UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE COMPUTAÇÃO

Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2024

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)
  - $\rightarrow$ Viés alto em geral está relacionado a *underfitting* (modelo muito simples).
  - $\rightarrow$ 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)
  - $\rightarrow$ Viés alto em geral está relacionado a *underfitting* (modelo muito simples).
  - $\rightarrow$ 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

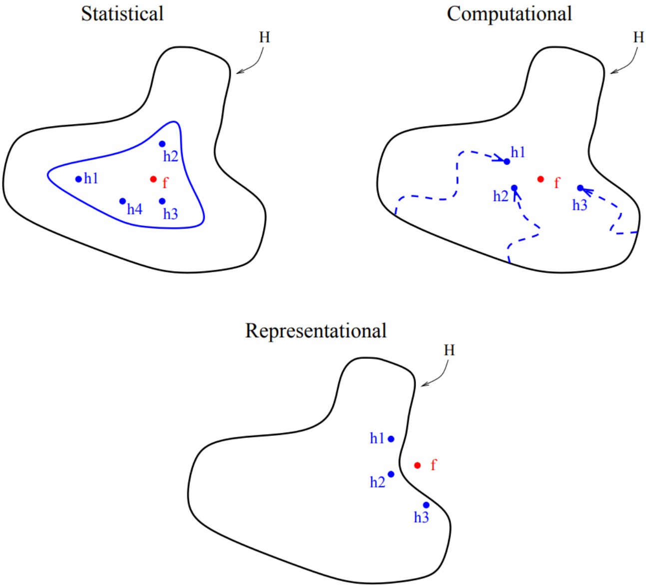
Dilema viés-variância

- 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 θ^* 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*)



DIETTERICH, Thomas G. Ensemble methods in machine learning, 2000.

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.
  - $\rightarrow$ Ponderação pelo desempenho em um conjunto de validação.

Comitês de classificadores por votação majoritária

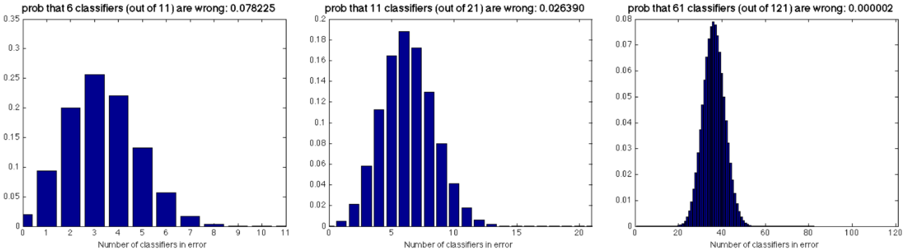
- Considere um problema de classificação binária e taxas de erros individuais independentes e iguais a ϵ.
- **Questão**: Qual a probabilidade de um comitê por votação majoritária de M classificadores retorne um erro?

Comitês de classificadores por votação majoritária

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

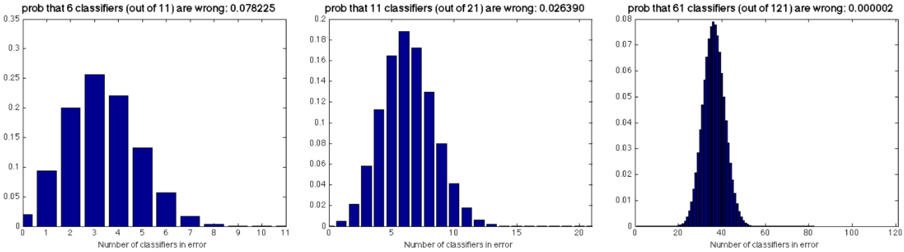
$$\sum_{k=\lfloor M/2 \rfloor+1}^{M} P(\text{num erros} = k) = \sum_{k=\lfloor M/2 \rfloor+1}^{M} \binom{M}{k} \epsilon^k (1-\epsilon)^{M-k}.$$

Motivação para o uso de comitês (*ensembles*)

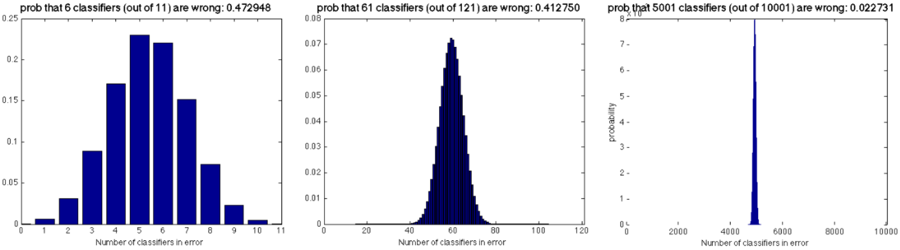


$\epsilon = 0.3$; $M = 11$, $M = 21$ e $M = 121$.

