Aprendizagem de Máquina

Comitês (Ensembles) de Modelos

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA

Sumário

Viés e Variância

- Ao escolher $\hat{f}$ nos deparamos com um balanço entre
  - Aproximar f usando o conjunto de treinamento
  - Generalizar nos novos dados

Viés e Variância

- B conjuntos de dados são usados para produzir B hipóteses

$$\hat{f}_1 = \text{learn}(D1)$$
$$\hat{f}_2 = \text{learn}(D2)$$
$$\ldots$$
$$\hat{f}_B = \text{learn}(D_B)$$

- Assim, podemos estimar uma função média $\bar{f}$ para qualquer x

$$\bar{f}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)$$

Viés e Variância

- O **viés** mede quanto o modelo diverge da função objetivo

$$\text{viés}(x) = \left(\bar{f}(x) - f(x)\right)^2$$

- A **variância** indica a dispersão entre o modelo médio e cada modelo treinado com um dos B conjuntos de dados

$$\text{variância}(x) = \mathbb{E}_D[\left(f^{(D)}(x) - \bar{f}(x)\right)^2]$$

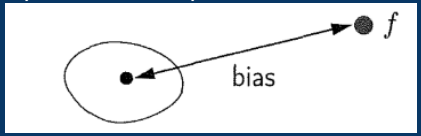
- A **variância** pode ser vista como uma medida de **"instabilidade"** no modelo de aprendizado, que se manifesta como uma reação a variações no conjunto de dados e resulta na geração das mais variadas hipóteses

Viés e Variância

$$\text{viés}(x) = \left(\bar{f}(x) - f(x)\right)^2$$

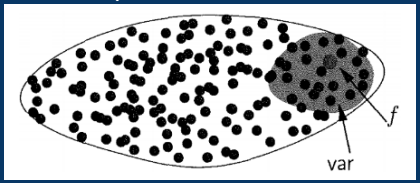
$$\text{variância}(x) = \mathbb{E}_D\left[\left(f^{(D)}(x) - \bar{f}(x)\right)^2\right]$$

Apenas uma hipótese



$\bar{f}(x) = f^{(D)}(x)$ para qualquer x, logo variância $= 0$
O viés dependerá apenas de quão próximo o modelo estará da f, logo, espera-se um viés alto.

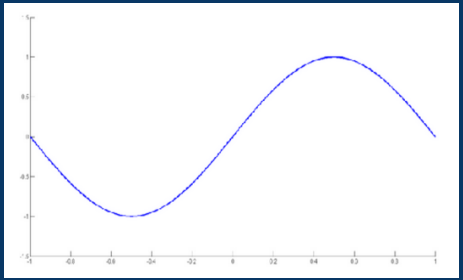
Muitas hipóteses



f está no conjunto
Viés ≈ 0, pois $\bar{f}$ deve estar perto de f
Variância alta

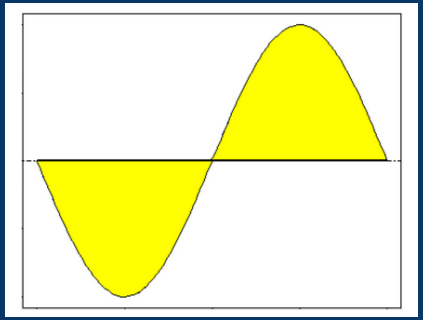
Viés e variância

- Função seno $f(x) = \text{seno}(\pi x)$

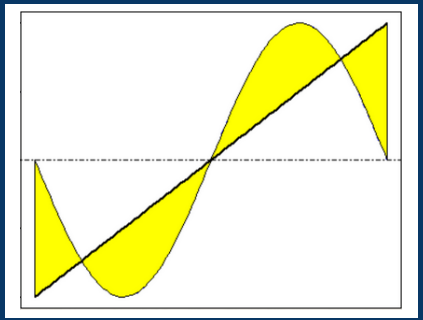


- Treinar duas hipóteses
  - $\mathcal{H}_0$: conjunto de todas as retas da forma $\hat{f}(x) = b$
  - $\mathcal{H}_1$: conjunto de todas as retas da forma $\hat{f}(x) = ax + b$

