Aprendizado de Máquina

Aula 5

http://www.ic.uff.br/~bianca/aa/

Tópicos

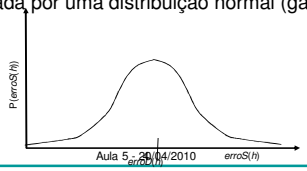
1. Introdução – Cap. 1 (16/03)
2. Classificação Indutiva – Cap. 2 (23/03)
3. Árvores de Decisão – Cap. 3 (30/03)
4. Ensembles - Artigo (13/04)
5. **Avaliação Experimental – Cap. 5 (20/04)**
6. Teoria do Aprendizado – Cap. 7 (27/04)
7. Aprendizado de Regras – Cap. 10 (04/05)
8. Redes Neurais – Cap. 4 (11/05)
9. Máquinas de Vetor de Suporte – Artigo (18/05)
10. Aprendizado Bayesiano – Cap. 6 e novo cap. online (25/05)
11. Aprendizado Baseado em Instâncias – Cap. 8 (01/05)
12. Classificação de Textos – Artigo (08/06)
13. Aprendizado Não-Supervisionado – Artigo (15/06)

Avaliação de hipóteses

- A acurácia da hipótese nos dados de treinamento é **tendenciosa** porque ela foi construída para concordar com esses dados.
- A acurácia deve ser avaliada nu**m conjunto de teste** independente (normalmente sem interseção).
- Quanto maior o conjunto de teste, mais precisa será a medida de acurácia e menos ela variará para conjuntos de teste diferentes.

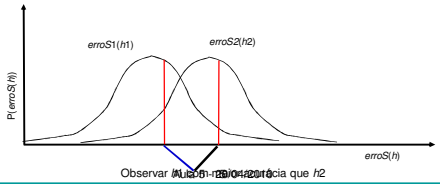
Variância na Acurácia de Teste

- Seja $erro_S(h)$ a porcentagem de exemplos num conjunto de teste S de tamanho n que são incorretamente classificados pela hipótese h.
- Seja $erro_D(h)$ a taxa de erro verdadeira para a distribuição de dados D.
- Quando n for pelo menos 30, o *teorema do limite central* garante que a distribuição de $erro_S(h)$ para amostras aleatórias diferentes poderá ser aproximada por uma distribuição normal (gaussiana).



Comparando duas hipóteses

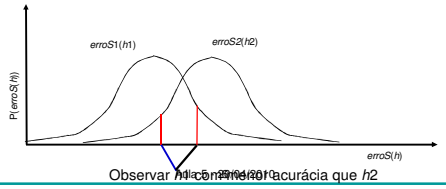
- Quando avaliamos duas hipóteses, a ordem observada com relação a acurácia pode ou não refletir a ordem verdadeira.
  - Suponha que h_1 seja testada em um conjunto de teste S_1 de tamanho n_1
  - Suponha que h_2 seja testada em um conjunto de teste S_2 de tamanho n_2



Observar $h1$ com menor acurácia que $h2$

Comparando duas hipóteses

- Quando avaliamos duas hipóteses, a ordem observada com relação a acurácia pode ou não refletir a ordem verdadeira.
  - Suponha que h_1 seja testada em um conjunto de teste S_1 de tamanho n_1
  - Suponha que h_2 seja testada em um conjunto de teste S_2 de tamanho n_2



Observar $h1$ com menor acurácia que $h2$

Teste Estatístico de Hipóteses

- Determina a probabilidade que uma diferença observada empiricamente em uma estatística seja devida somente ao acaso.
- Existem testes específicos para determinar a significância da diferença entre duas médias calculadas a partir de duas amostras.
- Os testes determinam a probabilidade da ***hipótese nula***, de que as duas amostras vieram da mesma distribuição.
- Pela convenção científica, ***rejeitamos a hipótese nula*** e dizemos que a diferença é ***estatisticamente significativa*** se a probabilidade da hipótese nula for menor que 5% ($p < 0.05$).

Teste lateral vs Teste bilateral

- O teste lateral supõe que a diferença esperada é em uma direção (A é melhor que B) e que a diferença observada é consistente com essa direção.
- O teste bilateral não supõe que existe uma direção esperada.
- O teste bilateral é mais conservador, já que requer uma diferença maior para acusar significância estatística.

