T7: Statistical model comparison

Fundamentos del Aprendizaje Automático

Curso 2025/2026

Structure

Outline

Open question

Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

Open question

Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

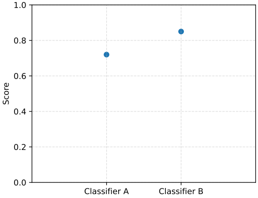
Classifier A	**Classifier B**

Open question

Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

1. Single partitioning: train $\mathcal{T}$ and test $\mathcal{S}$

Classifier A	**Classifier B**
Case 1) $s_A \leftarrow \mathcal{S} \Leftrightarrow f_A(\mathcal{S}; \mathcal{T})$	Case 1) $s_B \leftarrow \mathcal{S} \Leftrightarrow f_B(\mathcal{S}; \mathcal{T})$



Open question

Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

1. Single partitioning: train $\mathcal{T}$ and test $\mathcal{S}$
2. Cross-validation partitioning: train $[\mathcal{T}_1, \ldots, \mathcal{T}_k]$ and test $[\mathcal{S}_1, \ldots, \mathcal{S}_k]$

Classifier A

Case 1) $s_A \leftarrow \mathcal{S} \Leftrightarrow f_A(\mathcal{S}; \mathcal{T})$

Case 2) $s_{A1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_A(\mathcal{S}_1; \mathcal{T}_1)$

...

$s_{Ak} \leftarrow \mathcal{S}_k \Leftrightarrow f_A(\mathcal{S}_k; \mathcal{T}_k)$

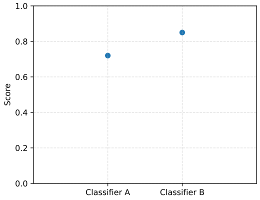
Classifier B

Case 1) $s_B \leftarrow \mathcal{S} \Leftrightarrow f_B(\mathcal{S}; \mathcal{T})$

Case 2) $s_{B1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_B(\mathcal{S}_1; \mathcal{T}_1)$

...

$s_{Bk} \leftarrow \mathcal{S}_k \Leftrightarrow f_B(\mathcal{S}_k; \mathcal{T}_k)$



Open question

Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

1. Single partitioning: train $\mathcal{T}$ and test $\mathcal{S}$
2. Cross-validation partitioning: train $[\mathcal{T}_1, \ldots, \mathcal{T}_k]$ and test $[\mathcal{S}_1, \ldots, \mathcal{S}_k]$

Classifier A

Case 1) $s_A \leftarrow \mathcal{S} \Leftrightarrow f_A(\mathcal{S}; \mathcal{T})$

Case 2) $s_{A1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_A(\mathcal{S}_1; \mathcal{T}_1)$

$$\ldots$$

$$s_{Ak} \leftarrow \mathcal{S}_k \Leftrightarrow f_A(\mathcal{S}_k; \mathcal{T}_k)$$

$$\bar{s}_A \pm \sigma_A$$

Classifier B

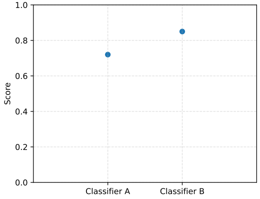
Case 1) $s_B \leftarrow \mathcal{S} \Leftrightarrow f_B(\mathcal{S}; \mathcal{T})$

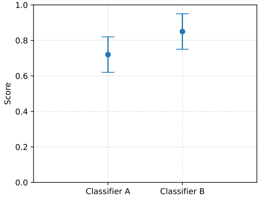
Case 2) $s_{B1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_B(\mathcal{S}_1; \mathcal{T}_1)$

$$\ldots$$

$$s_{Bk} \leftarrow \mathcal{S}_k \Leftrightarrow f_B(\mathcal{S}_k; \mathcal{T}_k)$$

$$\bar{s}_B \pm \sigma_B$$



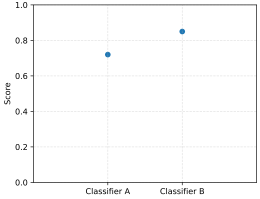


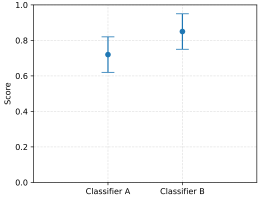
Open question

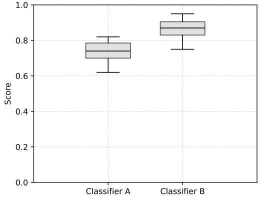
Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

