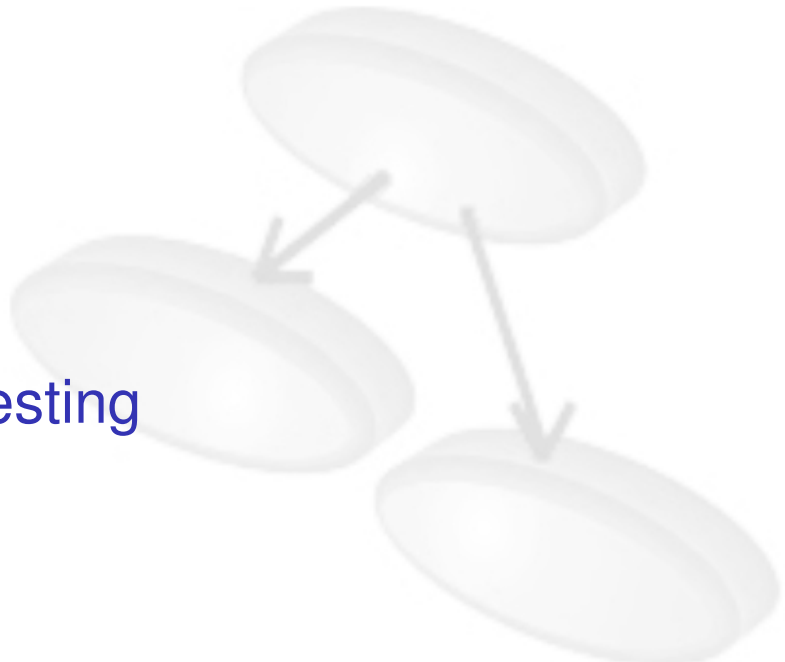
Honest Evaluation of Classification Models

Jose A. Lozano, Guzmán Santafé, Iñaki Inza

Intelligent Systems Group
The University of the Basque Country

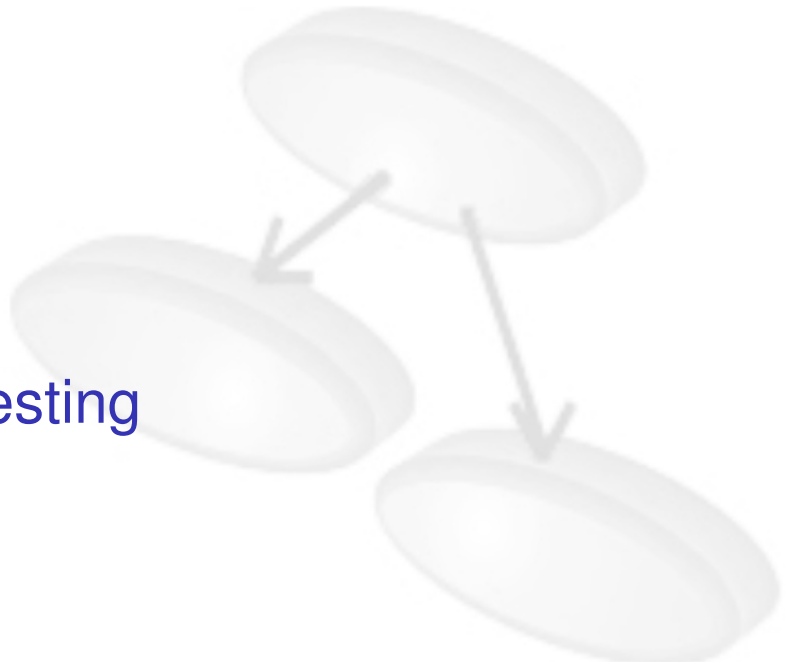
Asian Conference on Machine Learning (ACML'10)
November 8-10, 2010

Outline of the Tutorial



1. Hypothesis Testing

Outline of the Tutorial



Motivation

Basic Concepts

- Hypothesis testing form the basis of scientific reasoning in experimental sciences
- They are used to set scientific statements
- A hypothesis H_o called null hypothesis is tested against another hypothesis H_1 called alternative
- The two hypotheses are not at the same level: reject H_o does not mean acceptance of H_1
- The objective is to know when the differences in H_0 are due to randomness or not

