Avaliando Hipóteses

George Darmiton da Cunha Cavalcanti
Tsang Ing Ren
CIn/UFPE

Pontos importantes

- Erro da Amostra e Erro Real
- Como Calcular Intervalo de Confiança
  - Erros de hipóteses
- Estimadores
- Comparando Métodos de Aprendizagem
  - Distribuição Normal
  - Teste *t* Emparelhado
- Leave-one-out
- Bootstrap

Questões

- Dado o desempenho de uma hipótese sobre um conjunto de dados de tamanho limitado, como esse estimador se comporta diante de novos dados?
- Dado que um estimador supera o desempenho de outro estimador sobre um conjunto de dados de tamanho limitado, esse comportamento se manterá sobre novos dados?
- Quando poucos dados estão disponíveis, qual a melhor forma de usar esses dados para aprender a hipótese e para estimar seu desempenho?

Motivação

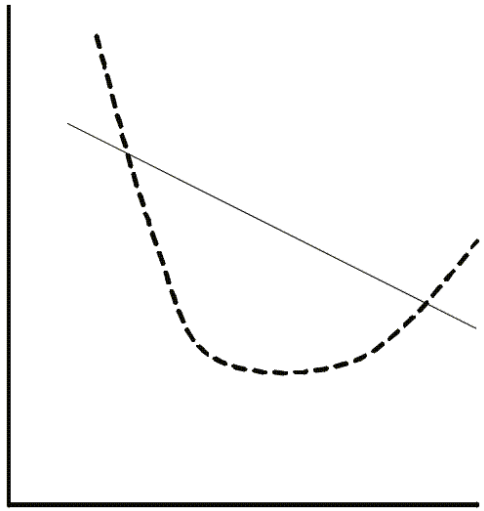
- Estimar o desempenho de uma hipótese quando se tem todos os dados é simples
- Entretanto, quando é necessário aprender a hipótese e avaliá-la usando uma pequena quantidade de dados, duas dificuldades surgem:
  - *Bias*
  - *Variance*

Bias *versus* Variance Dilemma

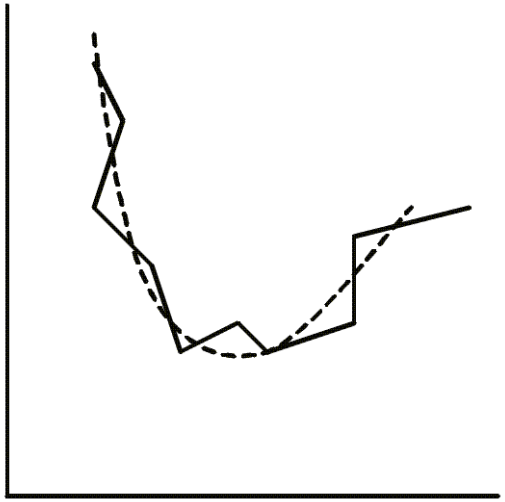
- *Bias*
  - O desempenho da hipótese aprendida sobre o conjunto de treinamento pode ser otimista.
  - *Statistical bias is the complexity restriction that the neural network architecture imposes on the degree of fitting accurately the target function.* The statistical bias accounts only for the degree of fitting the given training data, but not for the level of generalization.
- *Variance*
  - Ao avaliar a hipótese sobre um conjunto de teste, essa medida pode variar da real, devido à distribuição do conjunto de teste
  - *Statistical variance is the deviation of the neural network learning efficacy from one data sample to another sample that could be described by the same target function model. This is the statistical variance that accounts for the generalisation of whether or not the neural network fits the examples without regard to the specificities of the provided data.*

Bias *versus* Variance Dilemma

- O erro pode ser dividido em dois fatores: bias e variância.



