Machine Learning: Evaluation

Steffen Rendle

Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim

Wintersemester 2007 / 2008

Comparison of Algorithms

Comparison of Algorithms

Is algorithm „A“ better than algorithm „B“?

- depends on task/ dataset
- difference might be due to limited size of dataset
- Same quality measure and dataset(s) should be used to evaluate „A“ and „B“.

Comparison Schemes

Paired Test

- Both algorithms are trained on the same n datasets and use the same n test datasets.
- E.g. 10-fold CV:
  1. train each „A“ and „B“ on fold $f_2 \dots f_n$, evaluate each on f_1. We get quality measures a_1 and b_1.
  2. train each „A“ and „B“ on fold $f_1, f_3 \dots f_n$, evaluate each on f_2. We get quality measures a_2 and b_2.
  3. ...
- We get results $a_1, \dots, a_n$ and $b_1, \dots, b_n$. The results a_i and b_i are paired.
- Reduces variance in the quality estimations.

Comparison Schemes

Unpaired Test

- Given are n results for „A“ and m for method „B“.
- E.g. one researcher has implemented logistic regression and evaluated it on the iris dataset with 10-fold CV. Another researcher has implemented a decision tree and evaluates it on iris with 5-fold CV. They can compare their results with an unpaired test.

Hypothesis testing

HYPOTHESIS TESTING is a statistical method for comparing two test series.

- Hypothesis H_0: „There is no difference."
- This hypothesis is tested with a significance level α. E.g 5%.
- H_0 is also called the „null hypothesis".
- A hypothesis test tries to falsify/ reject the null hypothesis.
- If we can reject the null hypothesis, the results are SIGNIFICANTLY different.

Hypothesis testing

We deal with two testing schemes:

- Wald test
  - for normal distributed data
  - general test
- t-Test
  - for testing means of normal distributed data
  - good for small sample sizes

Paired Wald-Test

- Data:
  - n samples
  - $X_1, \ldots, X_n$ for algorithm „A“
  - $Y_1, \ldots, Y_n$ for algorithm „B“
  - X_i and Y_i are paired!
- Hypothesis: $\delta = 0$

Paired Wald-Test

- $Z_i := X_i - Y_i$ (because X_i and Y_i are paired!)
- $\delta = E(Z) = E(X) - E(Y)$
- δ is estimated by $\hat{\delta} = \overline{X} - \overline{Y}$
- $\hat{\delta}$ is a random variable
- the estimated standard error $\hat{\text{se}}(\hat{\delta})$ of $\hat{\delta}$ is

$$\hat{\text{se}}(\hat{\delta}) = \sqrt{\frac{s_Z^2}{n}}$$

$$\text{with} \quad s_Z^2 := \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z})^2$$

Paired Wald-Test

- The normalized random variable W describing the error is:

$$W := \frac{\hat{\delta}}{\hat{\text{se}}(\hat{\delta})} = \frac{\overline{Z}}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}}$$

- As $W \rightsquigarrow N(0,1)$ we can conclude that the difference is with probability of at least α in:

$$C_n = \overline{Z} \pm z_{\alpha/2}\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}$$

- We can test our hypothesis $\delta = 0$:
  - Method 1: If $0 \notin C_n$ than reject H_0. I.e. there should be a difference between the two algorithms.
  - Method 2: If $|W| > z_{\alpha/2}$ then reject H_0.

Unpaired Wald-Test

- Data:
  - n samples of „A“, m samples of „B“
  - $X_1, \ldots, X_n$ for algorithm „A“
  - $Y_1, \ldots, Y_m$ for algorithm „B“
  - X_i and Y_i are independent!
- Hypothesis: $\delta = 0$

Unpaired Wald-Test

- $\delta = E(X) - E(Y)$
- δ is estimated by $\hat{\delta} = \overline{X} - \overline{Y}$
- $\hat{\delta}$ is a random variable
- the estimated standard error $\hat{\text{se}}(\hat{\delta})$ of $\hat{\delta}$ is

$$\hat{\text{se}}(\hat{\delta}) = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

$$\text{with} \quad s_U^2 := \frac{1}{k-1} \sum_{i=1}^{k} (U_i - \overline{U})^2$$

Unpaired Wald-Test

- The normalized random variable W describing the error is:

$$W := \frac{\hat{\delta}}{\hat{\text{se}}(\hat{\delta})} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

- As $W \rightsquigarrow N(0,1)$ we can conclude that the difference is with probability of at least α in:

$$C_n = \overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

- We can test our hypothesis $\delta = 0$:
  - Method 1: If $0 \notin C_n$ than reject H_0. I.e. there should be a difference between the two algorithms.
  - Method 2: If $|W| > z_{\alpha/2}$ then reject H_0.

t-Tests

- The Wald-test is based on $W \rightsquigarrow N(0,1)$.
- For small sample sizes, the approximation with a normal is inaccurate.
- In fact for small sample sizes W is t-distributed: $W \sim t_{n-1}$

$$f(x) = \alpha_n \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\text{with} \quad \alpha_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

t-Tests

- t-Tests can be performed by using $t_{n,\alpha}$ instead of z_α. Where n are the „degrees of freedom“.
- Paired t-Test:
  - degrees of freedom: $k = n - 1$
  - Method: If $|W| > t_{k,\alpha/2}$ then reject H_0.
- Unpaired t-Test:
  - degrees of freedom: $k = \min\{n, m\} - 1$
  - Method: If $|W| > t_{k,\alpha/2}$ then reject H_0.

Example

Paired t-Test with level $\alpha = 0.05$ for the data:

	fold 1	fold 2	fold 3	fold 4	fold 5
Method A	88	89	92	90	90
Method B	92	90	91	89	91

. . . *see blackboard* . . .