

# Mineração de Dados 2018.2

## Testes Estatísticos

Thiago Ferreira Covões

(slides baseados no material do Prof. Ethem Alpaydin e da Profa.  
Debora Medeiros [UFABC])

# Teste de Hipóteses

- Rejeitar a **hipótese nula** se não for suportada pelas amostradas com confiança suficiente

- $X = \{ x^t \}_t$

$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0 \text{ [} H_0 = \text{hipótese nula]}$$

Aceitar  $H_0$  com nível de significância  $\alpha$  se  $\mu_0$  está no  $100(1-\alpha)$  intervalo de confiança

	Decisão	
Verdade	Aceitar	Rejeitar
Sim	Correto	Erro Tipo I
Não	Erro Tipo II	Correto (poder)

# Testes de hipótese não-paramétricos

- Não tem premissas sobre as distribuições dos dados
  - Teste T assume normalidade
  - Difícil de ser verificado com poucas amostras
- Muito usados para avaliar classificadores em mais de uma base de dados
  - Erros podem não ser comensuráveis
- Vamos ver 2:
  - Wilcoxon Signed Rank: compara 2 métodos
  - Friedman: compara um grupo de métodos
    - Nemenyi como teste *post-hoc*

# *Wilcoxon signed-rank*

- 1) Assume que dados são pareados e oriundos da mesma população
- 2) Pares escolhidos aleatoriamente
- 3) Dados em escala intervalar
  - Diferença entre valores tem sentido

Hipótese nula:

*Diferença entre pares segue uma distribuição simétrica em volta do zero*

# *Wilcoxon signed-rank*

1) Cálculo das diferenças  $d_i$  das medidas de desempenho nos conjuntos de dados  $i$

2) Ranqueamento de  $|d_i|$

- Menores diferenças assumem primeiras posições
- Caso de empate: valores médios das posições

$$R^+ = \sum_{d_i > 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i)$$

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

# Exemplo

	C4.5	C4.5+m	difference	rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	-0.008	7
breast cancer 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
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary 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

# *Wilcoxon signed-rank*

3) Seja  $T = \min\{R+, R-\}$

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

- Hipótese de equivalência rejeitada, com 95% de confiança, se  $z < -1.96$

# *Wilcoxon signed-rank*

- No exemplo:
  - $R^- = 12$
  - $R^+ = 93$
  - $T = -2,542$
  - Portanto, podemos rejeitar a hipótese nula

# Comparando diversos modelos

- Maior número de comparações
  - Maior probabilidade de 1 deles detectar diferença estatística quando não há
    - Para  $J$  testes:  $1-(1-\alpha)^J$

# Comparando diversos modelos: Friedman

- Assume que dados são pelo menos ordinais
- Grupos são selecionados aleatoriamente

# Comparando diversos modelos: Friedman

Seja  $k$  o número de algoritmos e  $N$  o número de conjuntos de dados (idealmente  $k > 5$  e  $N > 10$ )

- 1)  $r_j^i$ : posição do desempenho algoritmo  $j$  no conjunto de dados  $i$
- 2)  $R_j$ : ranqueamento médio do algoritmo  $j$ 
  - $H_0$  afirma que todos os algoritmos são equivalentes

# Exemplo

	C4.5	C4.5+m	C4.5+cf	C4.5+m+cf
adult (sample)	0.763 (4)	0.768 (3)	0.771 (2)	0.798 (1)
breast cancer	0.599 (1)	0.591 (2)	0.590 (3)	0.569 (4)
breast cancer wisconsin	0.954 (4)	0.971 (1)	0.968 (2)	0.967 (3)
cmc	0.628 (4)	0.661 (1)	0.654 (3)	0.657 (2)
ionosphere	0.882 (4)	0.888 (2)	0.886 (3)	0.898 (1)
iris	0.936 (1)	0.931 (2.5)	0.916 (4)	0.931 (2.5)
liver disorders	0.661 (3)	0.668 (2)	0.609 (4)	0.685 (1)
lung cancer	0.583 (2.5)	0.583 (2.5)	0.563 (4)	0.625 (1)
lymphography	0.775 (4)	0.838 (3)	0.866 (2)	0.875 (1)
mushroom	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)	1.000 (2.5)
primary tumor	0.940 (4)	0.962 (2.5)	0.965 (1)	0.962 (2.5)
rheum	0.619 (3)	0.666 (2)	0.614 (4)	0.669 (1)
voting	0.972 (4)	0.981 (1)	0.975 (2)	0.975 (3)
wine	0.957 (3)	0.978 (1)	0.946 (4)	0.970 (2)
average rank	3.143	2.000	2.893	1.964

# Comparando diversos modelos: Friedman

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

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2}$$

- $H_0$  é rejeitada quando:
  - $F_F > F_{k-1, (k-1)(N-1)}$
  - Se  $H_0$  é rejeitada existe diferença de desempenhos, porém o teste não indica entre quais algoritmos
    - Pós-teste

Valores para distribuição F: <http://users.sussex.ac.uk/~grahamh/RM1web/F-ratio%20table%202005.pdf>

# Teste de Friedman

- No exemplo:
  - $\chi^2 = 9,28$
  - $F = 3,69$
  - Valor crítico  $F(3,39)$  para  $\alpha = 0,05$  é 2,85
  - Hipótese nula rejeitada

# Comparando diversos modelos: Pós- teste Nemenyi

- Desempenho de 2 algoritmos  $i$  e  $j$  é estatisticamente diferente se  $|R_i - R_j| \geq CD$  (*Critical Difference*):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

- onde  $q_\alpha$  são baseados na Studentized range. Por exemplo, para  $q_{0.05}$ :

$k$	2	3	4	5	6	7	8	9	10
Nemenyi	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164

# Nemenyi - Exemplo

- Nosso exemplo temos:
  - 4 algoritmos:  $q_\alpha = 2,569$
  - $CD = 1,25$
  - Maior diferença entre rankings:
    - $|3,143 - 1,964| = 1,179$
  - Se a maior diferença é menor que a diferença crítica, nosso pós-teste não é tem poder suficiente para detectar diferença entre os algoritmos

# Referências

- E. Alpaydin, Introduction to Machine Learning.
- Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7 (December 2006), 1-30.