Outline

Motivation

- **Premise**: comparing C different classifiers $\Rightarrow f_1, \ldots, f_C$
  - $\rightarrow$ Generalization of the pairwise comparison

Motivation

- **Premise**: comparing C different classifiers $\Rightarrow f_1, \ldots, f_C$
  - $\rightarrow$ Generalization of the pairwise comparison
- Pairwise tests are not directly applicable in this case
  - $\rightarrow$ Scenario may be *adapted*

Motivation

- **Premise**: comparing C different classifiers $\Rightarrow f_1, \ldots, f_C$
  - $\rightarrow$ Generalization of the pairwise comparison
- Pairwise tests are not directly applicable in this case
  - $\rightarrow$ Scenario may be *adapted*
- Possible approaches:
  1. One-VS-one comparison
  2. Specific multiple comparison tests

One-VS-one comparison

One-VS-one comparison

- Subdividing the multiple comparison into $\binom{C}{2}$ binary problems

One-VS-one comparison

- Subdividing the multiple comparison into $\binom{C}{2}$ binary problems
- Allows using pairwise comparison methods

One-VS-one comparison

- Subdividing the multiple comparison into $\binom{C}{2}$ binary problems
- Allows using pairwise comparison methods
- Difficult to state a global optimum for the task

One-VS-one comparison

- Subdividing the multiple comparison into $\binom{C}{2}$ binary problems
- Allows using pairwise comparison methods
- Difficult to state a global optimum for the task

- Consider the following scenario:
  - → Four classifiers: **A**, **B**, **C**, **D**
  - → Pairwise comparison: Wilcoxon signed-rank test

One-VS-one comparison

- Subdividing the multiple comparison into $\binom{C}{2}$ binary problems
- Allows using pairwise comparison methods
- Difficult to state a global optimum for the task

- Consider the following scenario:
  - → Four classifiers: **A**, **B**, **C**, **D**
  - → Pairwise comparison: Wilcoxon signed-rank test

Classifier	**Classifier**			
	A	**B**	**C**	**D**
A	–	=	>	>
B	=	–	=	<
C	<	=	–	=
D	<	>	=	–

Multiple comparison tests

- Compare all populations at the same time
  - $\rightarrow$ Easy to state global optima

Multiple comparison tests

- Compare all populations at the same time
  - $\rightarrow$ Easy to state global optima
- Typically, a **two-stage** analysis:
  1. Initial process to state *whether populations differ* among them
  2. Post-hoc analysis to state *which populations differ*

Multiple comparison tests

- Compare all populations at the same time
  - $\rightarrow$ Easy to state global optima

- Typically, a **two-stage** analysis:
  1. Initial process to state *whether populations differ* among them
  2. Post-hoc analysis to state *which populations differ*

- Consider the following conditions:
  - Set of C classifiers: $f_1, \ldots, f_C$
  - Collection M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
    - $\rightarrow$ Matrix of $M \times C$ values

The parametric case: ANOVA

- Acronym for *Analisys of Variance*

The parametric case: ANOVA

- Acronym for *Analisys of Variance*
- Analyzes whether three or more models (significantly) differ in their mean performance
  - → Null hypothesis (H_0): All population **means are equal**
  - → Relies on the F-test

The parametric case: ANOVA

- Acronym for *Analisys of Variance*
- Analyzes whether three or more models (significantly) differ in their mean performance
  - → Null hypothesis (H_0): All population **means are equal**
  - → Relies on the F-test
- Assumptions on the measurements to be compared:
  - → Follow a **normal distribution**
  - → Are **independent** among them

The non-parametric case: Friedman test

- Non-parametric alternative to the ANOVA test
  - $\rightarrow$ No normality assumption

The non-parametric case: Friedman test

- Non-parametric alternative to the ANOVA test
  - $\rightarrow$ No normality assumption

- Relies on ranking procedures
  - $\rightarrow$ Requires paired measurements

The non-parametric case: Friedman test

- Non-parametric alternative to the ANOVA test
  - $\rightarrow$ No normality assumption
- Relies on ranking procedures
  - $\rightarrow$ Requires paired measurements
- States whether there exist differences among the measurements
  - $\rightarrow$ Post-hoc analysis to state which are the different measurements

Friedman test - Procedure

Friedman test - Procedure

1. Rank models for each assortment $1 \leq i \leq M$:
   - Sort $f_1, \ldots, f_C \Rightarrow$ Best (pos. #1) to worst (pos. $\#C$)