1. Single partitioning: train $\mathcal{T}$ and test $\mathcal{S}$
2. Cross-validation partitioning: train $[\mathcal{T}_1, \ldots, \mathcal{T}_k]$ and test $[\mathcal{S}_1, \ldots, \mathcal{S}_k]$

Classifier A	Classifier B
Case 1) $s_A \leftarrow \mathcal{S} \Leftrightarrow f_A(\mathcal{S};\mathcal{T})$	Case 1) $s_B \leftarrow \mathcal{S} \Leftrightarrow f_B(\mathcal{S};\mathcal{T})$
Case 2) $s_{A1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_A(\mathcal{S}_1;\mathcal{T}_1)$	Case 2) $s_{B1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_B(\mathcal{S}_1;\mathcal{T}_1)$
...	...
$s_{Ak} \leftarrow \mathcal{S}_k \Leftrightarrow f_A(\mathcal{S}_k;\mathcal{T}_k)$	$s_{Bk} \leftarrow \mathcal{S}_k \Leftrightarrow f_B(\mathcal{S}_k;\mathcal{T}_k)$
$\bar{s}_A \pm \sigma_A$	$\bar{s}_B \pm \sigma_B$





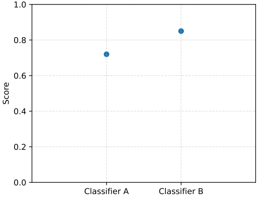


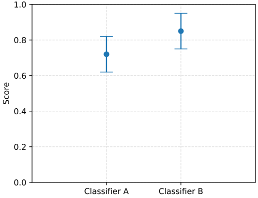
Open question

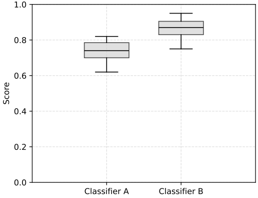
Task: Given set $\mathcal{D} = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{|\mathcal{D}|}$, which is the best classification option?

1. Single partitioning: train $\mathcal{T}$ and test $\mathcal{S}$
2. Cross-validation partitioning: train $[\mathcal{T}_1, \ldots, \mathcal{T}_k]$ and test $[\mathcal{S}_1, \ldots, \mathcal{S}_k]$
3. Several datasets: $\mathcal{D}_1, \ldots, \mathcal{D}_M$

Classifier A	Classifier B
Case 1) $s_A \leftarrow \mathcal{S} \Leftrightarrow f_A(\mathcal{S}; \mathcal{T})$	Case 1) $s_B \leftarrow \mathcal{S} \Leftrightarrow f_B(\mathcal{S}; \mathcal{T})$
Case 2) $s_{A1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_A(\mathcal{S}_1; \mathcal{T}_1)$	Case 2) $s_{B1} \leftarrow \mathcal{S}_1 \Leftrightarrow f_B(\mathcal{S}_1; \mathcal{T}_1)$
...	...
$s_{Ak} \leftarrow \mathcal{S}_k \Leftrightarrow f_A(\mathcal{S}_k; \mathcal{T}_k)$	$s_{Bk} \leftarrow \mathcal{S}_k \Leftrightarrow f_B(\mathcal{S}_k; \mathcal{T}_k)$
$\bar{s}_A \pm \sigma_A$	$\bar{s}_B \pm \sigma_B$







Open question

Model performance is influenced by several sources of randomness:

Open question

Model performance is influenced by several sources of randomness:

1. Model stochasticity
   - **Initialization** of the **weights**

Open question

Model performance is influenced by several sources of randomness:

