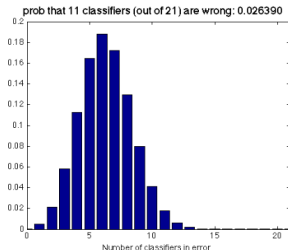
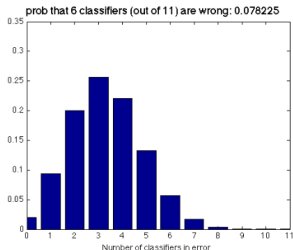
- Considere um problema de classificação binária e taxas de erros individuais independentes e iguais a  $\epsilon$ .
- **Questão:** Qual a probabilidade de um comitê por votação majoritária de  $M$  classificadores retorne um erro?

# Comitês de classificadores por votação majoritária

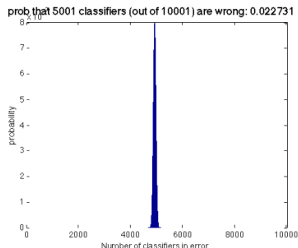
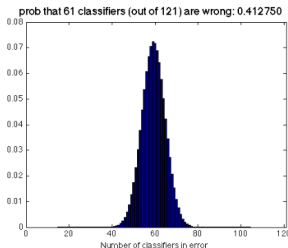
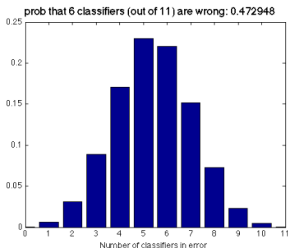
- Considere um problema de classificação binária e taxas de erros individuais independentes e iguais a  $\epsilon$ .
- **Questão:** Qual a probabilidade de um comitê por votação majoritária de  $M$  classificadores retorne um erro?

$$\sum_{k=\lfloor M/2 \rfloor + 1}^M P(\text{num erros} = k) = \sum_{k=\lfloor M/2 \rfloor + 1}^M \binom{M}{k} \epsilon^k (1 - \epsilon)^{M-k}.$$

# Motivação para o uso de comitês (*ensembles*)



$\epsilon = 0.3$ ;  $M = 11$ ,  $M = 21$  e  $M = 121$ .



$\epsilon = 0.49$ ;  $M = 11$ ,  $M = 121$  e  $M = 10001$ .

# Comitês de classificadores por votação majoritária

- **Problema:** Na prática é difícil garantir a completa independência das taxas de erro dos classificadores individuais.

# Comitês de classificadores por votação majoritária

- **Problema:** Na prática é difícil garantir a completa independência das taxas de erro dos classificadores individuais.
- Estratégias comuns para promoção de **diversidade**:
  - **Bagging:** Cada classificador é treinado com um conjunto de treinamento um pouco diferente.
  - **Boosting:** Cada classificador é treinado com pesos diferentes para cada exemplo.

# Agenda

- ① Comitês de modelos
- ② Bagging
- ③ Boosting
- ④ Tópicos adicionais
- ⑤ Referências



# Promoção de diversidade

- **Problema:** Classificadores base idênticos não apresentam melhoria quando combinados.

# Promoção de diversidade

- **Problema:** Classificadores base idênticos não apresentam melhoria quando combinados.
- **Ideia:** Promover diversidade através de modificações no conjunto de treinamento.

# Comitês de modelos

## Bootstrap Aggregating (Bagging) (BREIMAN, 1996)

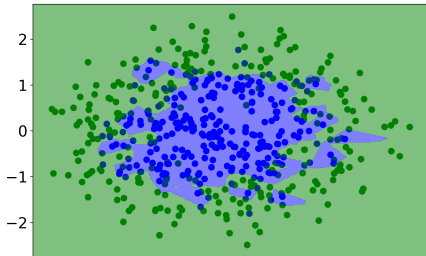
- Cria  $L$  subconjuntos a partir do conjunto de treinamento original via **amostragens aleatórias com reposição**.
- Cada modelo base é treinado com dados um pouco diferentes.
- Realiza votação majoritária ou média das saídas do comitê.
- Amostragem com reposição (*bootstrapping*) de um conjunto de  $N$  exemplos:
  - Cada exemplo possui probabilidade  $\left(\frac{N-1}{N}\right)^N$  de não ser adicionado em um subconjunto.
  - Para  $N$  grande, temos  $1 - \left(\frac{N-1}{N}\right)^N \approx 1 - \exp(-1) \approx 63.2\%$  de exemplos únicos em cada subconjunto.