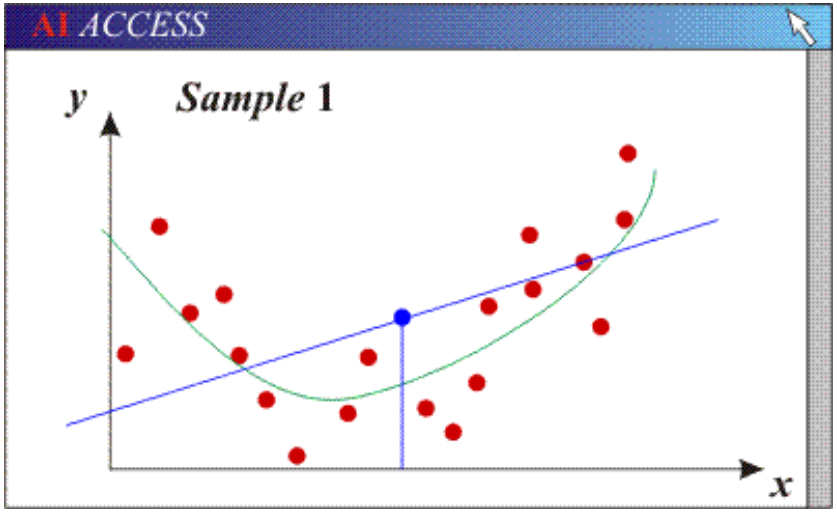
⇑ **Bias**
⇓ **Variância**



⇓ **Bias**
⇑ **Variância**

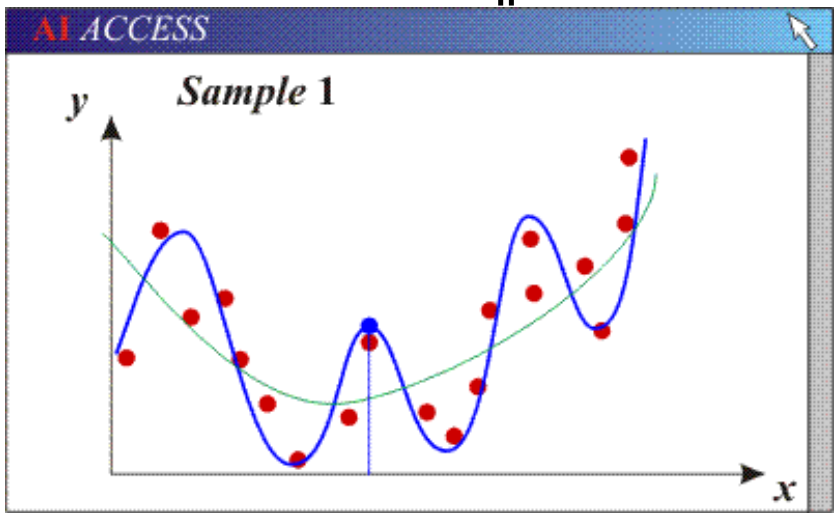
Um estimador que se ajusta muito bem aos exemplos de treinamento possui baixo bias e alta variância.
Se a variância é reduzida, o nível de ajuste aos dados diminui também.

Bias *versus* Variance Dilemma



⇑ **Bias**
⇓ **Variância**

⇓ **Bias**
⇑ **Variância**



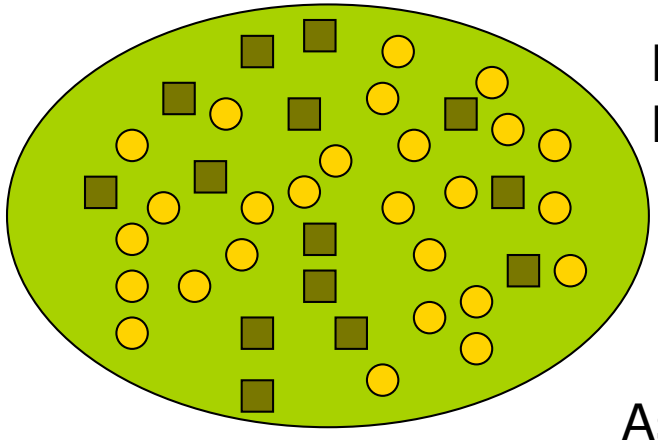
http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_bias_variance.htm

Objetivo

- Métodos para avaliar hipóteses
- Métodos para comparar o desempenho de duas hipóteses
- Métodos para comparar o desempenho de dois algoritmo de aprendizagem
  - quando apenas um conjunto limitado de dados está disponível

Estimando a Precisão de Hipóteses

X – espaço de possíveis instâncias



Diversas funções objetivo podem ser definidas
Por exemplo: ■ gosta de IA
● gosta muito de IA

Instâncias em ***X*** terão diferentes freqüências

Assim, uma maneira conveniente de modelar isso é através de distribuição ***D*** que especifica a probabilidade de encontrar cada instância em ***X***

A tarefa de aprendizagem consiste em encontrar uma função objetivo ***f*** dentro do espaço de hipóteses ***H***

$$f : X \rightarrow \{0,1\}$$

Classifica cada pessoa em: 0 – "gosta de IA" ou 1 – "gosta muito de IA"

Duas Definições de Erros

- Erro Real

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

- Erro da Amostra

$$error_{S}(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

$\delta(f(x) \neq h(x)) = 1$ se $f(x) \neq h(x)$ e 0 caso contrário.