1. Model stochasticity
   - **Initialization** of the **weights**
2. Data sampling
   - **Different** train/test **splits** often yields different **performace scores**

Open question

Model performance is influenced by several sources of randomness:

1. Model stochasticity
   - **Initialization** of the **weights**
2. Data sampling
   - **Different** train/test **splits** often yields different **performace scores**
3. Evaluation procedures
   - **Cross-validation** produces **correlated** estimates

Open question

Model performance is influenced by several sources of randomness:

1. Model stochasticity
   - **Initialization** of the **weights**
2. Data sampling
   - **Different** train/test **splits** often yields different **performace scores**
3. Evaluation procedures
   - **Cross-validation** produces **correlated** estimates

Performance difference is not necessarily a real difference

Statistical model comparison

Statistical methods for model comparison overcome this issue:

Statistical model comparison

Statistical methods for model comparison overcome this issue:

1. Determine whether observed differences are statistically significant
   - Observed difference is **random** or **real**
   - Essential for **rigorous** and **reproducible** evaluation in machine learning

Statistical model comparison

Statistical methods for model comparison overcome this issue:

1. Determine whether observed differences are statistically significant
   - Observed difference is **random** or **real**
   - Essential for **rigorous** and **reproducible** evaluation in machine learning
2. Quantify uncertainty
   - Typically in terms of *p-values*, *confidence intervals*, *test statistics*

Statistical model comparison

Statistical methods for model comparison overcome this issue:

1. Determine whether observed differences are statistically significant
   - Observed difference is **random** or **real**
   - Essential for **rigorous** and **reproducible** evaluation in machine learning
2. Quantify uncertainty
   - Typically in terms of *p-values*, *confidence intervals*, *test statistics*
3. Avoid erroneous claims based on noisy estimates

Statistical model comparison

Statistical methods for model comparison overcome this issue:

1. Determine whether observed differences are statistically significant
   - Observed difference is **random** or **real**
   - Essential for **rigorous** and **reproducible** evaluation in machine learning
2. Quantify uncertainty
   - Typically in terms of *p-values*, *confidence intervals*, *test statistics*
3. Avoid erroneous claims based on noisy estimates

Comparison	**Parametric**	**Non-parametric**
Pairwise	Paired t-test	Wilcoxon signed-rank
Multiple	ANOVA	Friedman + post-hoc

Key concepts

Key concepts

Always two propositions to explain the posed scenario:

Key concepts

Always two propositions to explain the posed scenario:

- Null hypothesis (H_0): **No relation** between sets
- Alternative hypothesis (H_1): There **exists a difference** between sets

Key concepts

Always two propositions to explain the posed scenario:

- Null hypothesis (H_0): **No relation** between sets
- Alternative hypothesis (H_1): There **exists a difference** between sets

The tests compute a statistical significance or p-value:

- Probability of obtaining results as **extreme as or more extreme** than the ones observed if H_0 were true
- The concept of extremeness depends on the test
- Smaller p-value $\Rightarrow$ stronger evidence against H_0

Key concepts

Always two propositions to explain the posed scenario:

- Null hypothesis (H_0): **No relation** between sets
- Alternative hypothesis (H_1): There **exists a difference** between sets

The tests compute a statistical significance or *p*-value:

- Probability of obtaining results as **extreme as or more extreme** than the ones observed if H_0 were true
- The concept of extremeness depends on the test
- Smaller *p*-value $\Rightarrow$ stronger evidence against H_0

The significance level or α acts as threshold to accept/reject H_0:

- Maximum probability of rejecting H_0 when it is actually true
- The lower, the more strict

Procedure

Procedure

1. Select suitable statistical test

Procedure

1. Select suitable statistical test
2. Formulate hypotheses: H_0 and H_1

Procedure

1. Select suitable statistical test
2. Formulate hypotheses: H_0 and H_1
4. Compute the p-value

Procedure

1. Select suitable statistical test

2. Formulate hypotheses: H_0 and H_1

4. Compute the p-value

5. Compare p-value to significance level α:
   - If $p < \alpha \Rightarrow$ **reject** H_0
   - If $p \geq \alpha \Rightarrow$ **do not reject** H_0