# Bagging

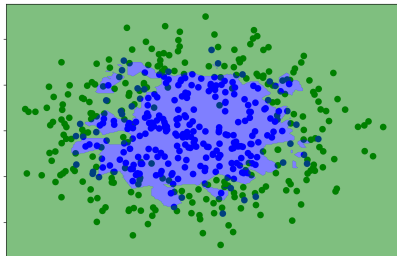
- Bagging é uma técnica de **redução de variância**.
- O viés de cada modelo base é um pouco maior, devido a menor quantidade de exemplos únicos de treinamento.
- **Out-of-bag**: Exemplos que não foram selecionados em um dado conjunto ( $\approx 1/3$  dos dados) e que podem servir de conjunto de validação.

# Exemplo de aplicação de Bagging com 3-NN

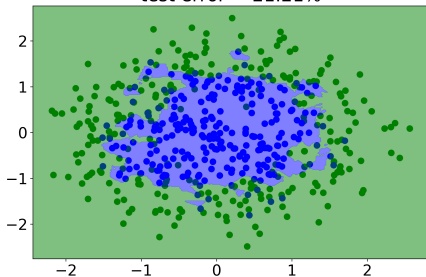
$N_c = 1$ , train error = 13.72%,  
test error = 25.25%



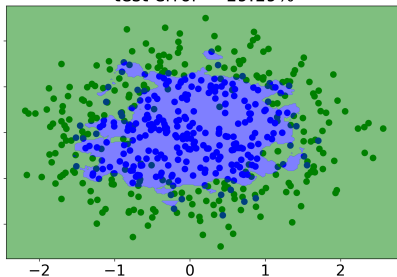
$N_c = 3$ , train error = 13.22%,  
test error = 23.23%



$N_c = 10$ , train error = 11.72%,  
test error = 21.21%



$N_c = 20$ , train error = 11.47%,  
test error = 19.19%



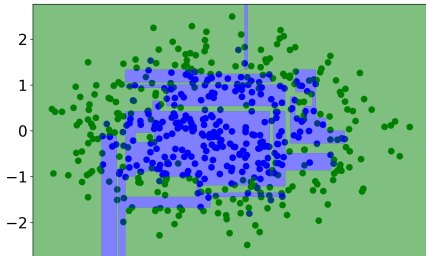
# Bagging

## Random Forest

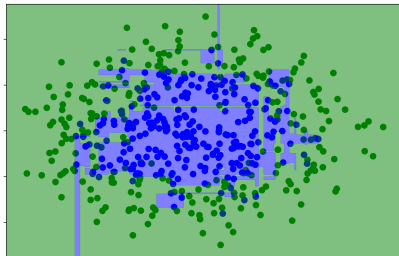
- Bagging de modelos de árvore de decisão.
- **Bagging de atributos:** Cada classificador base tem acesso a um subconjunto aleatório dos atributos disponíveis.
- Grande diminuição na variância do modelo final ao custo de um aumento no viés dos modelos base.
- Pode ser usado tanto em classificação quanto para regressão.