Quão boa é a estimativa de $error_D(h)$ dada por $error_S(h)$?

Problemas para estimar o erro

- *Bias*
  - Se *S* é o conjunto de treinamento, $error_S(h)$ depende de S

$$bias \equiv E[error_S(h)] - error_D(h)$$

  - Para estimativas não-enviesadas, *h* e *S* devem ser escolhidos independentemente

- *Variance*
  - Mesmo com S não-enviesado, $error_S(h)$ pode variar de $error_D(h)$

Exemplo 1

- Hipótese *h* classifica incorretamente 12 dos 40 exemplos em *S*.

$$error_S(h) = \frac{12}{40} = .30$$

$$error_{\mathcal{D}}(h)?$$

Estimadores

- Experimento
  1. Escolha uma amostra S de tamanho n de acordo com a distribuição D
  2. Mensure $error_S(h)$

- $error_S(h)$ é uma variável aleatória
  - Ou seja, resultado de um experimento

- Dado $error_S(h)$, o que pode ser concluído acerca de $error_D(h)$?

Intervalo de Confiança

- Se
  - *S* contém *n* exemplos, selecionados independentemente de *h*
  - *n* ≥ *30* (número mágico)
- Então
  - Com aproximadamente *95%* de probabilidade, $error_D(h)$ pertence ao intervalo

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

Exemplo 2

- Hipótese *h* classifica incorretamente 12 dos 40 exemplos em *S*.

$$error_S(h) = \frac{12}{40} = .30 \qquad error_{\mathcal{D}}(h)?$$

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

$$0.30 \pm 1.96\sqrt{\frac{0.3 \times 0.7}{40}} = 0.30 \pm 1.96\sqrt{0.00525} = 0.30 \pm 0.142$$

$error_D(h)$ *está no intervalo* $[0.158, 0.442]$ *com 95% de probabilidade*

Intervalo de Confiança

- Se
  - *S* contém *n* exemplos, selecionados independentemente de *h*
  - *n* ≥ *30* (número mágico)
- Então
  - Com aproximadamente *N%* de probabilidade, $error_D(h)$ pertence ao intervalo

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Diferença entre Hipóteses

- Teste h_1 sobre a amostra S_1, teste h_2 sobre S_2
  1. Determine um parâmetro a ser estimado

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

  2. Escolha um estimador

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

  3. Determine a distribuição de probabilidade do estimador

$$\sigma_{\hat{d}} \approx \sqrt{\frac{\text{error}_{S_1}(h_1)(1-\text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1-\text{error}_{S_2}(h_2))}{n_2}}$$

  4. Encontre um intervalo *(L, U)* de confiança

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2)}{n_2}}$$

Comparando Algoritmos de Aprendizagem

- Deseja-se estimar

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

  - Sabendo que *L(S)* é a saída do estimador *L* usando o conjunto de treinamento *S*

- Mas, dado um conjunto limitado D_0, como determinar um bom estimador?

  - Pode-se dividir D_0 em conjunto de treinamento S_0 e de teste T_0, e medir

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

  - E melhor, repetir isso várias vezes e calcular a média dos resultados

Comparando Algoritmos de Aprendizagem

1. Dividir D_0 em k conjuntos T_1, T_2, …, T_k de tamanhos iguais (com pelos menos 30 exemplos).

2. Para i=1:k

   use T_i para testar e o restante para treinar

$$S_i \leftarrow \{D_0 - T_i\}$$
$$h_A \leftarrow L_A(S_i)$$
$$h_B \leftarrow L_B(S_i)$$
$$\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

3. Retorne o valor

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

Teste *t* Emparelhado
(Comparar h_A, h_B)