Consideration about data collections

Ideally, these **principles** should be followed:

Consideration about data collections

Ideally, these **principles** should be followed:

1. Providing paired results:
    - Evaluation on the **same** data instances, folds, or datasets
    - All models are equally affected by random fluctuations in data
    - A must in some tests

Consideration about data collections

Ideally, these **principles** should be followed:

1. Providing paired results:
   - Evaluation on the **same** data instances, folds, or datasets
   - All models are equally affected by random fluctuations in data
   - A must in some tests

2. Considering independent data assortments $(\mathcal{D}_1, \ldots, \mathcal{D}_M)$:
   - Most statistical tests assume independent datasets
   - Cross-validation strategies violate this independence assumption

     $\rightarrow$ Optimistic estimations
   - Possible solutions:
     a) Specific tests that compensate the bias
     b) Averaging across folds

Outline

Parametric case: the paired t-test

- Assumes data are approximately normally distributed
- Types of tests:

Type	**Use case**
One-sample t-test	Compare a sample mean to a known value
Independent two-sample t-test	Compare means of two independent groups
Paired t-test	Compare means of related or paired data

- We focus on the paired t-test:
  - $\rightarrow$ Statistic t represents the ratio between **difference** and **variability**
  - $\rightarrow$ Larger $|t|$ represents more evidence against H_0

Formulation

Consider the following conditions:

Formulation

Consider the following conditions:

- Consider two classifiers: f_A and f_B
- Consider M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
- We have M performace scores for each classifier:
  - → Classifier f_A: $s_{A1}, \ldots, s_{AM}$
  - → Classifier f_B: $s_{B1}, \ldots, s_{BM}$

Formulation

Consider the following conditions:

- Consider two classifiers: f_A and f_B
- Consider M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
- We have M performace scores for each classifier:
  - $\rightarrow$ Classifier f_A: $s_{A1}, \ldots, s_{AM}$
  - $\rightarrow$ Classifier f_B: $s_{B1}, \ldots, s_{BM}$

Procedure:

Formulation

Consider the following conditions:

- Consider two classifiers: f_A and f_B
- Consider M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
- We have M performace scores for each classifier:
  - → Classifier f_A: $s_{A1}, \ldots, s_{AM}$
  - → Classifier f_B: $s_{B1}, \ldots, s_{BM}$

Procedure:

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$

Formulation

Consider the following conditions:

- Consider two classifiers: f_A and f_B
- Consider M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
- We have M performace scores for each classifier:
  - $\rightarrow$ Classifier f_A: $s_{A1}, \ldots, s_{AM}$
  - $\rightarrow$ Classifier f_B: $s_{B1}, \ldots, s_{BM}$

Procedure:

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$
2. Compute mean ($\bar{d}$) and standard deviation (σ_d) as:

$$\bar{d} = \frac{1}{M} \sum_{i=1}^{M} d_i, \qquad \sigma_d = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} \left(d_i - \bar{d}\right)^2}$$

Formulation

Consider the following conditions:

- Consider two classifiers: f_A and f_B
- Consider M data assortments: $\mathcal{D}_1, \ldots, \mathcal{D}_M$ with $\mathcal{D}_i = \mathcal{T}_i \cup \mathcal{S}_i$
- We have M performace scores for each classifier:
  - $\rightarrow$ Classifier f_A: $s_{A1}, \ldots, s_{AM}$
  - $\rightarrow$ Classifier f_B: $s_{B1}, \ldots, s_{BM}$

Procedure:

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$
2. Compute mean $(\bar{d})$ and standard deviation (σ_d) as:

$$\bar{d} = \frac{1}{M}\sum_{i=1}^{M} d_i, \qquad \sigma_d = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(d_i - \bar{d}\right)^2}$$