Hypothesis Testing

Possible Outcomes of a Test

- Given a sample, a decision is taken about the null hypothesis (H_0)
- The decision is taken under uncertainty

	H_0 TRUE	H_0 FALSE
Decision: ACCEPT	$\surd$	Type II error (β)
Decision: REJECT	Type I error (α)	$\surd$

Hypothesis Testing: An Example

A Simple Hypothesis Test

- A natural process is given in nature that follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$
- We have a sample of this process $\{x_1, \ldots, x_n\}$ and a decision must be taken about the following hypotheses:

$$\begin{cases} H_0 : \mu = 60 \\ H_1 : \mu = 50 \end{cases}$$

- A statistic (function) of the sample is used to take the decision. In our example $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Hypothesis Testing: An Example

Accept and Reject Regions

- The possible values of the statistic are divided in accept and reject regions

$$A.R. = \{(x_1, \ldots, x_n) | \overline{X} > 55\}$$

$$R.R. = \{(x_1, \ldots, x_n) | \overline{X} \leq 55\}$$

- Assuming a probability distribution on the statistic $\overline{X}$ (it depends on the distribution of $\{x_1, \ldots, x_n\}$) the probability of each error type can be calculated:

$$\alpha = P_{H_0}(\overline{X} \in R.R.) = P_{H_0}(\overline{X} \leq 55)$$

$$\beta = P_{H_1}(\overline{X} \in A.R.) = P_{H_1}(\overline{X} > 55)$$

Hypothesis Testing: An Example

Accept and Reject Regions

- The A.R. and R.R. can be modified in order to have a particular value of α:

$$0{,}1 = \alpha = P_{H_0}(\overline{X} \in R.R.) = P_{H_0}(\overline{X} \leq 51)$$

$$0{,}05 = \alpha = P_{H_0}(\overline{X} \in R.R.) = P_{H_0}(\overline{X} \leq 50{,}3)$$

- *p*-value. Given a sample and the specific value of the test statistic $\overline{x}$ for the sample:

$$p\text{-value} = P_{H_0}(\overline{X} \leq \overline{x})$$

Hypothesis Testing: Remarks

Power: $(1 - \beta)$

- Depending on the hypotheses the type II error (β) can not be calculated:

$$\begin{cases} H_0 : \mu = 60 \\ H_1 : \mu \neq 60 \end{cases}$$

- In this case we do not know the value of μ for H_1 so we can not calculate the power $(1 - \beta)$
- A good hypothesis test: given an α the test maximises the power $(1 - \beta)$

Parametric test vs non-parametric test

Hypothesis Testing in Metaheuristic Optimization

Scenarios

- Two algorithms vs More than two
- One instance (problem) vs More than one instance (problem)
- .....

Testing Two Algorithms in Several Instances

Initial Approaches

- Averaging Over Datasets
- Paired t-test
  - $c^i = c_1^i - c_2^i$ and $\overline{d} = \frac{1}{N}\sum_{i=1}^{N} c^i$ then $\overline{d}/\sigma_{\overline{d}}$ follows a t distribution with $N-1$ degrees of freedom

Problems

- Commensurability
- Outlier susceptibility
- (t-test) Gaussian assumption

Testing Two Algorithms in Several Instances

Wilcoxon Signed-Ranks Test

- It is a non-parametric test that works as follows:
  1. Rank the module of the performance differences between both algorithms
  2. Calculate the sum of the ranks R^+ and R^- where the first (resp. the second) algorithm outperforms the other
  3. Calculate $T = min(R^+, R^-)$
- For $N \leq 25$ there are tables with critical values
- For $N > 25$

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \rightsquigarrow \quad \mathcal{N}(0,1)$$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598		
Instance2	0.599	0.591		
Instance3	0.954	0.971		
Instance4	0.628	0.661		
Instance5	0.882	0.888		
Instance6	0.936	0.931		
Instance7	0.661	0.668		
Instance8	0.583	0.583		
Instance9	0.775	0.838		
Instance10	1.000	1.000		