Friedman test - Procedure

1. Rank models for each assortment $1 \leq i \leq M$:
   - Sort $f_1, \ldots, f_C \Rightarrow$ Best (pos. #1) to worst (pos. #C)
2. Average rank per model:

$$\bar{R}_j = \frac{1}{M} \sum_{i=1}^{M} R_{ij} \quad \text{with} \quad 1 \leq j \leq C$$

Friedman test - Procedure

1. Rank models for each assortment $1 \leq i \leq M$:
   - Sort $f_1, \ldots, f_C \Rightarrow$ Best (pos. #1) to worst (pos. $\#C$)
2. Average rank per model:

$$\bar{R}_j = \frac{1}{M} \sum_{i=1}^{M} R_{ij} \quad \text{with} \quad 1 \leq j \leq C$$

3. Friedman statistic:

$$\chi_F^2 = \frac{12 \cdot M}{C \cdot (C+1)} \left[\sum_{j=1}^{C} \bar{R}_j^2 \right] - 3 \cdot M \cdot (C+1)$$

Friedman test - Procedure

1. Rank models for each assortment $1 \leq i \leq M$:
   - Sort $f_1, \ldots, f_C \Rightarrow$ Best (pos. $\#1$) to worst (pos. $\#C$)
2. Average rank per model:

$$\bar{R}_j = \frac{1}{M} \sum_{i=1}^{M} R_{ij} \quad \text{with} \quad 1 \leq j \leq C$$

3. Friedman statistic:

$$\chi_F^2 = \frac{12 \cdot M}{C \cdot (C+1)} \left[\sum_{j=1}^{C} \bar{R}_j^2 \right] - 3 \cdot M \cdot (C+1)$$

4. Obtain chi-square critical value: $\chi^2_{\alpha, C-1}$ ($\alpha \rightarrow$ Significance threshold)

Friedman test - Procedure

1. Rank models for each assortment $1 \leq i \leq M$:
   - Sort $f_1, \ldots, f_C \Rightarrow$ Best (pos. #1) to worst (pos. $\#C$)
2. Average rank per model:

$$\bar{R}_j = \frac{1}{M} \sum_{i=1}^{M} R_{ij} \quad \text{with} \quad 1 \leq j \leq C$$

3. Friedman statistic:

$$\chi_F^2 = \frac{12 \cdot M}{C \cdot (C+1)} \left[\sum_{j=1}^{C} \bar{R}_j^2 \right] - 3 \cdot M \cdot (C+1)$$

4. Obtain chi-square critical value: $\chi^2_{\alpha, C-1}$ ($\alpha \rightarrow$ Significance threshold)
5. Reject H_0 if $\chi_F^2 > \chi^2_{\alpha, C-1}$

Procedure - Chi-square critical value table

$C-1$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	42.980
25	34.382	37.652	44.314

Example

Dataset	Classifiers 1	2	3	4
$\mathcal{D}_1$	70	73	78	82
$\mathcal{D}_2$	68	76	75	80
$\mathcal{D}_3$	72	74	79	85
$\mathcal{D}_4$	69	72	78	81
$\mathcal{D}_5$	71	74	77	82
$\mathcal{D}_6$	67	70	73	79

Are there any **statistical differences** among the classifiers considering a **significance threshold** of $\alpha = 0.05$?

Example (solution)

1. Rank models for each assortment:

Dataset	Classifiers 1	2	3	4
$\mathcal{D}_1$	70 (4)	73 (3)	78 (2)	82 (1)
$\mathcal{D}_2$	68 (4)	76 (2)	75 (3)	80 (1)
$\mathcal{D}_3$	72 (4)	74 (3)	79 (2)	85 (1)
$\mathcal{D}_4$	69 (4)	72 (3)	78 (2)	81 (1)
$\mathcal{D}_5$	71 (4)	74 (3)	77 (2)	82 (1)
$\mathcal{D}_6$	67 (4)	70 (3)	73 (2)	79 (1)

Example (solution)

1. Rank models for each assortment:

Dataset	Classifiers 1	2	3	4
$\mathcal{D}_1$	70 (4)	73 (3)	78 (2)	82 (1)
$\mathcal{D}_2$	68 (4)	76 (2)	75 (3)	80 (1)
$\mathcal{D}_3$	72 (4)	74 (3)	79 (2)	85 (1)
$\mathcal{D}_4$	69 (4)	72 (3)	78 (2)	81 (1)
$\mathcal{D}_5$	71 (4)	74 (3)	77 (2)	82 (1)
$\mathcal{D}_6$	67 (4)	70 (3)	73 (2)	79 (1)

2. Average rank per model:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$

Example (solution)

