Comparação de classificadores

Procedimento típico

- Validação cruzada k-fold estratificado
  - A cada iteração todos os modelos usam as mesmas partições de treinamento e validação
- Média de medida de desempenho obtida sob igualdade de condições: *acurácia, precisão, cobertura, medida-F, AUC,* etc..
- O modelo que obtiver melhor média, vence!

Problema

- O modelo que obtiver melhor média, vence?
- Muitas vezes, as diferenças não são significativas.

Teste de hipótese

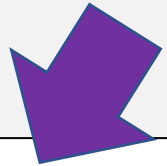
- Hipótese estatística é uma alegação sobre o valor de um ou mais parâmetros
  - Medidas de desempenho (alg. 1 e alg. 2): μ_1 e μ_2
  - Hipóteses (H): $\mu_1 - \mu_2 = 0$ ou $\mu_1 - \mu_2 > 0$ ou etc..
- Normalmente há duas suposições contraditórias:
  - $H_0 : \mu_1 - \mu_2 = 0$ (hipótese *nula,* inicialmente assumida como verdadeira)
  - *vs.* $H_1 : \mu_1 - \mu_2 \neq 0$ (*hipótese alternativa*):
- O teste visa então rejeitar ou não a Hipótese *nula*

Procedimento de teste

- Regra para decidir se H_0 deve ser aceita
- Possui:
  - Estatística de teste (em função dos dados da amostra em que a decisão se baseia)
  - Região de rejeição (representa o conj. de val. da estatística de teste p/ os quais H_0 é rejeitada. H_0 é rejeitada se valor da estat. calc. cair na região de rejeição.)
- Dois tipos de erro: Tipo I (H_0 =V rejeitada) e Tipo II (H_0 =F ñ rejeitada)
- A região de rejeição é calculada de modo a manter a probabilidade α de ocorrência de erro **Tipo I** sob controle

Procedimento de teste

- …
- Dois tipos de erro: Tipo I (H_0 =V rejeitada) e Tipo II (H_0 =F ñ rejeitada)
- A região de rejeição é calculada de modo a manter a probabilidade α de ocorrência de erro **Tipo I** sob controle



Nível de significância

α = 0.05 → 95% de confiança de não ter cometido erro Tipo I

Procedimentos de teste para ML

- *OBS*: ainda **não há consenso** sobre melhor procedimento, pois amostras apresentam dependências!
- Testes **não paramétricos** → ñ há restrição de que as amostras sigam alguma distribuição conhecida (ex: Normal)
- Bastante utilizado: *Wilcoxon signed-rank*
  - Baseado em ranqueamento e permite adição de outras medidas de desempenho (ex: tempo de treinamento)
- Dois cenários de testes:
  - Conjunto de dados específico é o alvo → usar único conjunto de dados
  - Algoritmo é o alvo → usar vários conjuntos de dados (preferível)

Comparação de dois modelos: *Wilcoxon signed-rank*

H_0 : modelos A e B são equivalentes

1. Aplica-se A e B a alguns conjuntos de dados ($i = 1...N$)
2. Calcula-se $d_i = \mu_B - \mu_A$
3. Ranqueia-se via $|\mu_B - \mu_A|$. Havendo empate, atribui-se valores médios

Conj. dados	C4.5	C4.5+m	Diferença	Dif_absoluta	Posição
Pulmão	0,583	0,583	0,000	0,000	1,5
Fungo	0,583	0,583	0,000	0,000	1,5
Atmosfera	0,882	0,888	+0,006	0,006	3,0
Mama	0,599	0,591	−0,008	0,008	4,0

Comparação de dois modelos: *Wilcoxon signed-rank*

4. Calcula-se $R+$ e $R-$

$$R+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$ (B melhor que A)

$$R- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i)$$ (A melhor que B)

5. Seja S a menor dessas somas. Livros de Estatística trazem tabelas com os valores críticos exatos para S, com N variando de 1 até 25. Para mais conjuntos de dados, a estatística do teste seria:

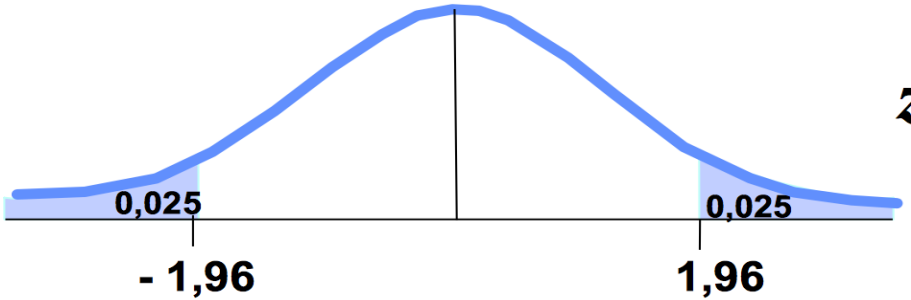
$$z = \frac{S - \frac{1}{4}N(N-1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$$

Comparação de dois modelos: *Wilcoxon signed-rank*

Com $\alpha = 0.05$,

H_0 pode ser rejeitada se $z < -1.96$

Grau de Confiança	α	Valor Crítico $z_{\alpha/2}$
90%	0,10	1,645
95%	0,05	1,96
99%	0,01	2,575



$z_{\alpha/2} = \pm 1,96$

Divulgação científica

saense

http://www.saense.com.br
https://www.facebook.com/saense/

Hendrik Macedo
Escreve sobre Inteligência Artificial no Saense.
http://www.saense.com.br/autores/artigos-publicados-por-hendrik-macedo/