Aprendizado de Máquina: Comparação de Classificadores

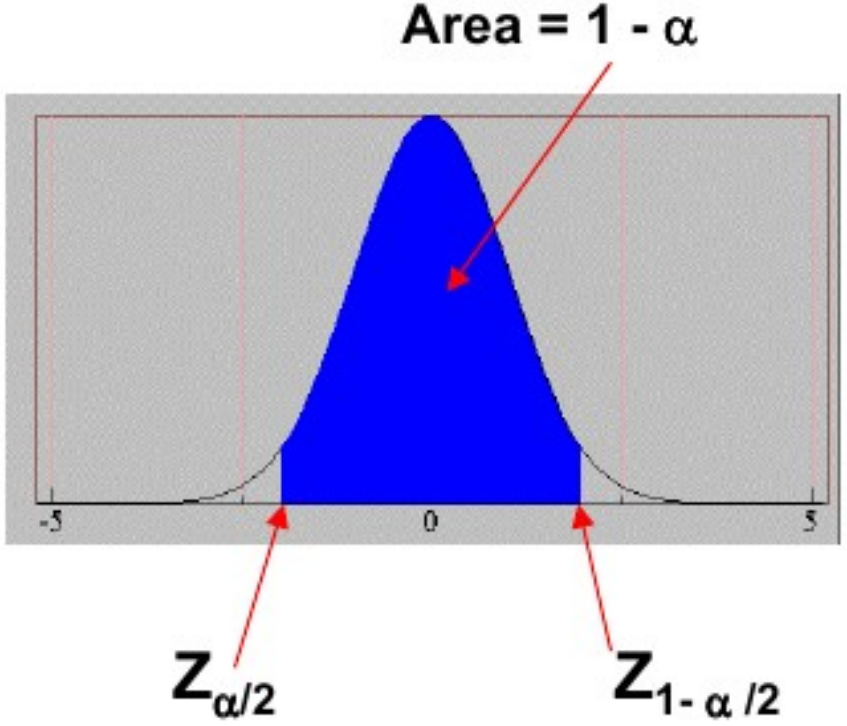
Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Intervalo de Confiança

- Conjunto de teste é uma pequena fatio do mundo real
- Quão confiável são medidas como acurácia, medida-f, precisão, etc, extraídas desse conjunto?
- A medida a ser avaliada é considerada uma variável aleatória
  - Retorna diferentes valores quando avaliada com diferentes instâncias do dataset

Intervalo de Confiança (2)

- Ideia geral: calcular intervalo de confiança
- Caso particular: medições formam uma distribuição normal (ou gaussiana)
  - Média 0
  - Desvio padrão 1



Intervalo de Confiança (3)

- Convenções:
  - **x**: variável aleatória sendo medida
  - **c**: a confiança que desejamos do intervalo (quanto maior, maior o intervalo)
  - **z**: representa os limites do intervalo
  - $p(-z \leq x \leq +z) = c$

Intervalo de Confiança (4)

- Exemplo: desejamos uma confiança de 90%
  - $p(-1.65 \leq x \leq 1.65) = 0.9$
- De onde vêm os valores de z?
  - Previamente calculados em tabelas estatística, conforme a distribuição dos dados
  - Convenção usada nos livros: $p(x \geq z) = (1-c)/2$

Intervalo de Confiança (5)

- No exemplo, confiança de 90% resulta que (1-c)/2 = 0.05, logo z=1.65

Table 5.1 Confidence Limits for the Normal Distribution

$\Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

Confiança da Acurácia

- Técnica apresentada pode ser adaptada para as medidas de avaliação de classificadores estudadas
- Exemplo para acurácia: classificador ou acerta ou erra cada instância do dataset
  - Processo parecido com lançar uma moeda
  - Classificador bom é como uma moeda viciada: acerta mais do que erra
  - Nome técnico: ensaio de Bernoulli

Confiança da Acurácia (2)

- Distribuição de Bernoulli (ou binomial): se a taxa de acerto do classificador sobre **n** instâncias de teste é **p**, então:
  - Média esperada: p
  - Variância esperada: p(1-p)/n
  - Desvio padrão: $\sqrt{(p(1-p)/n)}$

Confiança da Acurácia (3)

- Dataset suficientemente grande
  - A distribuição de Bernouli da acurácia aproxima razoavelmente bem uma distribuição normal
  - Porém, não tem media 0 e desvio 1 (conforme slide anterior)
  - Precisamos adaptar o cálculo do intervalo de confiança apresentado anteriormente

Confiança da Acurácia (4)

- Opção 1: subtrair a média e dividir pelo desvio padrão

$$p\left(-z < \frac{acc - p}{\sqrt{p\frac{(1-p)}{n}}} < z\right) = c$$

  - Onde **acc** é a acurácia real e **p** é a calculada no conjunto de teste
  - Desvantagem: mais difícil de interpretar