Teste *z-score* para comparar hipóteses

- Suponha que h_1 seja testada em um conjunto S_1 de tamanho n_1 e que h_2 seja testada em um conjunto S_2 de tamanho n_2.
- Calcule a diferença entre a acurácia de h_1 e h_2

$$d = \left| erro_{S_1}(h_1) - erro_{S_2}(h_2) \right|$$

- Calcule o desvio padrão da estimativa da diferença

$$\sigma_d = \sqrt{\frac{erro_{S_1}(h_1)\cdot(1-erro_{S_1}(h_1))}{n_1} + \frac{erro_{S_2}(h_2)\cdot(1-erro_{S_2}(h_2))}{n_2}}$$

- Calcule *o z-score* da diferença $z = \frac{d}{\sigma_d}$

Teste *z-score* para comparar hipóteses

- Determine a confiança da diferença olhando a maior confiança, *C*, para um dado z-score na tabela.

confidence level	50%	68%	80%	90%	95%	98%	99%
z-score	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- Essa é a confiança para um teste de dois lados, para um teste de um lado aumenta-se a confiança de acordo com $C' = (100 - \frac{(100-C)}{2})$

Exemplo 1

Suponha que testamos duas hipóteses em conjuntos diferentes de tamanho 100 e observamos:

$$erro_{S_1}(h_1) = 0.20 \qquad erro_{S_2}(h_2) = 0.30$$

$$d = \left| erro_{S_1}(h_1) - erro_{S_2}(h_2) \right| = |0.2 - 0.3| = 0.1$$

$$\sigma_d = \sqrt{\frac{erro_{S_1}(h_1)\cdot(1-erro_{S_1}(h_1))}{n_1} + \frac{erro_{S_2}(h_2)\cdot(1-erro_{S_2}(h_2))}{n_2}}$$

$$= \sqrt{\frac{0.2\cdot(1-0.2)}{100} + \frac{0.3\cdot(1-0.3)}{100}} = 0.0608$$

$$z = \frac{d}{\sigma_d} = \frac{0.1}{0.0608} = 1.644$$

Confiança para o teste de duas caudas: 90%
Confiança para o teste de uma cauda: (100 – (100 – 90)/2) = 95%

Exemplo 2

Suponha que testamos duas hipóteses em conjuntos diferentes de tamanho 100 e observamos:

$$erro_{S_1}(h_1) = 0.20 \qquad erro_{S_2}(h_2) = 0.25$$

$$d = \left| erro_{S_1}(h_1) - erro_{S_2}(h_2) \right| = |0.2 - 0.25| = 0.05$$

$$\sigma_d = \sqrt{\frac{erro_{S_1}(h_1)\cdot(1-erro_{S_1}(h_1))}{n_1} + \frac{erro_{S_2}(h_2)\cdot(1-erro_{S_2}(h_2))}{n_2}}$$

$$= \sqrt{\frac{0.2\cdot(1-0.2)}{100} + \frac{0.25\cdot(1-0.25)}{100}} = 0.0589$$

$$z = \frac{d}{\sigma_d} = \frac{0.05}{0.0589} = 0.848$$

Confiança para o teste de duas caudas: 50%
Confiança para o teste de uma caudas: (100 – (100 – 50)/2) = 75%

Suposições do teste z-score

- Hipóteses podem ser testadas em conjuntos de teste diferentes; se o mesmo conjunto for usado, conclusões mais fortes podem ser tiradas.
- Conjunto de testes tem pelo menos 30 exemplos.
- Hipóteses foram aprendidas a partir de conjuntos de treinamento independentes.
- Apenas compara duas hipóteses sem levar em consideração o método de aprendizado utilizado. Não compara os métodos em geral.

Comparando algoritmos de aprendizado

- Para comparar a acurácia de hipóteses produzidas por algoritmos diferentes, idealmente, queremos medir:

$$E_{S \subset D}(erro_D(L_A(S)) - erro_D(L_B(S)))$$

onde $L_X(S)$ representa a hipótese aprendida pelo método L a partir de dados S.

- Para estimar isso, temos que ter múltiplos conjuntos de treinamento e teste.
- Como os dados são limitados, normalmente fazemos várias divisões dos dados em treinamento e teste.