3. Friedman statistic:

$$\chi_F^2 = \frac{12 \cdot 6}{4 \cdot (4+1)} \left[4^2 + 2.83^2 + 2.17^2 + 1^2\right] - 3 \cdot 6 \cdot (4+1) = 17$$

Example (solution)

3. Friedman statistic:

$$\chi^2_F = \frac{12 \cdot 6}{4 \cdot (4+1)} \left[4^2 + 2.83^2 + 2.17^2 + 1^2\right] - 3 \cdot 6 \cdot (4+1) = 17$$

4. Chi-square critical value $\Rightarrow \chi^2_{\alpha,C-1} = \chi^2_{0.05,4-1} = 7.815$

Example (solution)

3. Friedman statistic:

$$\chi_F^2 = \frac{12 \cdot 6}{4 \cdot (4+1)} \left[4^2 + 2.83^2 + 2.17^2 + 1^2\right] - 3 \cdot 6 \cdot (4+1) = 17$$

4. Chi-square critical value $\Rightarrow \chi^2_{\alpha,C-1} = \chi^2_{0.05,4-1} = 7.815$

5. Check possible H_0 rejection $\rightarrow \chi_F^2 > \chi^2_{\alpha,C-1}$:
   $\rightarrow$ $17 > 7.815 \rightarrow H_0$ rejected!

Post-hoc analysis

- Required to clarify the measurement/s that significantly differ
  - $\rightarrow$ Previous analysis proved a statistical difference among them

Post-hoc analysis

- Required to clarify the measurement/s that significantly differ
  - → Previous analysis proved a statistical difference among them

- Two methods that rely on the principle of **Critical Difference**:
  1. Nemenyi test:
     - → Compares all measurements among them
     - → Which specific **pairs of models** differ
  2. Bonferroni-Dunn test:
     - → Compares all measurements against a reference
     - → Comparison against a single **control model**

Nemenyi test - Procedure

1. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$):
   - Compute the mean rank across assorments for each classifier
   - Same as Friedman test

Nemenyi test - Procedure

1. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$):
   - Compute the mean rank across assorments for each classifier
   - Same as Friedman test
2. Compute the Critical Difference:

$$CD = q_\alpha(C) \cdot \sqrt{\frac{C \cdot (C+1)}{6 \cdot M}}$$

$\rightarrow$ q_α: Studentized Range critical values

Nemenyi test - Procedure

1. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$):
   - Compute the mean rank across assorments for each classifier
   - Same as Friedman test
2. Compute the Critical Difference:

$$CD = q_\alpha(C) \cdot \sqrt{\frac{C \cdot (C+1)}{6 \cdot M}}$$

$\rightarrow$ q_α: Studentized Range critical values

3. Pairwise comparison of the models ($1 \leq i, j \leq C$ with $i \neq j$):
   - Hypotheses posed:

     H_0: $f_i = f_j$

     H_1: $f_i \neq f_j$
   - Reject condition: $\left|\bar{R}_i - \bar{R}_j\right| > CD$

Nemenyi test - q_α

C	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
2	1.960	2.241	2.807
3	2.052	2.343	2.949
4	2.108	2.403	3.020
5	2.146	2.444	3.069
6	2.174	2.475	3.105
7	2.195	2.499	3.133
8	2.211	2.518	3.157
9	2.224	2.534	3.176
10	2.235	2.548	3.192
11	2.244	2.559	3.206
12	2.252	2.569	3.218
13	2.259	2.577	3.228
14	2.265	2.584	3.237
15	2.270	2.590	3.245
16	2.275	2.596	3.252
17	2.279	2.601	3.258
18	2.283	2.605	3.264
19	2.286	2.609	3.269
20	2.289	2.613	3.274

Example

Dataset	Classifiers			
	1	2	3	4
$\mathcal{D}_1$	70	73	78	82
$\mathcal{D}_2$	68	76	75	80
$\mathcal{D}_3$	72	74	79	85
$\mathcal{D}_4$	69	72	78	81
$\mathcal{D}_5$	71	74	77	82
$\mathcal{D}_6$	67	70	73	79

Example (solution)

1. Average rank per model:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$

Example (solution)

1. Average rank per model:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$
2. Compute de Critical Difference:

$$CD = q_{\alpha=0.05}(C = 4) \cdot \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 6}} = 2.403 \cdot 0.745 = 1.79$$

Example (solution)

1. Average rank per model:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$
2. Compute de Critical Difference:

$$CD = q_{\alpha=0.05}(C = 4) \cdot \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 6}} = 2.403 \cdot 0.745 = 1.79$$