1. Particione os dados em *k* conjuntos de teste disjuntos T_1, T_2, …, T_k de tamanhos iguais (pelo menos 30 exemplos)
2. Para i=1:k

$$\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

3. Retorne o valor

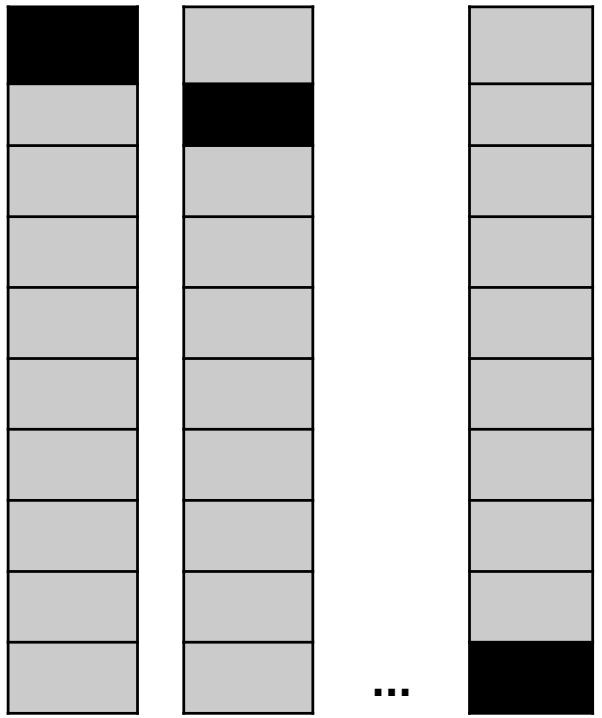
$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

- O intervalo de confiança

$$\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}} \qquad s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

Cross-Validation (n-fold cross-validation)

- O banco de dados é dividido em *n* partes disjuntas
- Uma das partes é escolhida para ser o conjunto de teste
- Todas as outras partes são usadas para treinar o sistema (construir a máquina de aprendizagem)
- O erro do conjunto de teste é aferido
- O procedimento é repetido de forma que cada um dos subconjuntos seja o conjunto de teste por exatamente uma vez
- No final do procedimento o erro médio é calculado



Cross-Validation (n-fold cross-validation)

1. Dividir D_0 em k conjuntos T_1, T_2, …, T_k de tamanhos iguais (com pelos menos 30 exemplos).

2. Para i=1:k

   use T_i para testar e o restante para treinar

$$S_i \leftarrow \{D_0 - T_i\}$$
$$h \leftarrow L(S_i)$$
$$\delta_i \leftarrow error_{T_i}(h)$$

3. Retorne o valor

$$\bar{\delta}_i \equiv \frac{1}{k}\sum_{i=1}^{k}\delta_i$$

Teste de Hipótese

- **Hipótese Estatística**
  - Uma declaração acerca de parâmetros de uma ou mais populações

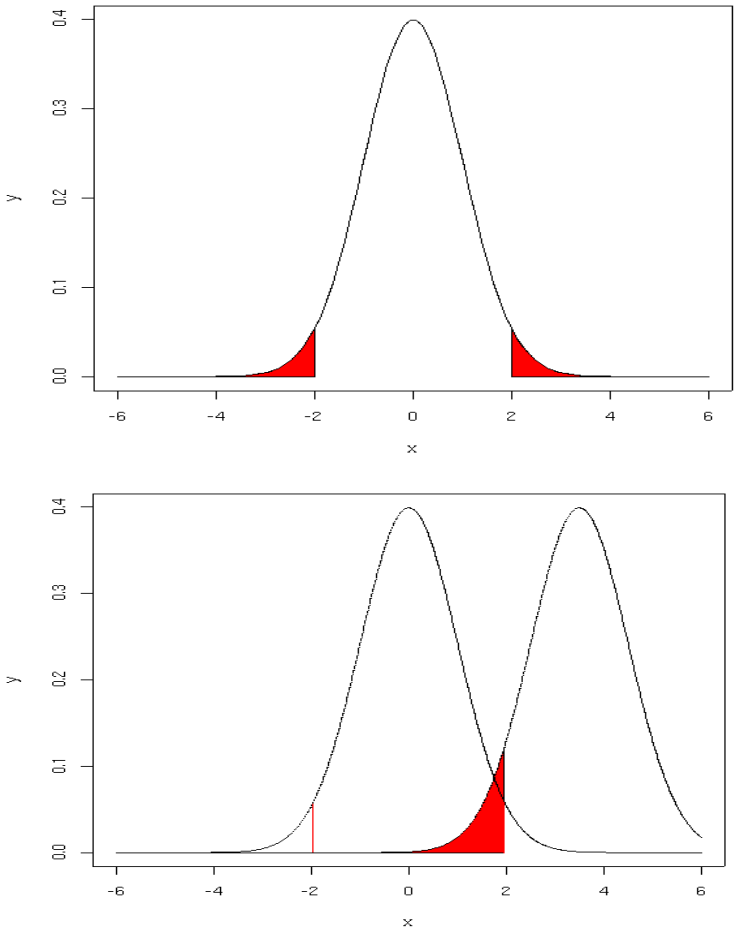
- **Teste de Hipótese**
  - Um procedimento para decidir entre a aceitação e a rejeição da hipótese
    - Identificar os parâmetros de interesse
    - Definir uma hipótese nula, H_0
    - Definir uma hipótese alternativa, H_1
    - Escolher um nível de significância α
    - Escolher um teste estatístico