Validação Cruzada

Particione os dados *D* em *k* conjuntos disjuntos de tamanho igual.

subconjuntos *P1…Pk*

Para *i* de 1 a *k faça*:
Use *Pi* como conjunto de teste e o resto dos dados para treinamento

Si = (*D* – *Pi*)
hA = *LA*(*Si*)
hB = *LB*(*Si*)
δ*i* = error*Pi*(*hA*) – error*Pi*(*hB*)

Retorne a diferença média de erros:

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

Comentários sobre validação cruzada

- Cada exemplo é usado uma vez como exemplo de teste e *k*–1 vezes como exemplo de treinamento.
- Todos os conjuntos de teste são independentes; mas os conjuntos de treinamento não são.
- Mede a acurácia de hipóteses treinadas com [(*k*–1)/*k*]·|D| exemplos.
- Método padrão é 10-fold.
- Se *k* for muito pequeno, não teremos um número suficiente de conjuntos; se *k* for muito grande, conjunto de teste é pequeno.
- Se *k*=|D|, método é chamado de validação cruzada ***leave-one-out***.

Teste de significância

- Tipicamente *k*<30, então não podemos usar o teste z.
- Podemos usar o ***t-test de Student***, que é melhor quando o número de amostras é menor.
- Podemos usar o ***t-test pareado***, que determina que diferenças menores são significativas quando os algoritmos usam os mesmos dados.
- Porém eles assumem que as amostras são independentes, o que não é verdade para a validação cruzada.
  - Conjuntos de teste são independentes
  - Conjuntos de treinamento NÃO são independentes
- Outros testes foram propostos, como o teste de McNemar.
- Apesar de nenhum teste ser perfeito, temos que de alguma forma levar a variância em consideração.

Exemplo

Que experimento dá mais evidência que SystemA é melhor que SystemB?

Experimento 1

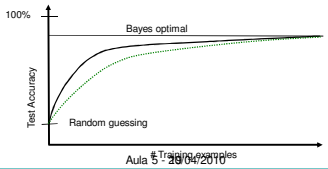
	SystemA	SystemB	Diff
Trial 1	87%	82%	+5%
Trail 2	83%	78%	+5%
Trial 3	88%	83%	+5%
Trial 4	82%	77%	+5%
Trial 5	85%	80%	+5%
Average	85%	80%	+5%

Experimento 2

	SystemA	SystemB	Diff
Trial 1	90%	82%	+8%
Trail 2	93%	76%	+17%
Trial 3	80%	85%	–5%
Trial 4	85%	75%	+10%
Trial 5	77%	82%	– 5%
Average	85%	80%	+5%

Curvas de Aprendizado

- Mostra acurácia vs. tamanho do conjunto de treinamento.
- Mostra se a acurácia máxima (ótimo de Bayes) está sendo alcançada.
- Mostra se um algoritmo é melhor pra dados limitados.
- A maioria dos algoritmos converge ao ótimo de Bayes com dados suficientes..

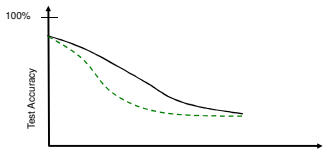


Curvas de Aprendizado com Validação Cruzada

Split data into *k* equal partitions
For trial *i* = 1 to *k* do:
  Use partition *i* for testing and the union of all other partitions for training.
  For each desired point *p* on the learning curve do:
    For each learning system *L*
      Train *L* on the first *p* examples of the training set and record training time, training accuracy, and learned concept complexity.
      Test *L* on the test set, recording testing time and test accuracy.
Compute average for each performance statistic across *k* trials.
Plot curves for any desired performance statistic versus training set size.
Use a paired t-test to determine significance of any differences between any two systems for a given training set size.

Curvas de Ruído

- Medem a resistência do algoritmo a ruído.
- Adicionar artificialmente ruído nas classes ou atributos, aleatoriamente substituindo uma fração dos valores por valores aleatórios. values.



Conclusões

- Important to test on a variety of domains to demonstrate a general bias that is useful for a variety of problems. Testing on 20+ data sets is common.
- Variety of freely available data sources
  - UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - KDD Cup (large data sets for data mining) http://www.kdnuggets.com/datasets/kddcup.html
  - CoNLL Shared Task (natural language problems) http://www.ifarm.nl/signll/conll/
- Data for real problems is preferable to artificial problems to demonstrate a useful bias for real-world problems.
- Many available datasets have been subjected to significant feature engineering to make them learnable.