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591		
Instance3	0.954	0.971		
Instance4	0.628	0.661		
Instance5	0.882	0.888		
Instance6	0.936	0.931		
Instance7	0.661	0.668		
Instance8	0.583	0.583		
Instance9	0.775	0.838		
Instance10	1.000	1.000		

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971		
Instance4	0.628	0.661		
Instance5	0.882	0.888		
Instance6	0.936	0.931		
Instance7	0.661	0.668		
Instance8	0.583	0.583		
Instance9	0.775	0.838		
Instance10	1.000	1.000		

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971	+0.017	
Instance4	0.628	0.661	+0.033	
Instance5	0.882	0.888	+0.006	
Instance6	0.936	0.931	-0.005	
Instance7	0.661	0.668	+0.007	
Instance8	0.583	0.583	0.000	
Instance9	0.775	0.838	+0.063	
Instance10	1.000	1.000	0.000	

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971	+0.017	
Instance4	0.628	0.661	+0.033	
Instance5	0.882	0.888	+0.006	
Instance6	0.936	0.931	-0.005	
Instance7	0.661	0.668	+0.007	
Instance8	0.583	0.583	0.000	
Instance9	0.775	0.838	+0.063	
Instance10	1.000	1.000	0.000	

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971	+0.017	
Instance4	0.628	0.661	+0.033	
Instance5	0.882	0.888	+0.006	
Instance6	0.936	0.931	-0.005	
Instance7	0.661	0.668	+0.007	
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	
Instance10	1.000	1.000	0.000	1.5

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971	+0.017	
Instance4	0.628	0.661	+0.033	
Instance5	0.882	0.888	+0.006	
Instance6	0.936	0.931	-0.005	
Instance7	0.661	0.668	+0.007	
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	
Instance10	1.000	1.000	0.000	1.5

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	
Instance2	0.599	0.591	-0.008	
Instance3	0.954	0.971	+0.017	
Instance4	0.628	0.661	+0.033	
Instance5	0.882	0.888	+0.006	
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	
Instance10	1.000	1.000	0.000	1.5

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$$R^+ =$$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$$R^+ = 7 + 8 + 4 + 5 + 9 + 1/2(1{,}5 + 1{,}5)$$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$$R^+ = 34.5$$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$R^+ = 34.5 \qquad R^- = 10 + 6 + 3 + 1/2(1{,}5 + 1{,}5)$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$$R^+ = 34.5 \qquad R^- = 20.5$$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$R^+ = 34.5 \qquad R^- = 20.5 \qquad T = min(R^+, R^-)$

Wilcoxon Signed-Ranks Test: Example

	ψ^1	ψ^2	diff	rank
Instance1	0.763	0.598	-0.165	10
Instance2	0.599	0.591	-0.008	6
Instance3	0.954	0.971	+0.017	7
Instance4	0.628	0.661	+0.033	8
Instance5	0.882	0.888	+0.006	4
Instance6	0.936	0.931	-0.005	3
Instance7	0.661	0.668	+0.007	5
Instance8	0.583	0.583	0.000	1.5
Instance9	0.775	0.838	+0.063	9
Instance10	1.000	1.000	0.000	1.5

$R^+ = 34.5 \qquad R^- = 20.5 \qquad T = min(R^+, R^-) = 20.5$

Testing Two Algorithms in Several Instances

Wilcoxon Signed-Ranks Test

- It also suffers from commensurability but only qualitatively
- When the assumptions of the *t* test are met, Wilcoxon is less powerful than *t* test

Testing Two Algorithms in Several Instances

Signed Test

- It is a non-parametric test that counts the number of losses, ties and wins
- Under the null the number of wins follows a binomial distribution $B(1/2, N)$
- For large values of N the number of wins follows $\mathcal{N}(N/2, \sqrt{N/2})$ under the null
- This test does not make any assumptions
- It is weaker than Wilcoxon

Testing Several Algorithms in Several Instances

Dataset (Demšar, 2006)

	ψ^1	ψ^2	ψ^3	ψ^4
D_1	0.84	0.79	0.89	0.43
D_2	0.57	0.78	0.78	0.93
D_3	0.62	0.87	0.88	0.71
D_4	0.95	0.55	0.49	0.72
D_5	0.84	0.67	0.89	0.89
D_6	0.51	0.63	0.98	0.55