Erros das hipóteses

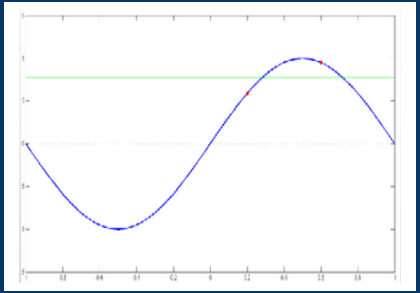


$E_{\mathcal{H}_0} = 0.5$

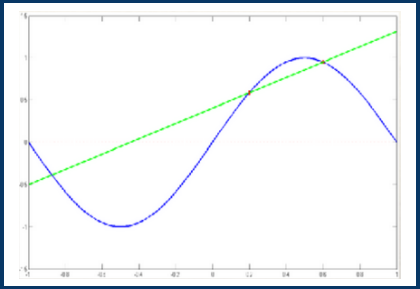


$E_{\mathcal{H}_1} = 0.2$

Treinar os modelos $\mathcal{H}_0$ e $\mathcal{H}_1$ usando duas instâncias: (x_1, y_1), (x_2, y_2)



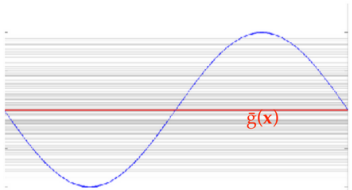
Para $\mathcal{H}_0$, a hipótese que melhor se ajusta aos dois pontos é $b = (y_1 + y_2)/2$



Para $\mathcal{H}_1$, a hipótese que melhor se ajusta é a reta que passa pelos dois pontos

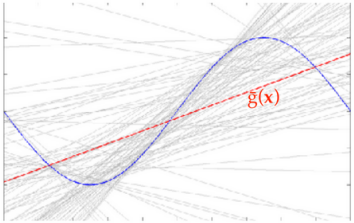
Repetindo esse processo com muitos pares de dados, podemos estimar viés e variância

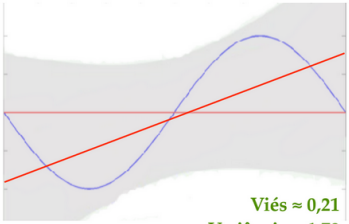
Analisando $\mathcal{H}_0$



Viés ≈ 0,50
Variância ≈ 0,25

Analisando $\mathcal{H}_1$





Viés ≈ 0,21
Variância ≈ 1,70

Viés e Variância

- Modelos com muito viés e pouca variância:
  - Modelos que assumem uma forma funcional para os dados, como regressões paramétricas, por exemplo
- Modelos com pouco viés e muita variância:
  - Modelos não-paramétricos que dependem muito de inicializações e dos dados de treinamento, como redes neurais e árvores de decisão

Como reduzir a variância?

Como reduzir a variância?

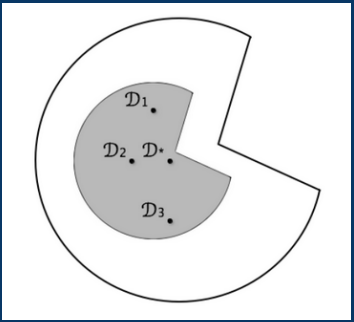
- Não há uma forma clara de escolher um bom método de aprendizagem
- Selecionar o melhor modelo de acordo com os dados de treinamento pode resultar no pior modelo para dados futuros
- Não há almoço grátis: não existe modelo que seja dominante para todas as distribuições de dados e a distribuição dos dados de treinamento é geralmente desconhecida

Resposta: combinar modelos

- Consiste em combinar as “opiniões” de modelos em um comitê na esperança de que a opinião combinada será melhor do que cada resposta individual
- Existem três justificativas para combinar modelos: Estatística, Computacional e Representacional

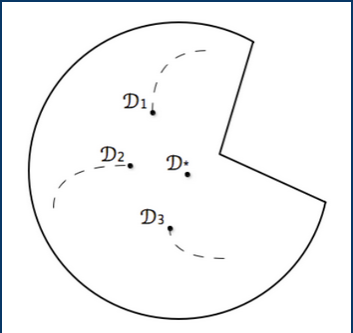
Justificativa estatística (pior caso)