# Exemplo de aplicação do Random Forest

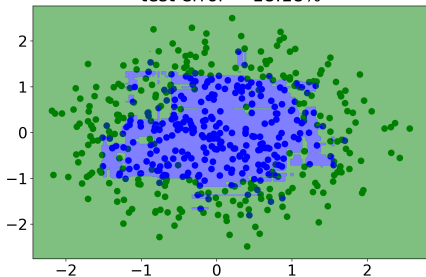
$N_c = 1$ , train error = 7.73%,  
test error = 25.25%



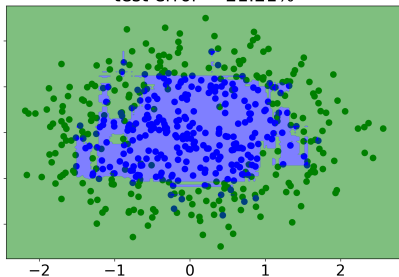
$N_c = 3$ , train error = 5.24%,  
test error = 22.22%



$N_c = 10$ , train error = 2.0%,  
test error = 18.18%



$N_c = 20$ , train error = 0.5%,  
test error = 21.21%



# Bagging

- **Vantagens:**

- Visa a diminuição da variância do modelo final.
- Em geral, resulta em melhor acurácia.
- Pode usar os exemplos *out-of-bag* para validação.
- Os modelos individuais podem ser treinados em paralelo.

- **Desvantagens:**

- Resulta no aumento de viés dos modelos base.
- Modelo final não é interpretável.
- Muitos modelos causam aumento do custo computacional.



# Agenda

- ① Comitês de modelos
- ② Bagging
- ③ Boosting**
- ④ Tópicos adicionais
- ⑤ Referências

# Comitês de modelos

## Boosting

- Estratégia de treinamento de múltiplos modelos em série, em que **o próximo modelo pondera mais os exemplos errados pelo modelo anterior**.
- Mantém um conjunto de pesos para cada padrão de treinamento.
- O modelo final agrega todos os modelos treinados a partir de uma soma ponderada.

# Boosting

- Boosting é uma técnica de **redução de viés**.
- Costuma ser aplicado com *weak learners*, como *decision stumps* (árvores de decisão com somente uma ramificação).
- Diversos algoritmos para aplicação de boosting no treinamento: **AdaBoost**, L2Boosting, Gradient Boosting, Logit Boosting...

# Boosting

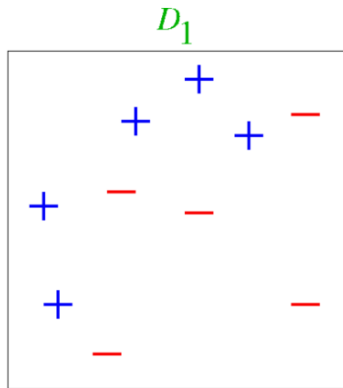
## AdaBoost

- 1 Inicializa os pesos dos exemplos de maneira uniforme:  $\beta_i = \frac{1}{N}$ ;
- 2 Para  $M$  classificadores:
  - 1 Treine um *weak learner*  $G_m$  com o conjunto de treinamento ponderado por  $\beta_i|_{i=1}^N$ ;
  - 2 Compute erros ponderados:  $e_m = \frac{\sum_i \beta_i \mathbb{I}(y_i \neq G_m(\mathbf{x}_i))}{\sum_i \beta_i}$ .
  - 3 Compute um peso para o classificador  $G_m$ :  $\alpha_m = \log \left( \frac{1-e_m}{e_m} \right)$ .
  - 4 Atualize os pesos dos exemplos de treinamento:

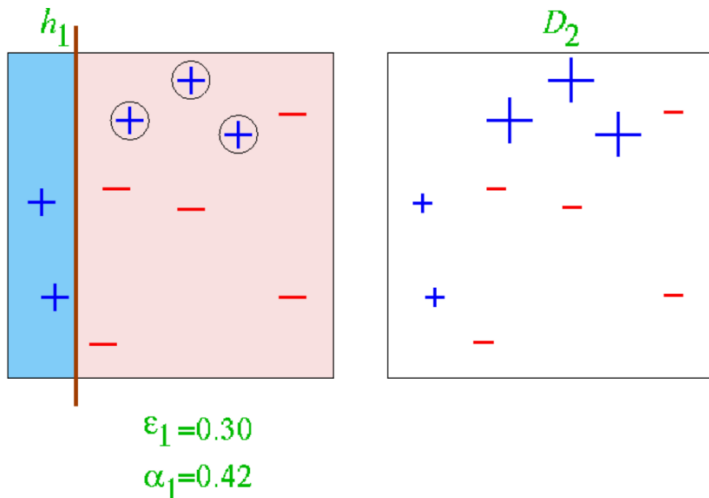
$$\beta_i \leftarrow \beta_i \exp \left( -\frac{1}{2} y_i \alpha_m G_m(\mathbf{x}_i) \right).$$

