Sistemas de Apoio à Decisão

Problema de Sobreajuste

Vahid Nikoofard

Universidade do Estado do Rio de Janeiro - Faculdade de Tecnologia
Laboratório Interdisciplinar Virtual de Inteligência Artificial (LIVIA)

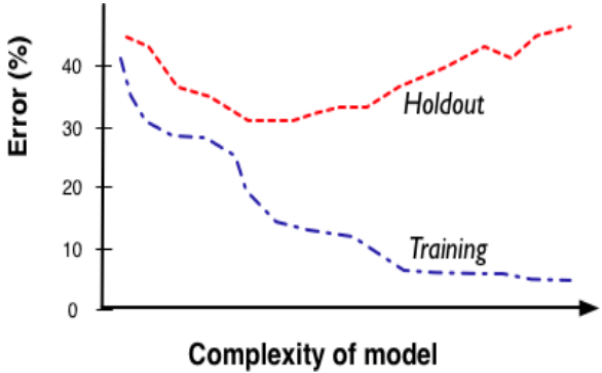
Novembro 2022

Generalização

- Um caso extremo: *table model* (memorização)
- Cada dataset é uma amostra da população
- complexidade *vs* overfitting

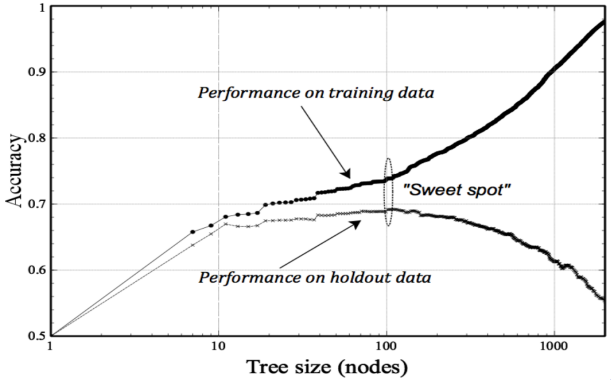
Dados de Retenção e Gráficos de Ajuste

- Dados de retenção (*holdout*) como um laboratório de teste
- Gráfico de ajusto: acurácia de um modelo como uma função de complexidade



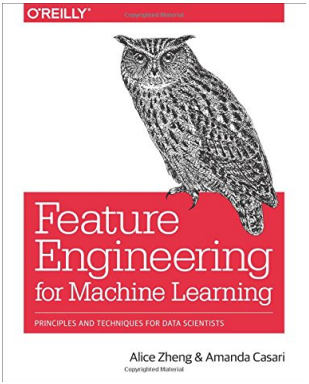
Sobreajuste na Indução de Árvore de Decisão

- Dividir até ficar um exemplo em cada folha $\rightarrow$ *table model*
- Um pouco melhor do que *table model* porque vai ter uma resposta não-trivial
- Determinar o **ponto ideal** é empírico.

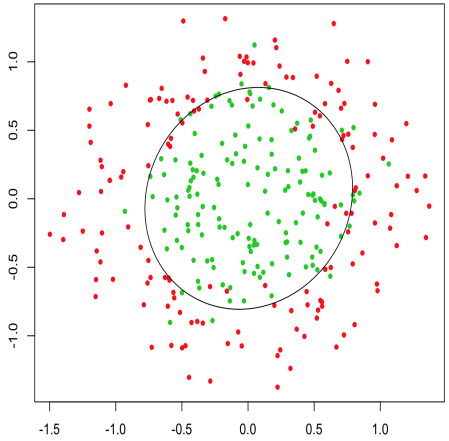


Sobreajuste em Funções Matemáticas

- Nosso modelo linear: $f(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2$
- Aumentando a complexidade: $f(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_4 x_4 + w_5 x_5$ onde $x_4 = x_1^2$ e $x_5 = x_2 / x_3$
- Aumentando a complexidade: Usar muitos atributos $\rightarrow$ solução: Feature Engineering



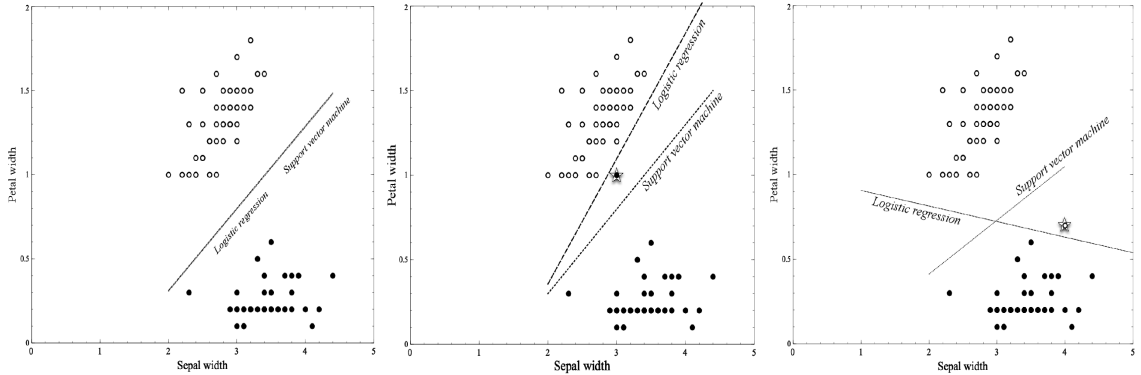
Feature Engineering: Um exemplo



$$f(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 \rightarrow f(\vec{x}) = w_0 + w_1 x_1^2 + w_2 x_2^2$$