3. Compute test statistic:

$$t = \frac{\bar{d}}{\sigma_d / \sqrt{M}}$$

Formulation

4. Obtain the value of the t-student distribution as t_{stu}:
   - Single / both directions:
     - $\rightarrow$ Single-tail when H_0 states $f_A > f_B$ or $f_A < f_B$
     - $\rightarrow$ Two-tail when H_0 states $f_A = f_B$
   - Degrees of freedom: $M - 1$
   - Significance value α

Formulation

4. Obtain the value of the t-student distribution as t_{stu}:
   - Single / both directions:
     - $\rightarrow$ Single-tail when H_0 states $f_A > f_B$ or $f_A < f_B$
     - $\rightarrow$ Two-tail when H_0 states $f_A = f_B$
   - Degrees of freedom: $M - 1$
   - Significance value α

5. Compare $|t|$ with t_{stu}:
   - If $|t| > t_{\text{stu}}$: **reject** H_0
   - If $|t| \leq t_{\text{stu}}$: **accept** H_0

Example

Dataset	Classifier A	Classifier B
$\mathcal{D}_1$	85	87
$\mathcal{D}_2$	70	68
$\mathcal{D}_3$	79	85
$\mathcal{D}_4$	78	75
$\mathcal{D}_5$	83	83

Is **Classifier A** equivalent to **Classifier B** for $\alpha = 0.05$[1]?

[1] $t_{\text{stu}} \approx 2.776$

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	+2
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	+3
$\mathcal{D}_5$	83	83	0

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	$+2$
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	$+3$
$\mathcal{D}_5$	83	83	0

S2) Mean $\bar{d} = -0.6$

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	+2
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	+3
$\mathcal{D}_5$	83	83	0

S2) Mean $\bar{d} = -0.6$

S2) Deviation $\sigma_d = 3.2$

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	$+2$
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	$+3$
$\mathcal{D}_5$	83	83	0

S2) Mean $\bar{d} = -0.6$

S2) Deviation $\sigma_d = 3.2$

S3) $t = \frac{-0.6}{3.2/\sqrt{(5-1)}} = -0.375$

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	+2
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	+3
$\mathcal{D}_5$	83	83	0

S2) Mean $\bar{d} = -0.6$

S2) Deviation $\sigma_d = 3.2$

S3) $t = \frac{-0.6}{3.2/\sqrt{(5-1)}} = -0.375$

S4) $t_{\text{stu}} = 2.776$

Example (solution)

Dataset	Classifier A	Classifier B	Difference
$\mathcal{D}_1$	85	87	-2
$\mathcal{D}_2$	70	68	$+2$
$\mathcal{D}_3$	79	85	-6
$\mathcal{D}_4$	78	75	$+3$
$\mathcal{D}_5$	83	83	0

S2) Mean $\bar{d} = -0.6$

S2) Deviation $\sigma_d = 3.2$

S3) $t = \frac{-0.6}{3.2/\sqrt{(5-1)}} = -0.375$

S4) $t_{\text{stu}} = 2.776$

S5) $\underbrace{|-0.375|}_{|t|} \not> \underbrace{2.776}_{t_{\text{stu}}}? \Rightarrow$ **Accept** H_0

Non-parametric case: the Wilcoxon signed-rank test

- Wilcoxon tests $\Rightarrow$ family of **non-parametric** alternatives to t-tests

Non-parametric case: the Wilcoxon signed-rank test

- Wilcoxon tests ⇒ family of **non-parametric** alternatives to t-tests

Wilcoxon test	**Parametric equivalent**
Signed-rank	Paired t-test
Rank-sum	Independent two-sample t-test

Non-parametric case: the Wilcoxon signed-rank test

- Wilcoxon tests $\Rightarrow$ family of **non-parametric** alternatives to t-tests

Wilcoxon test	**Parametric equivalent**
Signed-rank	Paired t-test
Rank-sum	Independent two-sample t-test

- Focus on Wilcoxon signed-rank test for paired data
  - $\rightarrow$ No normality assumption
  - $\rightarrow$ Robust to outliers

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$
2. Remove zeroes

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$
2. Remove zeroes
3. Rank the absolute differences $|d_i|$ (smallest to largest)
   - Assign sign to the ranked differences