- Dado um conjunto de modelos no espaço de modelos possíveis, podemos:
  - Escolher um modelo qualquer: risco de fazer uma má escolha
  - Obter o modelo médio:
    - não há garantia de ter desempenho melhor do que o melhor de todos os modelos D^*
    - evita a possibilidade de escolher os piores modelos



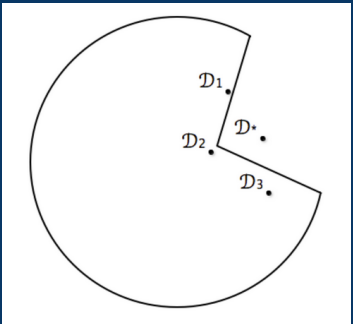
Justificativa computacional

- Algoritmos e inicializações diferentes levam a mínimos locais diferentes
- Modelos tendem a terminar o treinamento mais próximos do modelo ótimo D^*
- A agregação pode levar a um modelo que é uma melhor aproximação do que qualquer D_i



Justificativa representacional (melhor caso)

- D^* pode nem estar no espaço de modelos possíveis
  - Por exemplo, o espaço de modelos pode conter apenas modelos lineares
- Dessa forma, um ensemble de modelos lineares pode funções não lineares



Exemplo

- d_i é um classificador e cada coluna representa um padrão que pode ter sido classificado corretamente (1) ou não (0)
- A acurácia de cada classificador é 70%

d1	1	1	1	1	1	1	1	0	0	0
d2	1	1	1	1	1	1	1	0	0	0
d3	1	1	1	1	1	1	1	0	0	0

Exemplo

- E agora?

d1	1	1	1	1	1	1	1	0	0	0
d2	1	1	1	1	0	0	0	1	1	1
d3	1	0	1	1	0	1	1	0	1	1

Definições

- Dado um conjunto com B modelos $\{\hat{f}_1, \ldots, \hat{f}_B\}$
- O erro do b-ésimo modelo é dado por $\epsilon_b = f(x) - \hat{f}_b(x)$
- O erro médio quadrado é dado por: $MSE(\hat{f}_b) = \mathbb{E}[\epsilon_b^2]$
- O MSE médio é dado por $\overline{MSE} = \frac{1}{B} \sum_{i=b}^{B} \mathbb{E}[\epsilon_b^2]$

Definições

- Seja $\hat{f}_{comb}(x)$ o modelo combinado: $\hat{f}_{comb}(x) = \frac{1}{B}\sum_{i=b}^{B}\hat{f}_b(x)$
- Então: $MSE(\hat{f}_{comb}) = \mathbb{E}[(\frac{1}{B}\sum_{i=b}^{B}\epsilon_b)^2]$
- Assumindo que os ϵ_b são independentes:
  - $\mathbb{E}[\epsilon_b\epsilon_j] = \mathbb{E}[\epsilon_b]\mathbb{E}[\epsilon_j]$
  - $Cov(\epsilon_b, \epsilon_j) = 0$

Redução do MSE

$$
\begin{aligned}
MSE(\hat{f}_{comb}) &= \mathbb{E}[(\frac{1}{B}\sum_{i=b}^{B}\epsilon_b)^2] \\
&= \frac{1}{B^2}\sum_{i=b}^{B}\mathbb{E}[\epsilon_b^2] + \frac{1}{B^2}\sum_{j\neq b}\mathbb{E}[\epsilon_b\epsilon_j] \\
&= \frac{1}{B}\overline{MSE} + \frac{1}{B^2}\sum_{j\neq b}\mathbb{E}[\epsilon_b]\mathbb{E}[\epsilon_j] \\
MSE(\hat{f}_{comb}) &= \frac{1}{B}\overline{MSE}
\end{aligned}
$$

Redução do MSE

- Na prática, essa redução não é obtida porque a suposição $Cov(\epsilon_b, \epsilon_j) = 0$ não é satisfeita
- Portanto, precisamos encontrar os modelos mais diferentes possíveis

