Teste de Significância Estatística em AM

Profa. Elaine Faria

Junho/2017

Teste Estatístico

São regras de decisões, vinculadas a um fenômeno da população, que nos possibilitam avaliar, com o auxílio de uma amostra, se determinadas hipóteses podem ser rejeitadas, ou não

A maior parte das ciências se utiliza da técnica Estatística denominada Teste de Hipóteses

Definição retirada do material de aula da disciplina Estatística II da UFPR, Departamento de Estatística

Hipótese Estatística

- É uma suposição que pode ser verdadeira ou não relativa a uma ou mais populações

- Toma-se uma amostra aleatória da população de interesse e com base nela é estabelecida se a hipótese é verdadeira ou falsa

- Hipótese nula (H_0): hipótese de igualdade
  - Consiste em afirmar que os parâmetros ou características matemáticas de duas ou mais populações são idênticos

Exemplos de Aplicações para o Teste de Hipótese

- A média de altura das pessoas da cidade A é igual a da cidade B?
- Os homens e mulheres tem a mesma temperatura corporal?
- A pressão arterial das pessoas com uma dada doença diminuiu depois que eles tomaram um novo medicamento?
- Uma propaganda na televisão surtiu efeito?
- Uma metodologia de educação infantil está associado ao aprendizado?

Hipótese Nula

- Toma-se uma amostra a fim de inferirmos a respeito do valor paramétrico (θ)
- Por meio de um estimador obtém-se obtém-se a estimativa do parâmetro (θ')
- Verificar se a diferença observada entre θ−θ', é significativa, ou não
- Quanto menor a diferença, maior será a probabilidade de não rejeitarmos H0.
  - θ− θ' não foi significativa → diferença ocorreu por acaso.
- Caso contrário, devemos rejeitar H0
  - Diferença foi suficientemente grande para não ter, provavelmente, ocorrido ao acaso.

Definição retirada do material de aula da disciplina Estatística II da UFPR, Departamento de Estatística

Hipótese nula - Exemplo

- Será que a altura média ($\theta = \mu$) dos alunos da UFPR é de 1,71 m?

  Hipótese Nula: H0 : μ = 1, 71 m

- Deve-se colher uma amostra de tamanho n e obter a estimativa da média ($\theta' = \mu_{obs}$)

- Verificar a diferença entre μ e μ_{obs}
  - Caso H0 fosse rejeitada, concluiríamos que a diferença observada foi significativa e que não se deveu ao acaso

Exemplo retirado do material de aula da disciplina Estatística II da UFPR, Departamento de Estatística

Hipótese Alternativa

- É uma hipótese que, necessariamente, difere de H0.
- No exemplo anterior teríamos

  $H1 : \mu \neq 1,71$ m ou

  $H1 : \mu < 1,71$ m ou

  $H1 : \mu > 1,71$ m

Nível de significância α

- Definição
  - É o limite que se toma como base para afirmar que um certo desvio é decorrente do *acaso* ou não
  - É um limiar de confiança que informa se vamos ou não rejeitar a hipótese nula

- Deve ser previamente definido, sendo usualmente usado 0,05 (95% de certeza de que de fato existe uma diferença significativa) e 0,01 (90%)

- A partir de um nível de significância convencionado (α) os desvios são devidos à lei do acaso e o resultado é considerado não significativo

  O teste estatístico é convertido em uma probabilidade condicional chamada **p-value** (significância de um resultado)**.** Se P <= α, rejeita-se a hipótese nula

Nível de significância

- Ex: se você decidiu por um nível de significância de 0.05 (95% de certeza que há realmente uma diferença significativa), então o *p-value* menor que 0.05 indica que você deve rejeitar a hipótese nula

- Se o seu teste obtém p=0.07, significa que você não pode rejeitar a hipótese nula de igualdade → não há diferença significativa na análise conduzida

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Será que testes estatísticos podem nos ajudar nos experimentos de AM?

Vamos relembrar o que é AM ....

Aprendizado de Máquina

É uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.

Monard, M. C e Baranauskas J. A. Sistemas Inteligentes, Ed. Manole, 2003.