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi}$, $i = 1, \ldots, M$
2. Remove zeroes
3. Rank the absolute differences $|d_i|$ (smallest to largest)
   - Assign sign to the ranked differences
4. Sum positive (W^+) and negative (W^-) ranks
   - Test statistic $W = \min\{W^+, W^-\}$

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi},\ i = 1, \ldots, M$
2. Remove zeroes
3. Rank the absolute differences $|d_i|$ (smallest to largest)
   - Assign sign to the ranked differences
4. Sum positive (W^+) and negative (W^-) ranks
   - Test statistic $W = \min\{W^+, W^-\}$
5. Obtain the **critical value** from the critical values table:
   - Single / both directions:
     - $\rightarrow$ Single-tailed when H_0 states $f_A > f_B$ or $f_A < f_B$
     - $\rightarrow$ Two-tailed when H_0 states $f_A = f_B$
   - Significance value α

Procedure

1. Dataset-wise performance difference: $d_i = s_{Ai} - s_{Bi},\ i = 1, \ldots, M$
2. Remove zeroes
3. Rank the absolute differences $|d_i|$ (smallest to largest)
   - Assign sign to the ranked differences
4. Sum positive (W^+) and negative (W^-) ranks
   - Test statistic $W = \min\{W^+, W^-\}$
5. Obtain the **critical value** from the critical values table:
   - Single / both directions:
     - $\rightarrow$ Single-tailed when H_0 states $f_A > f_B$ or $f_A < f_B$
     - $\rightarrow$ Two-tailed when H_0 states $f_A = f_B$
   - Significance value α
6. Reject H_0 if:
   - Single-tailed test: $W \leq W_\alpha$
   - Two-tailed test: $W \leq W_{\alpha/2}$

Procedure

	Two-Tailed			One-Tailed		
n	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
5	0	0	0	0	0	0
6	2	2	0	2	2	1
7	3	3	1	4	3	2
8	5	4	2	6	5	3
9	8	6	3	8	7	5
10	10	8	5	11	9	6
11	13	10	7	14	11	8
12	17	13	9	18	14	11
13	21	16	11	22	18	13
14	25	19	14	26	21	16
15	30	23	17	31	25	19
16	35	27	21	36	29	22
17	40	31	24	41	33	26
18	46	36	28	47	38	29
19	52	40	32	53	42	33
20	59	45	37	60	47	38
21	66	51	41	67	53	43
22	73	56	46	74	58	48
23	81	62	51	82	64	53
24	89	68	56	90	70	58
25	98	75	62	99	77	64

Example

Dataset	Classifier A	Classifier B
$\mathcal{D}_1$	87	85
$\mathcal{D}_2$	68	70
$\mathcal{D}_3$	85	79
$\mathcal{D}_4$	75	78
$\mathcal{D}_5$	83	83
$\mathcal{D}_6$	90	85

Is **Classifier A** equivalent to **Classifier B** for $\alpha = 0.1$?

Example (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	87	85	$+2$	1.5
$\mathcal{D}_2$	68	70	-2	1.5
$\mathcal{D}_3$	85	79	$+6$	5
$\mathcal{D}_4$	75	78	-3	3
$\mathcal{D}_5$	83	83	0	-
$\mathcal{D}_6$	90	85	$+5$	4

Example (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	87	85	+2	1.5
$\mathcal{D}_2$	68	70	-2	1.5
$\mathcal{D}_3$	85	79	+6	5
$\mathcal{D}_4$	75	78	-3	3
$\mathcal{D}_5$	83	83	0	-
$\mathcal{D}_6$	90	85	+5	4

S4) $W^+ = 1.5 + 4 + 5 = 10.5$

S4) $W^- = 1.5 + 3 = 4.5$

Example (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	87	85	$+2$	1.5
$\mathcal{D}_2$	68	70	-2	1.5
$\mathcal{D}_3$	85	79	$+6$	5
$\mathcal{D}_4$	75	78	-3	3
$\mathcal{D}_5$	83	83	0	-
$\mathcal{D}_6$	90	85	$+5$	4