Métodos para criar modelos diversos e acurados

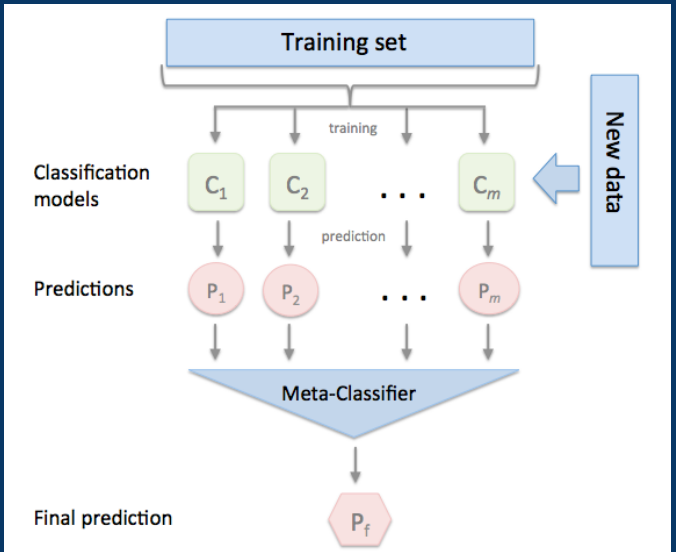
- Manipulação dos dados de treinamento:
  - Bagging, boosting, etc
- Aleatorização
  - Modelos treinados usando instâncias aleatórias diferentes
  - Inicializações diferentes dos pesos das redes neurais
- Variação de hiperparâmetros de modelos
- Variação de classificadores

Stacking

Stacking

- Às vezes chamado de *stacked generalization*
- Envolve treinar um modelo que combina as predições de vários outros modelos
- Primeiro, os outros modelos são treinados usando o conjunto de treinamento como entrada
- Depois, um modelo combinador é treinado tendo as saídas dos primeiros modelos como entradas

Stacking


Training set
training
New data
Classification models
C_1
C_2
C_m
prediction
Predictions
P_1
P_2
P_m
Meta-Classifier
Final prediction
P_f

Bagging

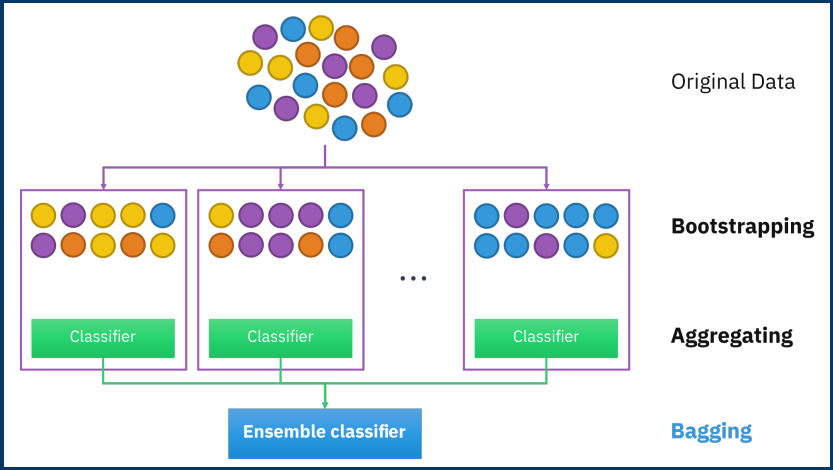
Bagging

- Contração de Bootstrap Aggregating
- Dado um conjunto de treinamento D de tamanho N, bagging gera B novos conjuntos de treinamento D_b selecionando N' observações de D uniformemente e com reposição (por isso o bootstrap no nome)
- Algumas observações podem ser repetidas em cada D_b
- Para cada conjunto de treinamento D_b, um modelo $\hat{f}_b$ é treinado
- Por fim suas saídas são combinadas:
  - Para regressão: fazemos a **média** das saídas de cada modelo
  - Para classificação: fazemos uma **votação** para cada classe ou calculamos a média das probabilidades estimadas por cada modelo para cada classe

Bagging

- Bagging reduz a variância, gerando melhores resultados para modelos instáveis, como redes neurais e árvores de decisão
- Mas pode piorar a performance de modelos estáveis como k-vizinhos mais próximos