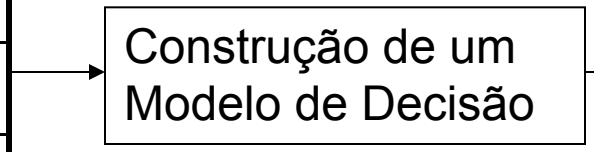
Introdução

- Duas importantes áreas do AM são
  - Classificação (aprendizado supervisionado)
  - Agrupamento (aprendizado não-supervisionado)

Exemplo de Classificação

Nome	Idade	Renda	Pagador
João	<30	Média	Bom
Ana	41..50	Alta	Bom
Pedro	41..50	Alta	Bom
Maria	41..50	Baixa	Ruim
Paulo	<30	Baixa	Ruim
Aldo	>60	Alta	Ruim

Base de Dados
Treinamento



Algoritmo

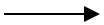
Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Classificador

Exemplo de Classificação

Nome	Idade	Renda	Pagador
Ivo	31..40	Baixa	**????**



Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Ruim

Novo Dado

Classificador

Classificação

Exemplos de Medidas de Avaliação usadas em Classificação

	Classe Verdadeira	
Classe Prevista	**P**	**N**
P	VP	FP
N	FN	VN

Matriz de Confusão

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

$$\text{Pr}ecisão = \frac{VP}{VP + FP}$$

$$\text{Re}call = \frac{VP}{VP + FN}$$

Exemplos de Técnicas de Amostragem usadas em Classificação

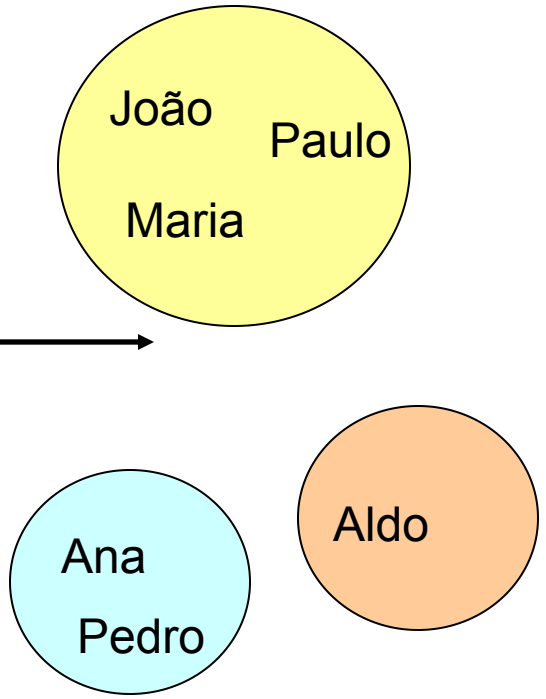
- *Holdout*
  - Treino: ½ ou 2/3
  - Teste: o resto dos dados
- Random Subsampling
  - Múltiplas execuções do *holdout*
  - Diferentes partições de treino e teste sem intersecção
- *Cross validation*
  - Dividir o conjunto de dados em k partições (ex: 10-fold)
  - Treino: k-1 partições
  - Teste: 1 partição
- Leave-One-Out
  - N-fold-cross validation
  - Treino: N-1 exemplos
  - Teste: 1 exemplo

Exemplo de Agrupamento

Nome	Idade	Peso
João	25	60
Ana	50	75
Pedro	60	90
Maria	22	65
Paulo	18	68
Aldo	15	80

Base de Dados

Aplicação de uma técnica agrupamento



Agrupamento

Medidas de Avaliação usadas em Agrupamento

- Critérios de validação externos
  - Avaliam o nível de compatibilidade entre uma partição obtida e uma partição de referência
  - Ex: Rand Index e Jaccard
- Critérios de validação internos
  - Avaliam a estrutura de grupos obtida utilizando apenas os próprios dados
  - Ex: SSE
- Critérios de validação relativos
  - Indicam qual a melhor dentre duas ou mais partições
  - Ex: Índice Dunn e Silhueta

Trabalhos em AM

- Comunidade de AM cresceu muito nos últimos anos
  - Novos algoritmos foram propostos
  - Nro crescente de novas aplicações reais
  - *Frameworks* que permitem a comparação entre algoritmos existentes
    - Modificações nesses algoritmos são facilmente realizadas

Trabalhos em AM

- Um artigo típico de AM propõe
  - Um novo algoritmo
  - Uma modificação em um algoritmo existente
  - Um novo pré-processamento
  - Um novo pós-processamento

  Hipótese implícita: a nova proposta produz uma melhora no desempenho em comparação aos algoritmos da literatura.