Hipóteses Estatísticas

- **Hipótese Nula (H_0):** uma declaração a qual se presume ser verdadeira antes que as evidências estatísticas mostrem o contrário
  - Geralmente especifica um valor exato para um parâmetro
  - Exemplo H_0: $\mu = 70$ Kg
- **Hipótese Alternativa (H_1):** deve ser contrária, oposta, antagônica à *hipótese* nula
- **Teste Estatístico:** estatística calculada de medidas de uma amostra/experimento
  - Um teste estatístico assume uma distribuição (normal, t, chi-quadrado, entre outras)

Erros nos Testes de Hipóteses

- **Erro Tipo I** ocorre quando H_0 é *rejeitada* mas ela é de fato verdadeira
  - *P(Erro Type I)=α ou* nível de significância
- **Erro Tipo II** ocorre quando há falha em rejeitar H_0 mas ela é de fato falsa
  - *P(Erro Type II)=β*
  - 1-*β* = Probabilidade de rejeitar corretamente H_0



Teste de diferenças entre médias (1/6) (um exemplo)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L_1	63,5	70,4	66,2	56,0	60,3	74,5	69,8	57,5	63,3	66,9
L_2	64,0	71,2	68,1	55,8	61,0	74,0	70,7	58,5	63,5	68,2

A tabela mostra a taxa de acerto de duas máquina de aprendizagem para experimentos realizados utilizando *10-fold-cross-validation*

Teste de diferenças entre médias (2/6) (um exemplo)

Enunciando as hipóteses

- Deseja-se verificar se a média de uma das máquinas é maior/menor do a outra
- Em outras palavras
  - H_0 diz que não existe diferença entre as máquinas

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

$$Sabendo\,que\ \mu_d = \mu_{L_1} - \mu_{L_2}$$

Teste de diferenças entre médias (3/6) (um exemplo)

Estabelecer o nível de significância

$$\alpha = 0{,}01 \qquad 1 - \alpha = 0{,}99$$

Identificar a variável de teste

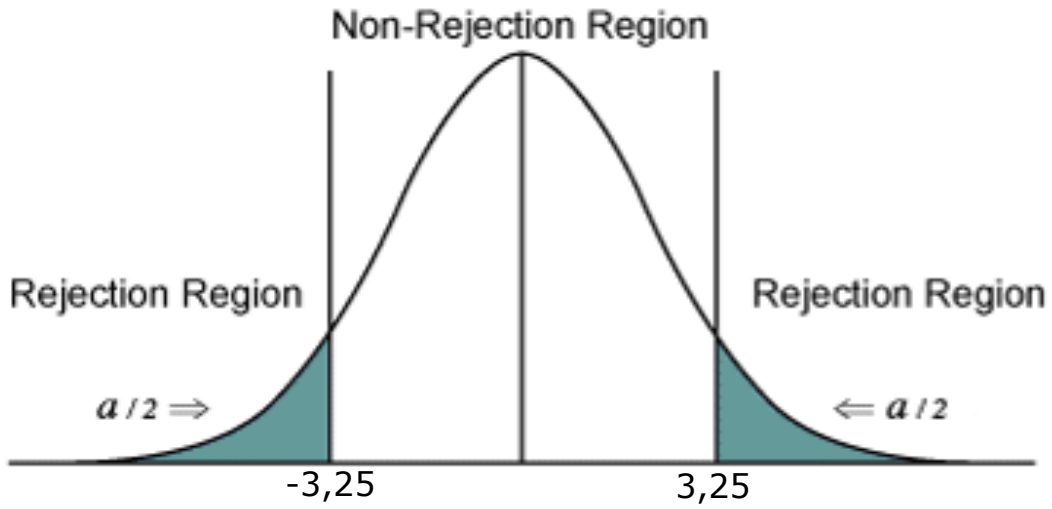
- Nesse problema tem-se uma amostra de apenas 10 elementos.