Bagging

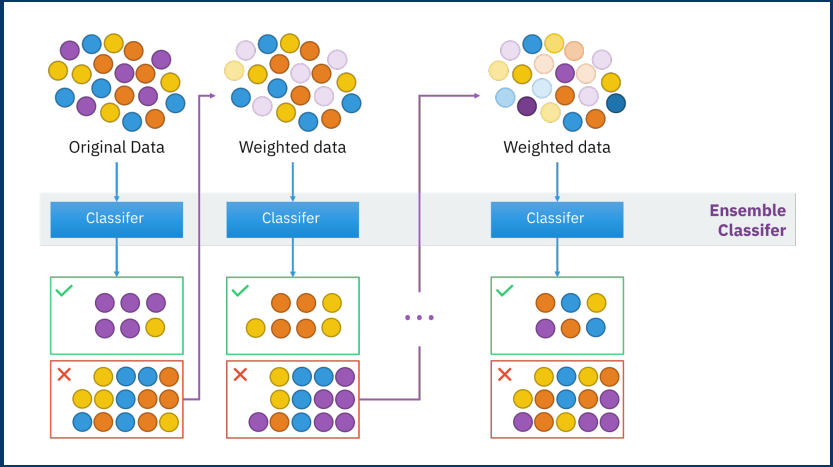

Original Data
Bootstrapping
Classifier
Classifier
Classifier
Aggregating
Ensemble classifier
Bagging

Boosting

Boosting

- Boosting foi proposto como uma resposta a uma pergunta de Kearns e Valiant (1988, 1989): “É possível que um conjunto de modelos fracos crie um único modelo forte?”
  - Um modelo é fraco quando suas respostas são apenas minimamente correlacionadas com a resposta verdadeira (pelo menos o suficiente para ser melhor que respostas aleatórias)
  - Um modelo forte produz respostas que são próximas às respostas esperadas
- A resposta afirmativa a essa pergunta veio em um paper de 1990 de Robert Schapire e levou ao desenvolvimento de algoritmos de boosting

Boosting


Original Data
Weighted data
Weighted data
Classifer
Classifer
Classifer
Ensemble Classifer

AdaBoost

- Algoritmo mais conhecido que implementa ensembles usando boosting
- Definições:
  - Na iteração t, temos um conjunto de classificadores $C_{(t-1)}(x_i) = \alpha_1 \hat{f}_1(x_i) + \ldots + \alpha_{(t-1)} \hat{f}_{(t-1)}(x_i)$
    - A classificação é dada pelo sinal de $C_{(t-1)}(x_i)$
  - O peso de cada instância é dado por $w_i^{(1)} = 1$ ou $w_i^{(t)} = e^{-y_i C_{(t-1)}(x_i)}$

AdaBoost – Passos

1. A cada iteração, escolha o modelo que minimiza o erro ponderado

$$\sum_{y_i \neq \hat{f}_t(x_i)} w_i^{(t)}$$

2. Use esse classificador para calcular a razão de erro

$$\epsilon_t = \frac{\sum_{y_i \neq \hat{f}_t(x_i)} w_i^{(t)}}{\sum_{i=1}^{N} w_i^{(t)}}$$

3. Calcule o peso do classificador

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

4. Adicione o classificador ao ensemble $C_t = C_{(t-1)}(x_i) + \alpha_t \hat{f}_t(x_i)$

Para Terminar

- Uma visão do boosting como uma minimização de uma função objetivo levou ao desenvolvimento do Gradient Boosting
- Implementações do Gradient Boosting, como XGBoost são hoje considerados alguns dos melhores modelos generalistas com valores-padrão de hiperparâmetros, sendo usados em várias aplicações práticas e vencendo competições no Kaggle
- Boosting não é um bom estimador de probabilidades (os modelos resultantes tendem a superestimar probabilidades)

Sugestão de Atividade

- Trabalhe no projeto **:D**

FIM

Aprendizagem de Máquina

Comitês (Ensembles) de Modelos

Telmo de Menezes e Silva Filho
tmfilho@gmail.com/telmo@de.ufpb.br
www.de.ufpb.br

UFPB Departamento de ESTATÍSTICA