Trabalhos em AM

- A avaliação em um artigo típico de AM
  - Usa um conjunto de *data sets* de teste
  - Usa medidas de avaliação apropriadas
  - Estabelece as conclusões com base nos *data sets* e nas medidas de avaliação

Lembrando que: não é possível encontrar um algoritmo que seja o melhor para todos os problemas → teorema "*no free lunch*"

Mas ....

Como podemos verificar estatisticamente a hipótese de que o desempenho será melhorado com a nova proposta?

Questões a serem tratadas

- Dados dois algoritmos de aprendizado A e B e um *data set*, qual é o melhor algoritmo?

- Dados vários algoritmos de aprendizado e vários *data sets*, qual deles é o melhor?
  - Há diferença estatística entre os resultados dos algoritmos?

Questões a serem tratadas

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	92.6
Avg	**44.8**	**80.0**	**82.4**	**78.0**	**82.1**	**81.8**	**81.7**

Grandes variações na acurácia de diferentes classificadores!

O algoritmo 3 é o melhor porque ele obtém a melhor média?

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Questões a serem tratadas

	Alg. 1	Alg. 2	Alg. 3	Alg. 4	Alg. 5	Alg. 6	Alg. 7
aud	25.3	76.0	68.4	69.6	79.0	**81.2**	57.7
aus	55.5	81.9	85.4	77.5	85.2	83.3	**85.7**
bal	45.0	76.2	87.2	**90.4**	78.5	81.9	79.8
bpa	58.0	63.5	60.6	54.3	65.8	65.8	**68.2**
bps	51.6	83.2	82.8	78.6	80.1	79.0	**83.3**
bre	65.5	96.0	**96.7**	96.0	95.4	95.3	96.0
cmc	42.7	44.4	46.8	50.6	52.1	49.8	52.3
gls	34.6	66.3	66.4	47.6	65.8	69.0	**72.6**
h-c	54.5	77.4	83.2	**83.6**	73.6	77.9	79.9
hep	79.3	79.9	80.8	83.2	78.9	80.0	83.2
irs	33.3	**95.3**	**95.3**	94.7	**95.3**	95.3	94.7
krk	52.2	89.4	94.9	87.0	98.3	98.4	98.6
lab	65.4	81.1	92.1	**95.2**	73.3	73.9	75.4
led	10.5	62.4	75.0	74.9	**74.9**	75.1	74.8
lym	55.0	83.3	83.6	**85.6**	77.0	71.5	79.0
mmg	56.0	63.0	**65.3**	64.7	64.8	61.9	63.4
mus	51.8	**100.0**	**100.0**	96.4	**100.0**	**100.0**	99.8
mux	49.9	78.6	99.8	61.9	99.9	**100.0**	**100.0**
pmi	65.1	70.3	73.9	75.4	73.1	72.6	76.0
prt	24.9	34.5	42.5	**50.8**	41.6	39.8	43.7
seg	14.3	**97.4**	96.1	80.1	97.2	96.8	96.1
sick	93.8	96.1	96.3	93.3	**98.4**	97.0	96.7
soyb	13.5	89.5	90.3	**92.8**	91.4	90.3	76.2
tao	49.8	**96.1**	96.0	80.8	95.1	93.6	88.4
thy	19.5	68.1	65.1	80.6	**92.1**	**92.1**	86.3
veh	25.1	69.4	69.7	46.2	73.6	72.6	72.2
vote	61.4	92.4	92.6	90.1	96.3	**96.5**	95.4
vow	9.1	99.1	**96.6**	65.3	80.7	78.3	87.6
wne	39.8	95.6	96.8	**97.8**	94.6	92.9	96.3
zoo	41.7	94.6	92.5	**95.4**	91.6	92.5	92.6
Avg	**44.8**	**80.0**	**82.4**	**78.0**	**82.1**	**81.8**	**81.7**

- O algoritmo 4 é o vencedor em 8 problemas com média de 78.0
- O algoritmo 2 é o vencedor para 4 problemas com média 80.0
- Qual o melhor entre eles?