- Como a amostra tem menos de 30 elementos a variável de teste que será utilizada é a variável $\mathbf{t_{n-1}}$ da distribuição t de *Student*.

Teste de diferenças entre médias (4/6) (um exemplo)

Definir a região de aceitação de H_0

$$t_{n-1;critico} = t_{10-1;0,01} = t_{9;0,01} = -t_{9;0,99} = 3,25$$



Para valores entre -3,25 e 3,25 aceita-se H_0.
Há uma chance de 1% de rejeitar H_0, mesmo sendo ela verdadeira.

Teste de diferenças entre médias (5/6) (um exemplo)

Calcular a diferença

	1	2	3	4	5	6	7	8	9	10
L_1	63,5	70,4	66,2	56,0	60,3	74,5	69,8	57,5	63,3	66,9
L_2	64,0	71,2	68,1	55,8	61,0	74,0	70,7	58,5	63,5	68,2
d_i	-0,5	-0,8	-1,9	0,2	-0,7	0,5	-0,9	-1,0	-0,2	-1,3
d_i^2	0,25	0,64	3,61	0,04	0,49	0,25	0,81	1	0,04	1,69

Calcular a média e o desvio padrão

$$\bar{d} = \frac{1}{n}\sum_i d_i = \frac{-6,6}{10} = -0,66$$

$$s_d = \sqrt{\frac{\sum d_i^2 - \left(\left(\sum d_i\right)^2 / n\right)}{n-1}} = 0,704273$$

Teste de diferenças entre médias (6/6) (um exemplo)

Calcular o valor da variável de teste

$$t_{n-1} = \frac{\bar{d}}{\left(s_d / \sqrt{n}\right)} \qquad t_{10-1} = t_9 = \frac{-0{,}66}{0{,}704273/\sqrt{10}} = -2{,}96349$$

Decidir pela aceitação ou rejeição de H_0

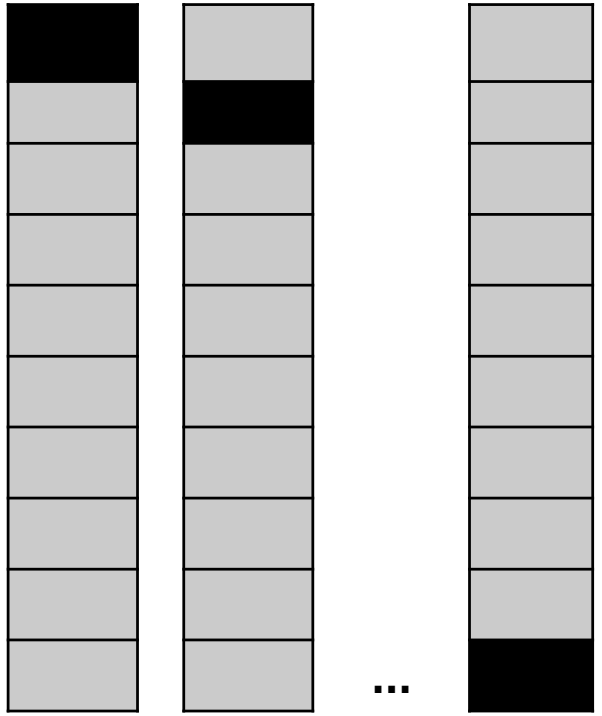
$$t_{n-1} = t_9 = -2{,}96349 \;>\; t_{n-1;critico} = t_{9;0,01} = -3{,}25$$

Como o valor da variável de teste foi MAIOR do que -3,25, a hipótese H_0 é aceita com 1% de significância

Assim, conclui-se com 99% de confiança (chance de erro de 1%) que a máquina L_2 possui uma taxa de acerto igual a L_1, na média.