3. Pairwise comparison:

Classifier	**Classifiers** 1	2	3	4
1	–	1.17 (✗)	1.83 (✓)	3.00 (✓)
2	1.17 (✗)	–	0.66 (✗)	1.83 (✓)
3	1.83 (✓)	0.66 (✗)	–	1.17 (✗)
4	3.00 (✓)	1.83 (✓)	1.17 (✗)	–

Bonferroni-Dunn - Procedure

1. Select the reference case $\Rightarrow f_{\text{ref}}$

Bonferroni-Dunn - Procedure

1. Select the reference case $\Rightarrow f_{\text{ref}}$
2. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$)

Bonferroni-Dunn - Procedure

1. Select the reference case $\Rightarrow f_{\text{ref}}$
2. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$)
3. Compute the Critical Difference:

$$CD = q_{\alpha/C-1} \cdot \sqrt{\frac{C \cdot (C+1)}{6 \cdot M}}$$

Bonferroni-Dunn - Procedure

1. Select the reference case $\Rightarrow f_{\text{ref}}$
2. Obtain the average ranks ($\bar{R}_i$ with $1 \leq i \leq C$)
3. Compute the Critical Difference:

$$CD = q_{\alpha/C-1} \cdot \sqrt{\frac{C \cdot (C+1)}{6 \cdot M}}$$

4. Compare with the reference case:
   - Hypotheses posed:

     H_0: $f_i = f_{\text{ref}}$
     H_1: $f_i \neq f_{\text{ref}}$
   - Reject condition: $\left|\bar{R}_i - \bar{R}_{\text{ref}}\right| > CD$

Bonferroni-Dunn test - $q_{\alpha/C-1}$

C	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	1.645	1.960	2.576
4	1.282	1.645	2.326
5	1.163	1.533	2.241
6	1.095	1.476	2.192
7	1.054	1.440	2.160
8	1.027	1.414	2.136
9	1.006	1.395	2.120
10	0.990	1.380	2.107
11	0.977	1.368	2.096
12	0.966	1.357	2.088
13	0.957	1.349	2.081
14	0.949	1.341	2.075
15	0.943	1.335	2.070
16	0.937	1.329	2.066
17	0.932	1.324	2.062
18	0.928	1.320	2.058
19	0.924	1.316	2.055
20	0.921	1.312	2.053

Example

Dataset	Classifiers 1	2	3	4
$\mathcal{D}_1$	70	73	78	82
$\mathcal{D}_2$	68	76	75	80
$\mathcal{D}_3$	72	74	79	85
$\mathcal{D}_4$	69	72	78	81
$\mathcal{D}_5$	71	74	77	82
$\mathcal{D}_6$	67	70	73	79

Which is the result of the **Bonferroni-Dunn** test with $\alpha = 0.05$ considering as reference **Classifier 4**?

Example (solution)

1. Reference case $\rightarrow f_4$

Example (solution)

1. Reference case $\rightarrow f_4$
2. Obtain the average ranks:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$

Example (solution)

1. Reference case $\rightarrow f_4$
2. Obtain the average ranks:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$
3. Compute the Critical Difference:

$$CD = q_{0.05/4-1} \cdot \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 6}} = 1.960 \cdot 0.745 = 1.46$$

Example (solution)

1. Reference case $\rightarrow f_4$
2. Obtain the average ranks:
   - $\bar{R}_1 = \frac{4+4+4+4+4+4}{6} = 4$
   - $\bar{R}_2 = \frac{3+2+3+3+3+3}{6} = 2.83$
   - $\bar{R}_3 = \frac{2+3+2+2+2+2}{6} = 2.17$
   - $\bar{R}_4 = \frac{1+1+1+1+1+1}{6} = 1$
3. Compute the Critical Difference:

$$CD = q_{0.05/4-1} \cdot \sqrt{\frac{4 \cdot (4+1)}{6 \cdot 6}} = 1.960 \cdot 0.745 = 1.46$$

4. Compare with f_4:

f_1) $\left|\bar{R}_1 - \bar{R}_4\right| > CD \Rightarrow |4-1| > 1.46 \Rightarrow 3 > 1.46$ ✓

f_2) $\left|\bar{R}_2 - \bar{R}_4\right| > CD \Rightarrow |2.83-1| > 1.46 \Rightarrow 1.83 > 1.46$ ✓

f_3) $\left|\bar{R}_3 - \bar{R}_4\right| > CD \Rightarrow |2.17-1| > 1.46 \Rightarrow 1.17 \not> 1.46$ ✗

T7: Statistical model comparison

Fundamentos del Aprendizaje Automático

Curso 2025/2026