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Importante

- O teste estatístico pode ser usando tanto em tarefas de classificação quanto agrupamento

.... Vamos falar sobre classificação

Taxonomia das questões estatísticas em AM (Dietterich, 1998)

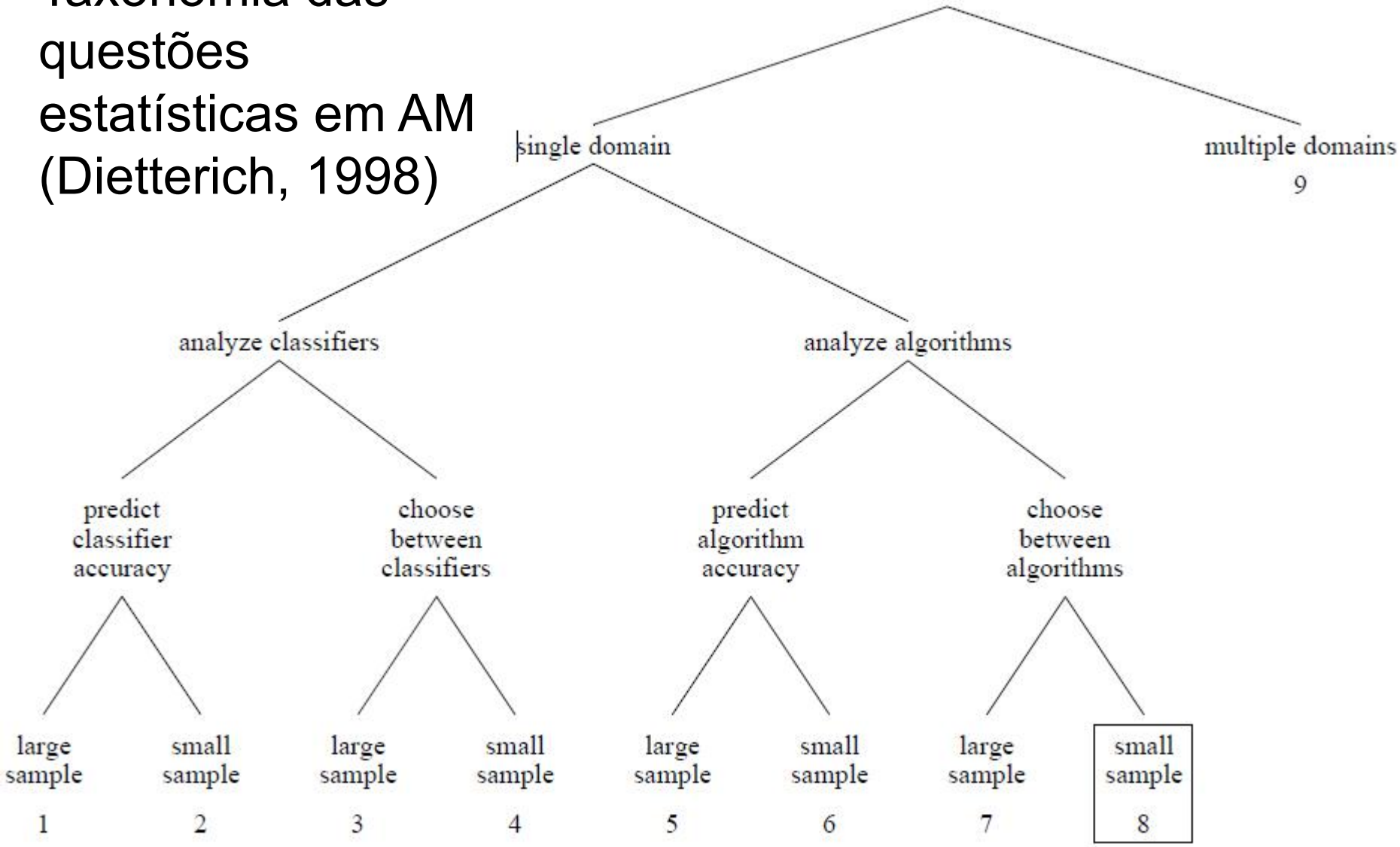


Imagem retirada do paper: Dietterich, T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Computation, Vol. 10, p. 1895-1923, 1998