- 3 A saída do comitê é dada por:  $f(\mathbf{x}) = \text{sign} \left( \sum_m \alpha_m G_m(\mathbf{x}) \right)$ .

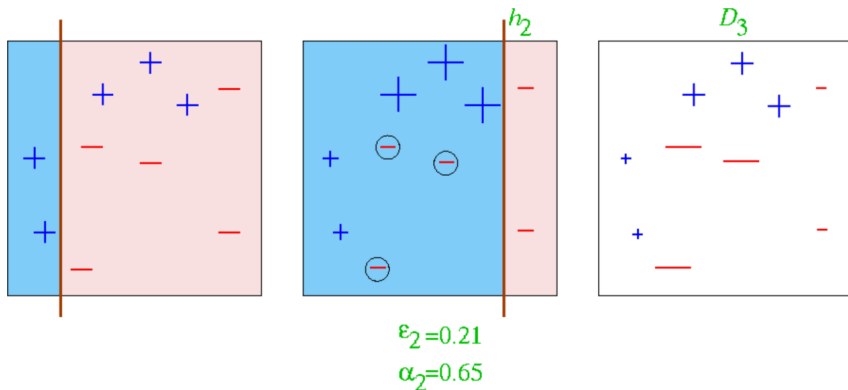
# Exemplo de aplicação do Boosting



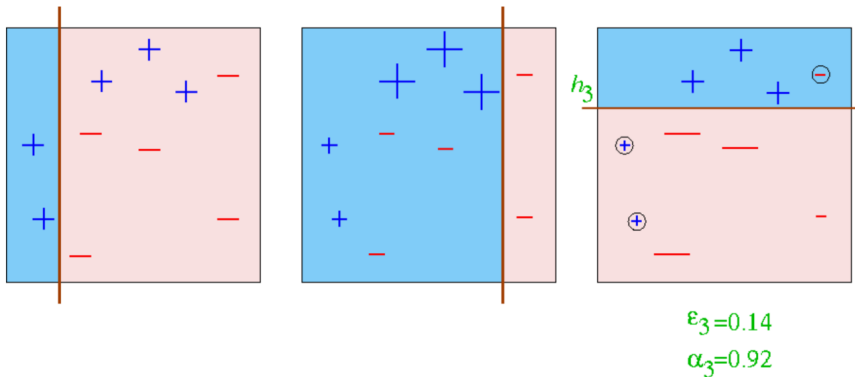
# Exemplo de aplicação do Boosting



# Exemplo de aplicação do Boosting

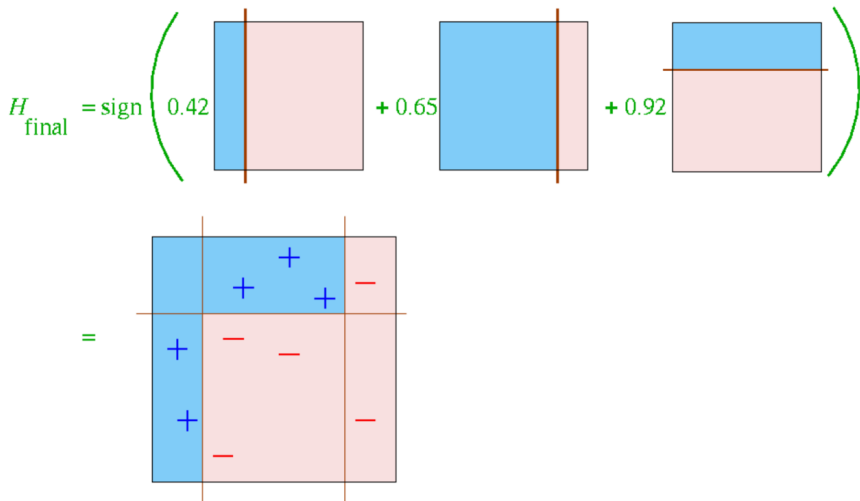


# Exemplo de aplicação do Boosting



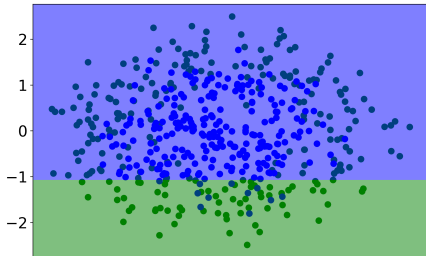


# Exemplo de aplicação do Boosting

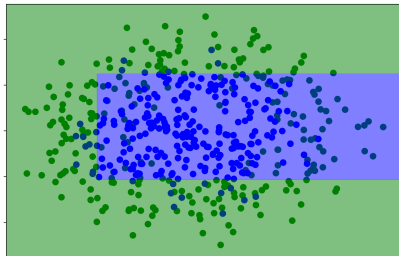


# Exemplo de aplicação do AdaBoost

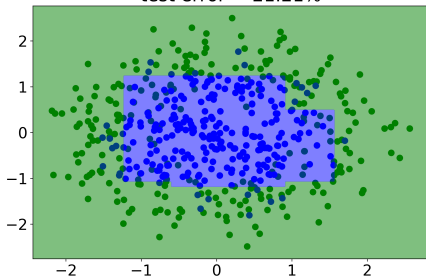
$N_c = 1$ , train error = 38.4%,  
test error = 40.4%



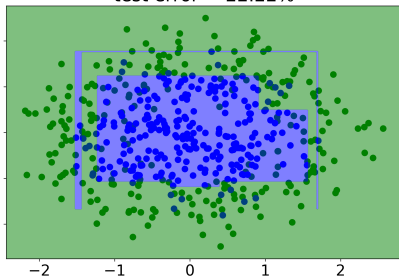
$N_c = 3$ , train error = 21.7%,  
test error = 27.27%



$N_c = 10$ , train error = 15.46%,  
test error = 21.21%



$N_c = 20$ , train error = 13.72%,  
test error = 22.22%



# Boosting

- **Vantagens:**

- Visa a diminuição incremental do viés.
- Em geral, resulta em melhor acurácia.
- Modelos aditivos (modelos anteriores não precisam ser retreinados)

- **Desvantagens:**

- Resulta no aumento da variância dos modelos base.
- Modelo final não é interpretável.
- Não pode ser treinado em paralelo.
- Pode ser sensível a dados ruidosos.

# Agenda

- ① Comitês de modelos
- ② Bagging
- ③ Boosting
- ④ Tópicos adicionais
- ⑤ Referências

# Tópicos adicionais

- Implementações open source: XGBoost, LightGBM, CatBoost.
- Bayesian Model Averaging:

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \sum_{k=1}^K p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \mathcal{M}_k)p(\mathcal{M}_k|\mathbf{X}, \mathbf{y}).$$

- Mistura de especialistas (mixture of experts):

$$p(y|\mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x})p_k(y|\mathbf{x}).$$

- Mistura hierárquicas de especialistas (hierarchical mixture of experts).

# Agenda

- ① Comitês de modelos
- ② Bagging
- ③ Boosting
- ④ Tópicos adicionais
- ⑤ Referências

# Referências bibliográficas

- **Cap. 16** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 14** - BISHOP, C. **Pattern recognition and machine learning**, 2006.