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

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

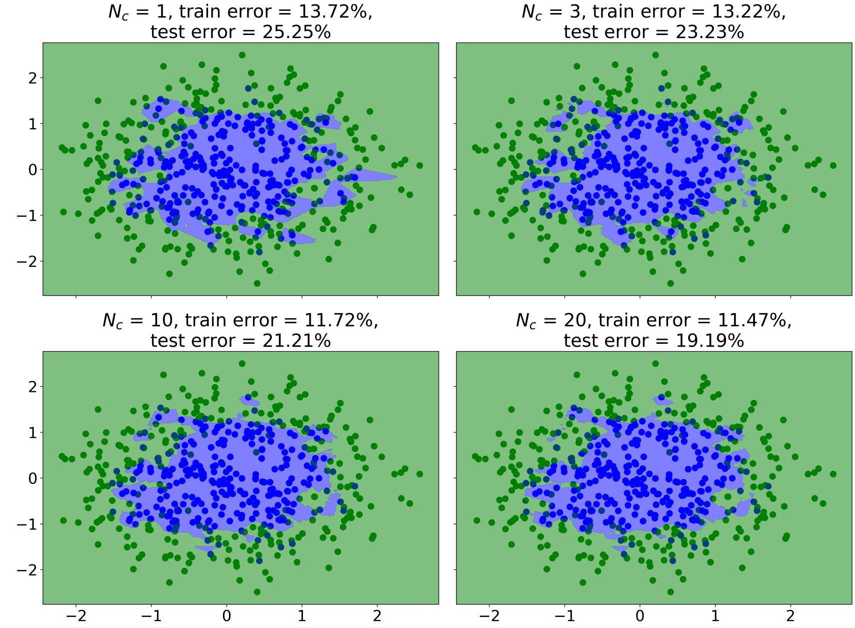
Bootstrap Aggregating (Bagging) (BREIMAN, 1996)

- 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

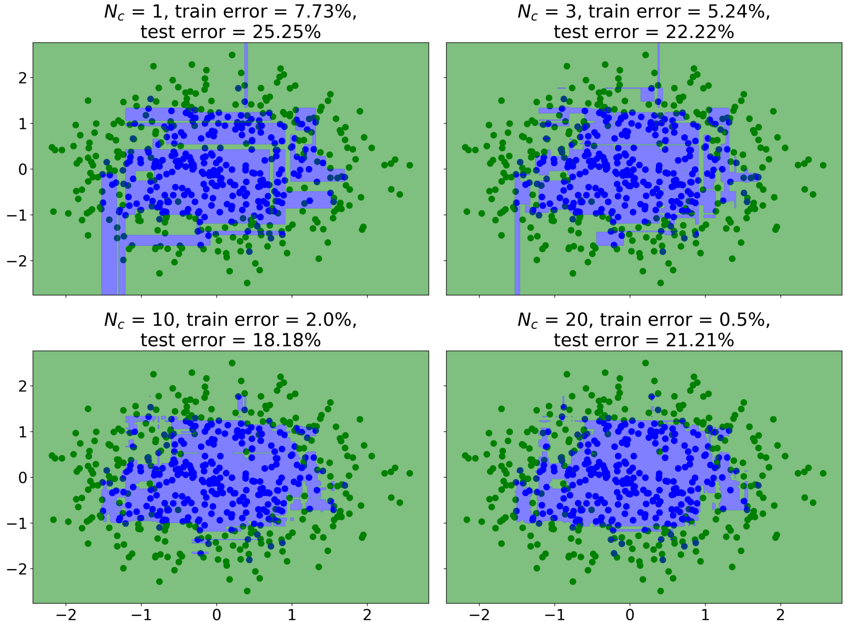


Bagging

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



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

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...