Diferenciando

- Classificador
  - Função que dado um exemplo (entrada) associa uma classe àquele exemplo
- Algoritmo de aprendizado
  - Dado um conjunto de exemplos rotulados, isto é, com as suas classes, constrói um classificador
- Em aplicações específicas, o objetivo principal é
  - Encontrar o melhor classificador e estimar sua acurácia em exemplos futuros

Questões estatísticas para problemas de classificação (Dietterich, 1998)

- *Q7: Dados dois algoritmos A e B e um *data set* grande S, qual algoritmo irá produzir o melhor classificador quando treinado em *data sets* de tamanho m?
  - Dividir S em conjuntos de treino disjuntos e um conjunto de treino

- *Q8: Dados dois algoritmos A e B e um *data set* pequeno S, qual algoritmo irá produzir o melhor classificador quando treinado em *data sets* de tamanho S?
  - Usar, por exemplo, *holdout*
  - Vários testes estatísticos

Questões estatísticas para problemas de classificação (Dietterich, 1998)

- *Q9: Dados dois algoritmos A e B e *data sets* de diferentes domínios qual algoritmo irá produzir o melhor classificador quando treinado com exemplos de diferentes domínios
  - Questão difícil em AM
  - Há testes estatísticos específicos

Definições

- Hipótese Nula
  - É hipótese que os parâmetros ou características matemáticas de duas ou mais populações são idênticas
- Erro tipo I
  - Ocorre quando a hipótese nula é rejeitada, sendo que na verdade ela é verdadeira
- Erro tipo II
  - Ocorre quando a hipótese nula não é rejeitada, sendo que na verdade ela é falsa

Definições

- Fontes de variação
  - Variação aleatória nos dados de teste usados para avaliar o algoritmo
    - Em um dado conjunto de teste, criado aleatoriamente, um classificador pode obter melhor desempenho que outro, mesmo que na população total eles tivessem desempenho idêntico
  - Variação aleatória devido a escolha do conjunto de treino
    - Em um dado conjunto de treino, criado aleatoriamente, um classificador pode obter melhor desempenho que outro, mesmo que em média os dois classificadores tenham a mesma acurácia

Definições

- Fontes de variação
  - Variação devido à aleatoriedade do algoritmo de aprendizado.
    - Ex: inicialização de pesos em redes neurais
  - Variação devido a erros de classificação aleatória
    - Se uma fração fixa de η elementos do conjunto de teste são aleatoriamente classificados incorretamente, então nenhum algoritmo de aprendizado pode alcançar uma taxa de erro menor que η

Formalizando o problema (Dietterich, 1998)

X: População

f: Função alvo - classifica cada elemento x em uma dentre K classes

Na aplicação: Uma amostra S é gerada aleatoriamente a partir de X de acordo com uma distribuição de probabilidade D

Dados de treino: dados contendo (x,f(x))

Algoritmo de aprendizado: recebe um conjunto de dados de treino R e produz um classificador f'

Erro: pegar a amostra S e subdividi-la em treino R e teste T. A taxa de erro f' em T fornece uma estimativa de erro de f' na população

Formalizando o problema (Dietterich, 1998)

Hipótese nula a ser testada:

Para um conjunto de treino R, de tamanho fixo e escolhido aleatoriamente, dois algoritmos de aprendizado terão o mesmo erro em um conjunto de teste, escolhido aleatoriamente a partir de X.

Tipos de teste

- Paramétricos – suposições
  - As observações devem ser independentes
  - As observações devem ser geradas por uma distribuição normal
  - As populações devem ter a mesma variância
  - Ex: t-test e ANOVA

Tipos de teste

- Não-Paramétricos – suposições
  - As observações devem ser independentes
  - Os dados devem ser representados por número ordinais

A maioria dos testes não paramétricos usam *ranks* ao invés dos dados brutos para testar suas hipóteses

A seguir .....