Cross Validated t-test

- Teste *t* emparelhado sobre *k* (10) taxas de acerto obtidas através do método *k-fold cross validation*



- Vantagens
  - Conjunto de treinamento grande
  - Mais poderoso (Diettrich, 98)

- Desvantagens
  - Os resultados (acerto ou erro) não são independentes (*overlap*)
  - Probabilidade elevada do erro Tipo-I (Diettrich, 98)
    - Erro Tipo-I
    - Rejeitar H_0 quando ela é verdadeira

Cross Validated t-test (exemplo)

#	Algoritmo 1 (precisão)	Algoritmo 2 (precisão)	Diferença
2	74,00	73,90	0,10
3	66,50	66,10	0,40
4	69,00	67,20	1,80
5	68,00	67,90	0,10
6	71,00	69,40	1,60
7	70,00	69,90	0,10
8	70,00	68,60	1,40
9	67,00	67,90	-0,90
10	68,00	67,60	0,40

Média	69,15	68,53

- O algoritmo 1 é melhor do que)?

Assim, existem evidências suficientes para afirmar que o algoritmo 1 obterá melhores resultados do que o algoritmo 2 em 95% dos casos.

$$\bar{d} = \frac{1}{n}\sum_i X_{1i} - X_{2i} = 0{,}62$$

$$s_d = std\_dev(X_{1i} - X_{2i}) = 0{,}85088$$

$$t = \frac{(\bar{d} - \mu_d)\sqrt{n}}{s_d} = 2{,}304$$

α= 0,05 n-1=10-1=9 $t_{0.05,\ 9}$=1,833

H_1: $\mu_d > 0$ $t > t_{\alpha/2,n-1}$

Outras Medidas

- Leave-one-out (LOO)
- Bootstrap

Leave-one-out (1/3)

- ***LOO cross-validation*** é simplesmente ***n-fold-cross-validation***, sendo que ***n*** é o número de instâncias no banco de dados

- Ou seja, é retirada uma instância do banco de dados e a máquina de aprendizagem é treinada com o restante dos dados.
  - O elemento retirado é usado para testar
  - Esse processo é repetido para todos os elementos do banco de dados
  - Depois o erro médio é calculado

Leave-one-out (2/3)

- Vantagens
  - Quase todo o conjunto de dados é usado para treinar o sistema
    - O que aumenta a chance do classificador ter uma taxa de acerto real
  - O procedimento é determinístico, ou seja, nenhuma seleção aleatória é usada

Leave-one-out (3/3)

- Desvantagens
  - Alto custo computacional
  - Conjunto de teste não-estratificado
    - Um problema artificial: duas classes com o mesmo número de instâncias cuja classificação é realizada pela maioria. O erro nesse caso seria de 100%, pois o padrão de teste estaria sempre em minoria no banco de dados de teste
- Estratificação
  - Garantir o mesmo número de elementos por classe no conjunto de teste

Bootstrap (1/3)

- É baseado no procedimento estatístico de amostragem com repetição
- Nas técnicas previamente apresentadas, os conjunto usados eram disjuntos
- Entretanto, alguns métodos de aprendizagem podem tirar proveito da duplicação de instâncias

Bootstrap (2/3)

- Assim, a idéia é construir um conjunto de treinamento com repetições
- 0.632 bootstrap
  - Dado um banco de dados com *n* instâncias, *n* instâncias são coletadas aleatoriamente, com repetição, gerando um conjunto de treinamento com *n* instâncias

$\frac{1}{n}$ ← Probabilidade de uma instância ser selecionada

$\left(1 - \frac{1}{n}\right)$ ← Probabilidade de uma instância não ser selecionada

$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0{,}368$ ← Probabilidade de uma instância não ser selecionada em nenhuma das n vezes

Sabendo que *e* é a base do logaritmo natural, 2,7183

Bootstrap (3/3)

Desta forma, o conjunto de teste irá conter, aproximadamente, 36,8% das instâncias, e o conjunto de treinamento, 63,2%.

Esse procedimento pode ser o melhor caminho de estimar o erro para bancos de dados que possuam poucas instâncias