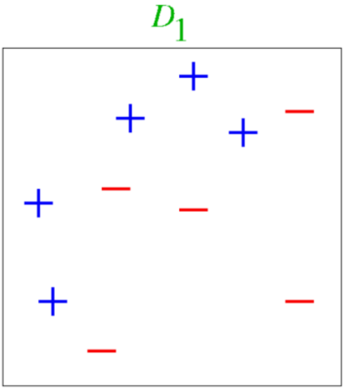
Boosting

AdaBoost

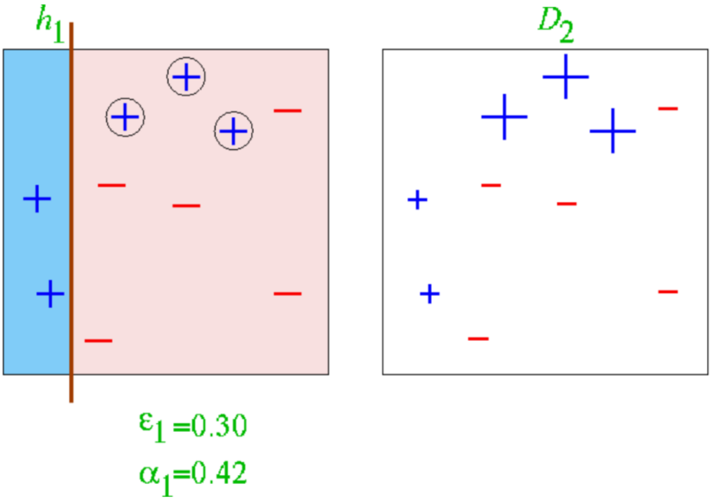
1. Inicializa os pesos dos exemplos de maneira uniforme: $\beta_i = \frac{1}{N}$;
2. Para M classificadores (considerando $y_i \in \{-1, 1\}$):
    1. Treine um *weak learner* G_m com o conjunto de treinamento ponderado por $\beta_i|_{i=1}^{N}$;
    2. Compute erros ponderados: $e_m = \frac{\sum_i \beta_i \mathbb{I}(y_i \neq G_m(\boldsymbol{x}_i))}{\sum_i \beta_i}$.
    3. Compute um peso para o classificador G_m: $\alpha_m = \frac{1}{2} \log\left(\frac{1-e_m}{e_m}\right)$.
    4. Atualize os pesos dos exemplos de treinamento:
    $$\beta_i \leftarrow \beta_i \exp\left(-y_i \alpha_m G_m(\boldsymbol{x}_i)\right).$$
3. A saída do comitê é dada por: $f(\boldsymbol{x}) = \text{sign}\left(\sum_m \alpha_m G_m(\boldsymbol{x})\right)$.

- Cada modelo dá mais ênfase aos padrões errados anteriormente.