Testing Several Algorithms in Several Instances

Multiple Hypothesis Testing

- Testing all possible pairs of hypotheses $\mu_{\psi i} = \mu_{\psi j} \ \forall \ i, \ j$. Multiple hypothesis testing
- Testing the hypothesis $\mu_{\psi 1} = \mu_{\psi 2} = \ldots = \mu_{\psi k}$

Testing Several Algorithms in Several Instances

Multiple Hypothesis Testing

- Testing all possible pairs of hypotheses $\mu_{\psi i} = \mu_{\psi j} \ \forall \ i, \ j$. Multiple hypothesis testing
- Testing the hypothesis $\mu_{\psi 1} = \mu_{\psi 2} = \ldots = \mu_{\psi k}$

Testing Several Algorithms in Several Instances

Multiple Hypothesis Testing

- Testing all possible pairs of hypotheses $\mu_{\psi i} = \mu_{\psi j} \ \forall \ i, \ j$. Multiple hypothesis testing
- Testing the hypothesis $\mu_{\psi 1} = \mu_{\psi 2} = \ldots = \mu_{\psi k}$

ANOVA vs Friedman

- *Repeated measures* ANOVA: Assumes Gaussianity and sphericity
- Friedman: Non-parametric test

Testing Several Algorithms in Several Instances

Freidman Test

1. Rank the algorithms for each dataset separately (1-best). In case of ties assigned average ranks
2. Calculate the average rank R_j of each algorithm ψ^j
3. The following statistic:

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

follows a χ^2 with $k-1$ degrees of freedom (N>10, k>5)

Testing Several Algorithms in Several Instances

Friedman Test: Example

	ψ^1	ψ^2	ψ^3	ψ^4
D_1	0.84 (2)	0.79 (3)	0.89 (1)	0.43 (4)
D_2	0.57 (4)	0.78 (2.5)	0.78 (2.5)	0.93 (1)
D_3	0.62 (4)	0.87 (2)	0.88 (1)	0.71 (3)
D_4	0.95 (1)	0.55 (3)	0.49 (4)	0.72 (2)
D_5	0.84 (3)	0.67 (4)	0.89 (1.5)	0.89 (1.5)
D_6	0.51 (4)	0.63 (2)	0.98 (1)	0.55 (3)
avr. rank	3	2.75	1.83	2.41

Testing Several Algorithms in Several Instances

Friedman Test: Example

	ψ^1	ψ^2	ψ^3	ψ^4
D_1	0.84 (2)	0.79 (3)	0.89 (1)	0.43 (4)
D_2	0.57 (4)	0.78 (2.5)	0.78 (2.5)	0.93 (1)
D_3	0.62 (4)	0.87 (2)	0.88 (1)	0.71 (3)
D_4	0.95 (1)	0.55 (3)	0.49 (4)	0.72 (2)
D_5	0.84 (3)	0.67 (4)	0.89 (1.5)	0.89 (1.5)
D_6	0.51 (4)	0.63 (2)	0.98 (1)	0.55 (3)
avr. rank	3	2.75	1.83	2.41

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

Testing Several Algorithms in Several Instances

Friedman Test: Example

	ψ^1	ψ^2	ψ^3	ψ^4
D_1	0.84 (2)	0.79 (3)	0.89 (1)	0.43 (4)
D_2	0.57 (4)	0.78 (2.5)	0.78 (2.5)	0.93 (1)
D_3	0.62 (4)	0.87 (2)	0.88 (1)	0.71 (3)
D_4	0.95 (1)	0.55 (3)	0.49 (4)	0.72 (2)
D_5	0.84 (3)	0.67 (4)	0.89 (1.5)	0.89 (1.5)
D_6	0.51 (4)	0.63 (2)	0.98 (1)	0.55 (3)
avr. rank	3	2.75	1.83	2.41

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] = 2{,}5902$$

Testing Several Algorithms in Several Instances

Iman & Davenport, 1980

- An improvement of Friedman test:

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

follows a F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom

Testing Several Algorithms in Several Instances

Post-hoc Tests

- Decision on the null hypothesis
- In case of rejection use of post-hoc tests to:
  1. Compare all pairs
  2. Compare all classifiers with a control