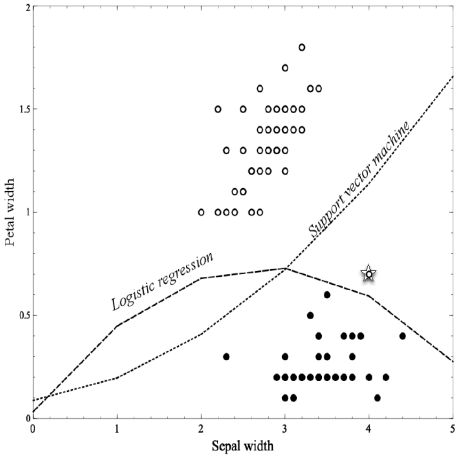
Sobreajuste em Funções Matemáticas: Funções Lineares

Efeito de um *outlier* nos modelos



Sobreajuste em Funções Matemáticas: Funções Lineares

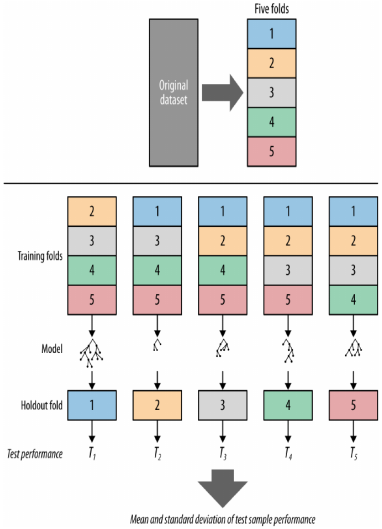
Efeito de mais atributos nos modelos: Quadrado de *Sepal Width*



Da Avaliação por Retenção até a Validação Cruzada (*cross-validation*)

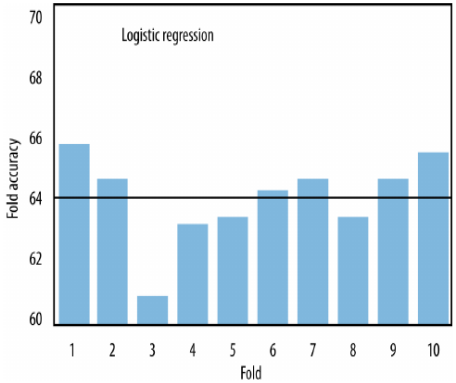
- *holdout* é apenas uma única estimativa → chance
- *cross-validation* é mais sofisticado e produz uma estatistica sobre a estimação
- A Construção de um "Laboratório" de Modelagem: eliminando as vies.

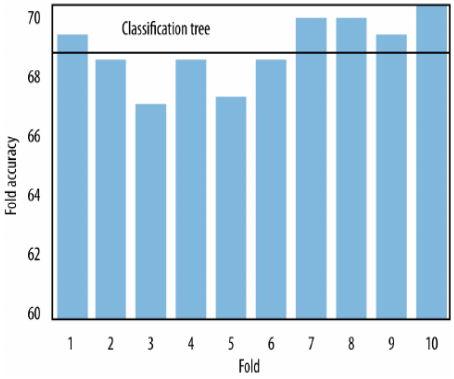
Cross-validation



Re-experimentar o Dataset de Rotatividade

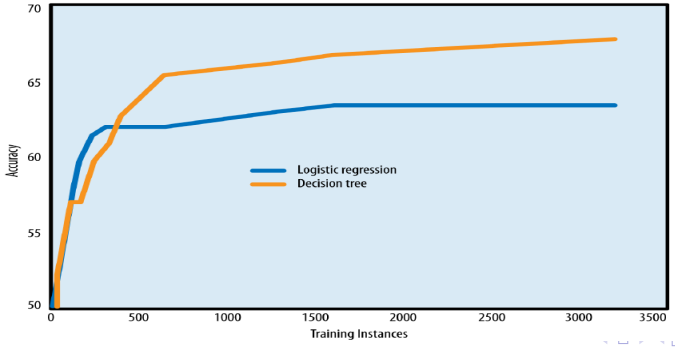
- Usar dataset inteiro para treino e teste $\rightarrow 73\%$
- 10-*fold cross validation*