Testes estatísticos para comparação entre classificadores em um único domínio → ver Dietterich (1998)

Testes Estatísticos

- *McNemar's Test*
  - Dividir a amostra S em treino R e teste T
  - Treinar os algoritmos A e B, produzindo os classificadores f'A e f'B
  - Testar os classificadores no conjunto de teste
  - Construir a tabela de contingência

Nro de exemplos incorretamente classificados em f'A e f'B	Nro de exemplos incorretamente classificados em f'A, mas não em f'B
Nro de exemplos incorretamente classificados em f'B, mas não em f'A	Nro de exemplos incorretamente classificados nem por f'A nem por f'B

Testes Estatísticos

n_{00}	n_{01}
n_{10}	n_{11}

- *McNemar's Test*
  - Considerando a hipótese nula, os dois algoritmos terão o mesmo erro ($n_{01}=n_{10}$)
  - É baseado no teste qui-quadrado

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Se o valor calculado é maior ou igual ao tabelado → Rejeita-se H0

Se o valor calculado é menor que o tabelado → Aceita-se H0

Testes Estatísticos

- *McNemar's Test* – Exemplo

101	121
59	33

$$\chi^2 = \frac{(|121-59|-1)^2}{121+59} = 20.67$$

Consultando a tabela, X = 3,841 com n=1 (graus de liberdade)

Conclusão: rejeitar a hipótese nula

Tabela do Teste Qui-Quadrado

	$P(\chi^2_n \leq x)$													
n	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995	
1	3,93E-05	0,000157	0,000982	0,003932	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879	1
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597	2
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838	3
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860	4
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,832	15,086	16,750	5
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548	6
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278	7
8	1,344	1,647	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955	8
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589	9
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188	10
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757	11
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300	12
13	3,565	4,107	5,009	5,892	7,041	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819	13
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319	14
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801	15
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267	16
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718	17
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156	18
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582	19
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997	20
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401	21
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796	22
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181	23
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,558	24
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928	25
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290	26
27	11,808	12,878	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645	27
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,994	28
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,335	29
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672	30
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766	40
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490	50
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952	60
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215	70
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321	80
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299	90
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,170	100

Testes Estatísticos

- *McNemar's Test* – Problemas
  - Não mede a variabilidade devido a escolha do conjunto de treino nem a aleatoriedade do algoritmo de aprendizado
    - Um único conjunto de treino
  - Compara dois algoritmos em um conjunto de treino de tamanho |R|
    - |R| deve ser menor que |S|, para garantir um conjunto de teste grande
    - Assume-se que diferença relativa observada nos conjuntos de treino de tamanho |R| se mantém para conjuntos de tamanho |S|

Quando usá-lo: quando acredita-se que essas fontes de variabilidade são pequenas

Testes Estatísticos

- *Resampled paired t test*
  - Um dos mais populares em AM
  - Uma série de *trials* é conduzida (30)
    - A amostra S é divida em treino R (ex: 2/3 da base) e teste T
    - Os algoritmos A e B são treinados em R e testados em T
  - $p_A^{(i)}$ e $p_B^{(i)}$ → nro de exemplos de teste incorretamente classificados pelos algoritmos A e B, respectivamente, no *trial* i
  - Assume-se que as diferenças $p^{(i)} = p_A^{(i)} - p_B^{(i)}$ foram definidas independentemente a partir de uma distribuição normal, então pode-se aplicar o *Student's t test*

$$\overline{p} = \frac{1}{n}\sum_{i=1}^{n} p^{(i)}$$

$$t = \frac{\overline{p}.\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(p^{(i)} - \overline{p})^2}{n-1}}}$$

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365
33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307
41.	1.303	1.683	2.020	2.421	2.701	3.301
42.	1.302	1.682	2.018	2.418	2.698	3.296
43.	1.302	1.681	2.017	2.416	2.695	3.291
44.	1.301	1.680	2.015	2.414	2.692	3.286
45.	1.301	1.679	2.014	2.412	2.690	3.281

Tabela do *Student t-test*

Testes Estatísticos

- *Resampled paired t test* – Problemas
  - Diferenças individuais não terão uma distribuição normal
    - $p_A^{(i)}$ e $p_B^{(i)}$ não são independentes
  - $p^{(i)}$'s não são independentes
    - Há sobreposição entre os conjuntos de teste (e dos conjuntos de treino também) nos *trials*