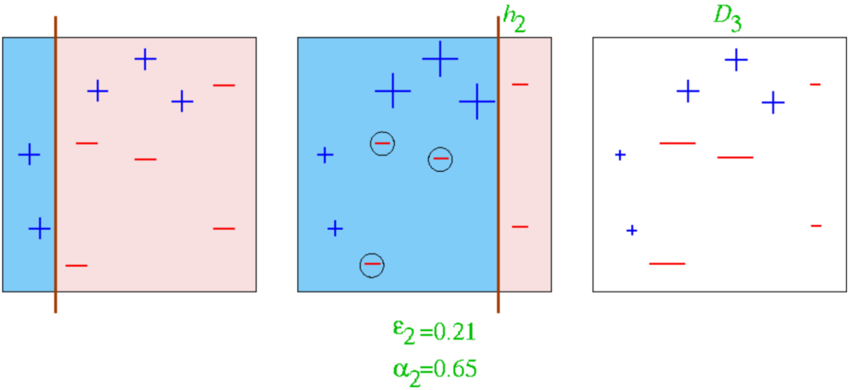
Exemplo de aplicação do Boosting



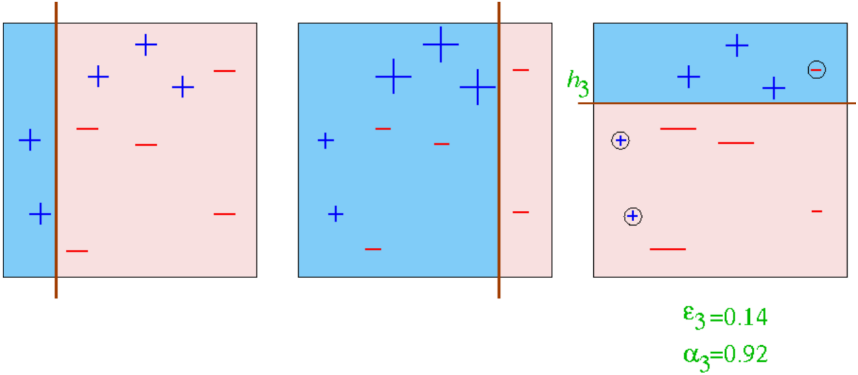
Exemplo de aplicação do Boosting



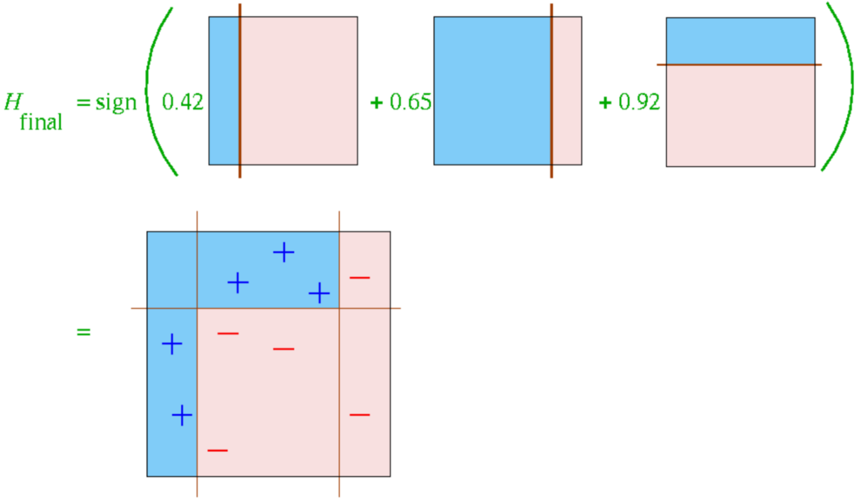
Exemplo de aplicação do Boosting



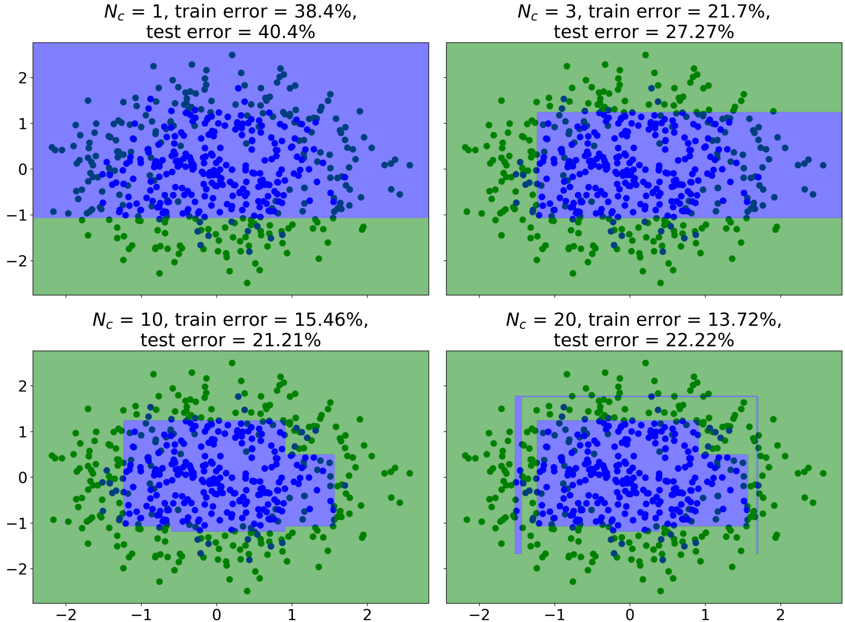
Exemplo de aplicação do Boosting



Exemplo de aplicação do Boosting



Exemplo de aplicação do AdaBoost



Boosting

Gradient Boosting

1. Inicializa com pesos uniformes: $\boldsymbol{f}_0 = \mathbf{1} \arg\min_\gamma \sum_{i=1}^N \mathcal{J}(y_i, \gamma)$;
2. Para M classificadores:
   1. Compute os gradientes residuais: $r_{im} = -\left[\frac{\partial \mathcal{J}(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{m-1}}$.

      Para $\mathcal{J}_{\mathsf{MSE}}(y_i, f_{m-1}(\boldsymbol{x}_i)) = \frac{1}{2}(y_i - f_{m-1}(\boldsymbol{x}_i))^2$:
      $$r_{im} = -\frac{\partial \mathcal{J}(y_i, f_{m-1}(\boldsymbol{x}_i))}{\partial f_{m-1}(\boldsymbol{x}_i)} = y_i - f_{m-1}(\boldsymbol{x}_i).$$
   2. Treine um *weak learner* G_m a partir do dataset $(\boldsymbol{x}_i, r_{im})|_{i=1}^N$.
   3. Calcule o multiplicador
      $\gamma_m = \arg\min_\gamma \sum_{i=1}^N \mathcal{J}(y_i, f_{m-1}(\boldsymbol{x}_i) + \gamma G_m(\boldsymbol{x}_i)).$
   4. Atualize os pesos: $\boldsymbol{f}_m = \boldsymbol{f}_{m-1} + \gamma_m G_m(\boldsymbol{X})$.
3. A saída será: $f(\boldsymbol{x}) = f_M(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \sum_{m=1}^M \gamma_m G_m(\boldsymbol{x})$.

- Cada modelo é treinado a partir dos "pseudo-resíduos" anteriores.
- $\mathcal{J}(y_i, \hat{y}_i)$ pode ser qualquer função custo diferenciável, como erro quadrático médio ou entropia cruzada.

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:

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

- Mistura de especialistas (mixture of experts):

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

- 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.