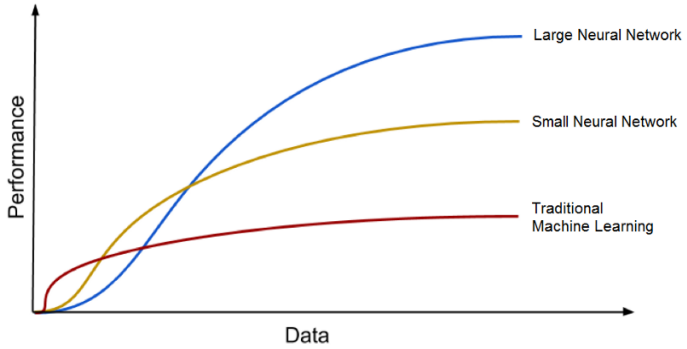


Curva de Aprendizagem

Uma curva de aprendizagem mostra o desempenho de generalização — desempenho apenas em dados de testes, representados graficamente contra a quantidade de dados de treinamento utilizados. Um gráfico de ajuste ilustra um desempenho de generalização, bem como o desempenho nos dados de treinamento, mas representados na complexidade do modelo. Gráficos de ajuste geralmente são mostrados para uma quantidade fixa de dados de treinamento.



Curva de Aprendizagem: *Deep Learning* vs algoritmos tradicionais



Como Evitar Sobreajuste com Indução de Árvore de Decisão

- Duas técnicas para evitar:
  - Parar de crescer antes de se tornar complexo demais
  - Deixar crescer até o final e depois podar a arvore.
- Qual é o numero mínimo de exemplos em cada folha?
- Teste de hipótese: *p-value* (Ex: para ganho de informação em cada divisão)

Teste de Hipótese

- **Hipótese nula**: Para fazer uma hipótese nula, significa que você está prevendo que não há significância estatística entre um resultado observado e o conjunto de dados ao qual ele pertence. Por exemplo, se a temperatura corporal média do grupo A e do grupo B for a mesma, você pode criar uma hipótese nula afirmando que não há significância estatística entre as temperaturas corporais médias do grupo A e do grupo B.
- **Significância**: No teste de hipóteses, a significância refere-se a quando é muito improvável que um resultado tenha ocorrido se a hipótese nula estiver correta.
- **Hipótese alternativa**: Este tipo de hipótese refere-se a quando há significância estatística entre um resultado observado e o conjunto de dados ao qual ele pertence, o que significa que seu teste rejeita uma hipótese nula que você fez. Por exemplo, você pode criar uma hipótese alternativa afirmando que há uma diferença entre as temperaturas corporais médias do grupo A e do grupo B com base em seus resultados.

Teste de Hipótese: cont.

O **valor-p** é uma medida que assume que a hipótese nula está correta, o que significa que, se o valor for pequeno, você pode rejeitar a hipótese nula em favor da hipótese alternativa. Um grande valor-p normalmente significa que o ponto de dados ou conjunto que você mediu se alinha com a hipótese nula, tornando-o o resultado mais provável.

Teste de Hipótese: Exemplo

João quer saber se a quantidade média de chuva para o mês de agosto é de nove centímetro. Ele encontra dados para o mês de agosto do ano passado e determina que a média amostral é de 8 centímetro, com um desvio padrão de 2 centímetros. Ele decide conduzir um teste-t bicaudal para encontrar o valor-p com um nível de $0,01$ para determinar se nove é a verdadeira média dos dados. Ele formula as seguintes hipóteses:

$H0 : \mu = 9$ polegadas
$H1 : \mu \neq 9$ polegadas

Depois de fazer a conta, ele chega no valor $0,006$, que é o valor-p para este teste. Como o valor-p é menor que o nível de significância de $0,01$, ele rejeita a hipótese nula que fez e aceita sua hipótese alternativa de que a quantidade média de chuva para o mês de agosto não é de nove centímetros.

Um Método Geral para Evitar Sobreajuste

- Dados de treino, validação e teste
- Depois de escolher a complexidade usar o todo dataset
- *Nested Cross Validation*
- Escolher a complexidade experimentalmente $\rightarrow$ custo computacional
- Seleção sequencial crescente (SFS, *sequential forward selection*) para selecionar os melhores atributos

Evitar Sobreajuste para Otimização de Parâmetros: Regularização

- Um balanço entre *fit* e simplicidade.
- $\arg\max_w \ \text{fit}(\vec{x}, \vec{w}) \rightarrow \arg\max_w \ [\text{fit}(\vec{x}, \vec{w}) - \lambda \cdot \text{penalty}(\vec{w})]$
- penalty: L2-norm (regressão *ridge*)$\rightarrow \ \ |\vec{w}|^2$ (elimina os pesos grandes)
- penalty: L1-norm (*lasso*) $\rightarrow \ \ |\vec{w}|$ (zera pesos menos importantes)

Cuidado com as "Comparações Múltiplas"

- Escolher o melhor conjunto entre todos os conjuntos aleatoriamente criados: exemplo de fundos de ações aleatórios.
- A raiz do problema de sobreajuste é a comparação múltipla.
- As tentativas de resolver o sobreajuste também são baseadas em comparação múltipla!!!
- formato de U no gráfico de ajuste $\rightarrow$ Bom Sinal