Confiança da Acurácia (5)

- Opção 2 (mais confiável): adaptar o cálculo para retornar os limites do intervalo de confiança

$$l = \frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p}{n} - \frac{p^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Confiança da Acurácia (6)

- Opção 3: quando n ≥ 30 e np(1-p) ≥ 5, o cálculo pode ser aproximado por:

$$l = p \pm z\sqrt{\frac{p(1-p)}{n}}$$

- Exercício: calcular os intervalos para os algoritmos estudados no semestre na base Iris com confiança de **90%**

Comparação de Modelos (2)

- Podemos estar interessados em comparar dois modelos ou dois algoritmos. Exemplos:
  - Modelo: rede neural treinada com topologia definida e pesos e biases fixados
  - Algoritmo: processo de treinamento para induzir uma dada rede neural

Comparação de Modelos (2)

- Comparando modelos
  - M_1: rede neural induzida pelo Backpropagation para resolver o Iris com hiperparâmetros fixos (ex.: 4 neurônios ocultos, semente 0, etc)
  - M_2: uma árvore de decisão induzida J48 para resolver o Iris
  - É possível dizer que M_1 é melhor/pior que M_2?

Comparação de Modelos (3)

- Opção 1: comparar diretamente intervalos de confiança. Análise mais trabalhosa
- Opção 2 (preferida): calcular d = p_1 - p_2
  - Se p_1 e p_2 aproximam distribuições normais, d também aproxima
  - Obs1: modelos não precisam ser treinados/testados sobre as mesmas instâncias
  - Obs2: estamos usando a análise bicaudal (diz se é melhor ou pior)

Comparação de Modelos (4)

- Variância esperada:

$$\sigma_d^2 = \sigma_{p_1}^2 + \sigma_{p_2}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- Desvio esperado

$$\sigma_d = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Comparação de Modelos (5)

- O intervalo de confiança de **d** é dado por:

$$l = d \pm z\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  z é o valor tabelado para (1-c)/2

- Intervalo contém 0: não é possível afirmar qual é melhor/pior e **nem mesmo se são equivalentes**
- Intervalo positivo: M_1 é melhor com confiança c
- Intervalo negativo: M_1 é pior com confiança c

Comparação de Modelos (5)

- Exemplo:
  - Modelo M_1: acurácia = 85%; testada em 30 exemplos
  - Modelo M_2: acurácia = 75%; testada em 5000 exemplos
  - É possível dizer que M_1 é melhor que M_2 com 90% de confiança?

Comparação de Modelos (6)

- M_1: n_1= 30 e p_1 = 0.85
- M_2: n_2= 5000 e p_2 = 0.75
- d = $p_1 - p_2$ = 0.1
- $\sigma^2 = \frac{0.85(1-0.85)}{30} + \frac{0.75(1-0.75)}{5000} = 0.0042875$
- $\sigma = \sqrt{0.0042875} = 0.065$

Comparação de Modelos (7)

- $d_t = 0.1 \pm 1.65 \times 0.065 = 0.1 \pm 0.107$
- Como intervalo contém 0, diferença não é estatisticamente significante

Comparação de Modelos (8)

- Modelo M_1: $p_1 = 70\%$; $n_1 = 100$
- Modelo M_2: $p_1 = 85\%$; $n_2 = 200$
- c = 99%
- M2 é melhor do que M1?

Table 5.1 Confidence Limits for the Normal Distribution

$Pr[X \geq z]$	z
0.1%	3.09
0.5%	2.58
1%	2.33
5%	1.65
10%	1.28
20%	0.84
40%	0.25

Comparação de Algoritmos

- Muitas vezes não desejamos saber qual modelo é melhor, mas sim qual algoritmo é capaz de induzir modelos melhores
- Estratégia
  - Induzir vários modelos usando método 1 (ex.: rede neural)
  - Induzir vários modelos usando método 2 (ex.: árvore de decisão)
  - Calcular média e desvio das acurácias e comparar resultados

Comparação de Algoritmos (2)

- Aplicação: útil para dizer qual algoritmo é melhor para resolver uma tarefa específica
  - Quebrar um grande dataset em pedaços menores
  - Ou usar datasets semelhantes (com atributos parecidos)

Comparação de Algoritmos (3)