Testes Estatísticos

- *K-fold cross-validation paired t test*
  - Dividir aleatoriamente S em k conjuntos disjuntos de igual tamanho $T_1,...,T_k$
  - Conduzir k *trials,* em que o conjunto de teste é T_i e o de treino é a união de todos os outros T_j, $j \neq i$
  - A mesma estatística *t* é computada

Vantagem:

- O conjunto de teste é independente dos outros, mas no de treino há sobreposição

Testes Estatísticos

- *The 5x2cv paired t test*
  - Realizar 5 replicações do 2-*fold-cross validation*
    - Os dados são particionados em dois conjuntos de tamanhos S_1 e S_2
    - Os algoritmo A e B são treinados em cada um dos conjunto e testados no outro
    - São produzidas quatro estimativas de erro: $p_A^{(1)}$ e $p_B^{(1)}$ (treinado em S_1 e testado em S_2) e $p_A^{(2)}$ e $p_B^{(2)}$ (treinado em S_2 e testado em S_1)
    - A partir das diferenças $p^{(1)}=p_A^{(1)} - p_B^{(1)}$ e $p^{(2)}=p_A^{(2)} - p_B^{(2)}$, a variância é estimada

$$s^2 = (p^{(1)} - \overline{p})^2 + (p^{(2)} - \overline{p})^2$$

$$\overline{p} = (p^{(1)} + p^{(2)})/2$$

$$\tilde{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}}$$

A seguir ....

Teste Estatísticos para comparação entre classificadores em múltiplos *data sets* → ver Demsar (2006)

- Comparação entre dois classificadores
- Comparação entre múltiplos classificadores

Diferenciando

- Testes usados para avaliar a diferença entre dois classificadores
  - **Único *data set***: computar o desempenho médio e sua variância sob repetidos conjuntos de treino e teste ou amostras aleatórias dos dados
  - **Múltiplos *data sets***: cada *data set* é usado para acessar o desempenho e não a variância. As fontes de variância são as diferenças no desempenho em (independentes) *data sets* e não só em amostras

Testes para comparar dois classificadores

- Comparar dois classificadores em múltiplos *data sets*
  - Exemplo
    - Comparar o C4.5 e um melhoramento do algoritmo C4.5
    - Medida de avaliação: AUC
    - Nro de *data sets*: 14

Testes Estatísticos

- *Wilcoxon Signed-Ranks Test*
  - Alternativa não-paramétrico ao *paired t-test*
  - Cria um *rank* com as diferenças no desempenho dos dois classificadores para cada conjunto de dados, ignorando o sinal, e compara os *ranks* para as diferenças positivas e negativas

$$R+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i) \qquad R- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$

d_i: diferença entre o *score* de desempenho de dois classificadores na i-ésima saída

$$T = \min(R+,R-)$$

Testes Estatísticos

- *Wilcoxon Signed-Ranks Test*
  - Há uma tabela para valores críticos exatos de T para N até 25
    - Se T <= valor tabela → Rejeitar Hipótese Nula
  - Para mais de 25 *data sets* usar

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

N: nro de graus de liberdade

Testes Estatísticos

- *Wilcoxon Signed-Ranks Test - Exemplo*

dataset	C4.5	C4.5m	Difference	Rank
Adult	0.763	0.768	+0.005	3.5
Breast	0.599	0.591	-0.008	7
Wisconsin	0.954	0.971	+0.017	9
Cmc	0.628	0.661	+0.033	12
Ionosphere	0.882	0.888	+0.006	5
Iris	0.936	0.931	-0.005	3.5
Bupa	0.661	0.668	+0.007	6
Lung	0.583	0.583	0.000	1.5
Lymphograph	0.775	0.838	+0.063	14
Mushroom	1.000	1.000	0.000	1.5
Tumor	0.940	0.962	+0.022	11
Rheum	0.619	0.666	+0.047	13
Voting	0.972	0.981	+0.009	8
Wine	0.957	0.978	+0.021	10