S4) $W^+ = 1.5 + 4 + 5 = 10.5$

S4) $W^- = 1.5 + 3 = 4.5$

S4) $W = \min(W^+, W^-) = \min(10.5, 4.5) = 4.5$

Example (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	87	85	+2	1.5
$\mathcal{D}_2$	68	70	-2	1.5
$\mathcal{D}_3$	85	79	+6	5
$\mathcal{D}_4$	75	78	-3	3
$\mathcal{D}_5$	83	83	0	-
$\mathcal{D}_6$	90	85	+5	4

S4) $W^+ = 1.5 + 4 + 5 = 10.5$

S4) $W^- = 1.5 + 3 = 4.5$

S4) $W = \min(W^+, W^-) = \min(10.5, 4.5) = 4.5$

S5) $W_{\alpha/2} = 0$

Example (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	87	85	+2	1.5
$\mathcal{D}_2$	68	70	-2	1.5
$\mathcal{D}_3$	85	79	+6	5
$\mathcal{D}_4$	75	78	-3	3
$\mathcal{D}_5$	83	83	0	-
$\mathcal{D}_6$	90	85	+5	4

S4) $W^+ = 1.5 + 4 + 5 = 10.5$

S4) $W^- = 1.5 + 3 = 4.5$

S4) $W = \min(W^+, W^-) = \min(10.5, 4.5) = 4.5$

S5) $W_{\alpha/2} = 0$

S6) $\underbrace{4.5}_{W} \not\leq \underbrace{0}_{W_{\alpha/2}} \Rightarrow$ **Accept** H_0

Exercise

Dataset	Classifier A	Classifier B
$\mathcal{D}_1$	95	80
$\mathcal{D}_2$	88	70
$\mathcal{D}_3$	90	80
$\mathcal{D}_4$	85	60
$\mathcal{D}_5$	92	78
$\mathcal{D}_6$	87	80
$\mathcal{D}_7$	91	85
$\mathcal{D}_8$	95	70
$\mathcal{D}_9$	82	60
$\mathcal{D}_{10}$	96	80
$\mathcal{D}_{11}$	88	85
$\mathcal{D}_{12}$	90	78
$\mathcal{D}_{13}$	85	80
$\mathcal{D}_{14}$	92	78
$\mathcal{D}_{15}$	87	80

Exercise (solution)

Dataset	Classifier A	Classifier B	Difference	Rank
$\mathcal{D}_1$	95	80	$+15$	1
$\mathcal{D}_2$	88	70	$+18$	2
$\mathcal{D}_3$	90	80	$+10$	3
$\mathcal{D}_4$	85	60	$+25$	4.5
$\mathcal{D}_5$	92	78	$+14$	4.5
$\mathcal{D}_6$	87	80	$+7$	6
$\mathcal{D}_7$	91	85	$+6$	7
$\mathcal{D}_8$	95	70	$+25$	8.5
$\mathcal{D}_9$	82	60	$+22$	8.5
$\mathcal{D}_{10}$	96	80	$+16$	10
$\mathcal{D}_{11}$	88	85	$+3$	11
$\mathcal{D}_{12}$	90	78	$+12$	12
$\mathcal{D}_{13}$	85	80	$+5$	13
$\mathcal{D}_{14}$	92	78	$+14$	14.5
$\mathcal{D}_{15}$	87	80	$+7$	14.5

- $W^+ = 120,\ W^- = 0 \Rightarrow \min\left(W^+, W^-\right) = 0$
- $n = 15$
- If $\alpha = 0.1 \Rightarrow W_{\alpha/2} = 23 \Rightarrow W < W_{\alpha/2} \Rightarrow$ **Reject**

Exercise (solution)

One-tailed test:

- Is **A** $>$ **B**?
- $W \leq W_\alpha \Rightarrow 0 \leq 25 \Rightarrow$ **Reject!**

Classifier **A** is significantly better than **B** (with $\alpha = 0.1$)