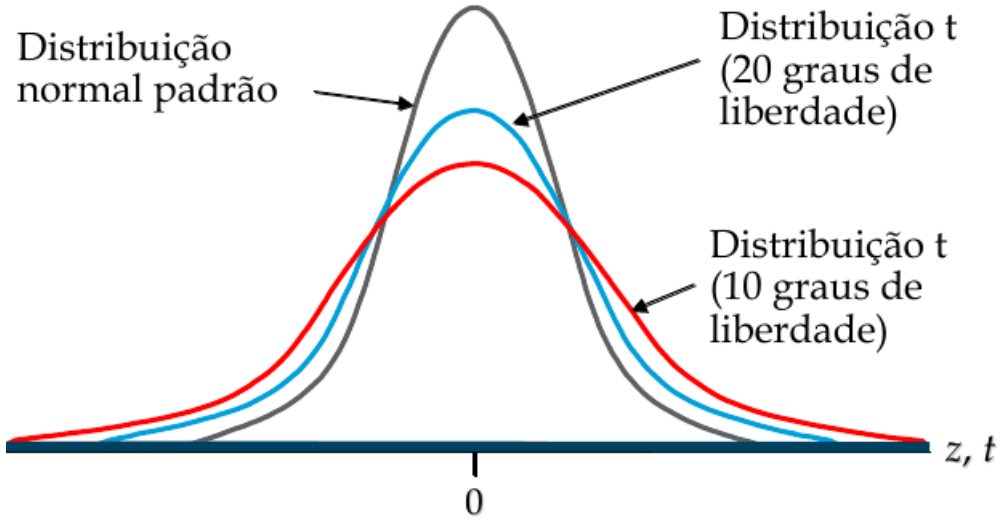
- Outra aplicação: dizer qual algoritmo se sai melhor uma área de aplicação (ex.: diagnóstico de doenças)
  - Usar datasets diferentes dentro da a área de aplicação
- Problema: com menos de 30 datasets, dificilmente acurácias medidas vão formar uma distribuição normal
- Alternativa: usar a distribuição T-student (Teste T)

Comparação de Algoritmos (4)

- Distribuição T-student
  - Família de distribuições de probabilidade semelhante a distribuição normal
  - Distribuição da família depende do parâmetro graus de liberdade (k-1)
    - Número k de datasets usados ou
    - Número k de folds quando um dataset grande é quebrado em pedaços menores

Comparação de Algoritmos (5)

- Quanto maior o número de graus
  - Menor a dispersão
  - Mais semelhante se torna a uma distribuição normal



Comparação de Algoritmos (6)

- Valores de z de acordo com o c desejado (versão com duas caudas)

	(I-α)				
K – 1	0.80	0.90	0.95	0.98	0.99
1	3.08	6.31	12.7	31.8	63.7
2	1.89	2.92	4.30	6.96	9.92
9	1.38	1.83	2.26	2.82	3.25
29	1.31	1.70	2.04	2.46	2.76

Comparação de Algoritmos (7)

- Exemplo: acurácias de teste estimadas para 2 algoritmos usando 30 datasets
  - d = 0.05
  - σ = 0.002
- As diferenças são significativas com 95% de confiança?

Comparação de Algoritmos (8)

- $d_t = 0.05 \pm 2.04 \times 0.002$
- Como intervalo não inclui valor 0, a diferença é estatisticamente significante com 95% de confiança

Comparação de Algoritmos (9)

- Diferença de acurácias extraída entre 2 algoritmos
  - Treinados em 10 datasets
  - d possui média 0.06 e desvio padrão 0.003
  - As diferenças são significativas com 99% de confiança?

	(I-α)				
K – 1	0.80	0.90	0.95	0.98	0.99
1	3.08	6.31	12.7	31.8	63.7
2	1.89	2.92	4.30	6.96	9.92
9	1.38	1.83	2.26	2.82	3.25
29	1.31	1.70	2.04	2.46	2.76

Comparação de Algoritmos (10)

- Diferença de acurácias extraída entre 2 algoritmos
  - Treinados em 10 datasets
  - d possui média 0.06 e desvio padrão 0.003
  - As diferenças são significativas com 99% de confiança?

Comparação de Algoritmos (11)

- Limitação do Teste-T original: não pode extrair todo o potencial da validação cruzada
  - Datasets precisam ser independentes, mas na validação cruzada, instâncias são reusadas para treino
- Teste-T corrigido: capaz usar a média e o desvio padrão de acurácias obtidas nos folds da validação cruzada
  - Como se fossem vários datasets independentes

Comparação de Vários Algoritmos

- O Teste-T serve para comparar algoritmos dois a dois, mas não consegue dizer quando um algoritmo é o melhor dentre vários
- Existem outros métodos capazes de comparar vários classificadores de uma só vez
  - Teste de hipóteses de Feelders e Verkooijen
  - Teste de hipóteses de Friedman
  - ANOVA

Créditos

- Parcialmente adaptado de:
  - Notas de aula do Prof. Dr. André C. P. L. F. de Carvalho - ICMC-USP