$R^+ = 3.5 + 9 + 12 + 5 + 6 + 14 + 11 + 13 + 8 + 10 + 1.5 = 93$

T = min(93,12) = 12

N=14

$R^- = 7 + 3.5 + 1.5 = 12$

Rejeitar Hipótese Nula

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Testes Estatísticos

	One Tailed Significance levels:		
	0.025	0.01	0.005
	Two Tailed significance levels:		
N	0.05	0.02	0.01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Tabela para o *Wilcoxon Test* com N < 25

Testes para comparar múltiplos classificadores

- Comparar múltiplos classificadores em múltiplos *data sets*
  - Ex: comparar 7 classificadores conduzindo 21 *paired t-test*
    - O algoritmo A é significativamente melhor que B e C
    - A e E são significativamente melhores que D

Problema: múltiplas hipóteses nula sendo testadas

Solução: usar testes como ANOVA ou Friedman

Teste Estatístico

- *Friedman test*
  - Teste não-paramétrico
  - Monta um *rank* dos algoritmos para cada *data set*
    - r_i^j é o rank do j-ésimo algoritmo (dentre k) no i-ésimo *data set* (dentre N)
  - Compara o *rank* médio dos algoritmos
  - Hipótese nula: todos os algoritmos são equivalentes e então os ranks Rj devem ser iguais

$$Rj = \frac{1}{N}\sum_i r_i^j \qquad X_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

Teste Estatístico

dataset	C4.5	C4.5m	C4.5cf	C4.5cf,m
Adult	0.763	0.768	0.771	0.798
Breast	0.599	0.591	0.590	0.569
Wisconsin	0.954	0.971	0.968	0.967
Cmc	0.628	0.661	0.654	0.657
Ionosphere	0.882	0.888	0.886	0.898
Iris	0.936	0.931	0.916	0.931
Bupa	0.661	0.668	0.609	0.685
Lung	0.583	0.583	0.563	0.625
Lymphography	0.775	0.838	0.866	0.875
Mushroom	1.000	1.000	1.000	1.000
Tumor	0.940	0.962	0.965	0.962
Rheum	0.619	0.666	0.614	0.669
Voting	0.972	0.981	0.975	0.975
Wine	0.957	0.978	0.946	0.970

Exemplo de uso do teste *Friedman*

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Teste Estatístico

dataset	C4.5	C4.5m	C4.5cf	C4.5cf,m
Adult	4	3	2	1
Breast	1	2	3	4
Wisconsin	4	1	2	3
Cmc	4	1	3	2
Ionosphere	4	2	3	1
Iris	1	2.5	4	2.5
Bupa	3	2	4	1
Lung	2.5	2.5	4	1
Lymphography	4	3	2	1
Mushroom	2.5	2.5	2.5	2.5
Tumor	4	2.5	1	2.5
Rheum	3	2	4	1
Voting	4	1	2	3
Wine	3	1	4	2
Average Rank	**3.143**	**2.000**	**2.893**	**1.964**

Ranks associados aos classificadores

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Teste Estatístico

- *Friedman test*

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] =$$

$$= \frac{12 \cdot 14}{4 \cdot 5}\left[9.878 + 4.000 + 8.369 + 3.857 - \frac{4 \cdot 25}{4}\right] =$$

$$= 9.28$$

Conclusão: observando os valores críticos → rejeitar a hipótese nula

Slide traduzido do original Statistical Analysis of Experiments in Data Mining and Computational Intelligence, Salvador García, Francisco Herrera

Teste Estatístico

- No *Friedman test*, se a hipótese nula é rejeitada, um pos-hoc teste pode ser aplicado
  - Quando todos os classificadores são comparados entre si → *Nemenyi test* pode ser usado
  - Quando todos os classificadores são comparados com classificador de controle (1xN) → *Bonferroni* pode ser usado

Ferramentas para realizar teste estatístico

- http://www.keel.es/

- R

Referências

- Demsar, J. Statistical Comparisons of Classifiers over Multiple Data Sets, Machine Learning Research, Vol. 7, p. 1-30, 2006.
- Dietterich, T. Supervised Classification Learning Algorithms, Neural Computation, Vol. 10, p. 1895-1923, 1998.
- García, S.; Fernández, A.; Luengo, J.; Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Information Sciencs, v. 180, 2044-2064, 2010.
- Apostila do curso de Estatística II – UFPR. Disponível em: http://www.est.ufpr.br/ce003/material/apostilace003.pdf