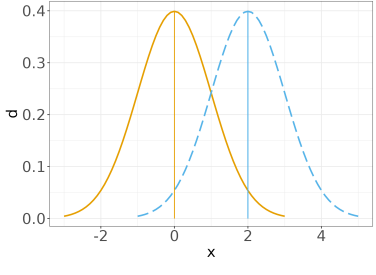
Applied Machine Learning

Performance Evaluation:
Model Selection & Hypothesis Testing



Learning goals

- What questions can we answer with benchmarks?
- What value do hypothesis tests add?
- Benchmarking scenarios and relevant tests

MODEL EVALUATION, MODEL SELECTION AND ALGORITHM SELECTION

- Model evaluation: estimate how well one model performs
- Model selection: select between two or more models
- Algorithm selection: select between two or more algorithms (inducers)

QUESTIONS TO ASK IN MODEL SELECTION TASKS

There are several considerations when selecting a model:

1. Do we only care about performance or do we have additional model preferences, e.g., complex vs. non-complex models, slow vs. fast computation?
2. Do we only care about performance on the given test data or whether differences in performance hold for other test data as well?
3. How large is the difference in model performance?
4. How shall benchmark studies and our own preferences guide our decision in model selection? For instance, if we care about limiting model complexity, and a non-complex model can achieve 90% of the predictive performance of a complex one, which model shall be selected?

$\Rightarrow$ Hypothesis tests are an additional tool in benchmarking to assist us in model selection tasks but only represent one aspect in our decision making process; important other considerations include our preferences regarding the

NEED FOR HYPOTHESIS TESTS IN BENCHMARKING

Hypothesis tests can (a) quantify evidence that a difference in model performance can be generalized to other data; and (b) assist us in weighing the trade-off in our model preferences. Consider the following two scenarios:

- Assume we are interested in both limiting model complexity and predictive performance. A non-complex linear model achieves 90% of the predictive performance of a complex random forest.
- **Scenario A**: A hypothesis test indicates that the difference in performance is **not significant**.
- **Scenario B**: A hypothesis test indicates that the difference in performance is **significant**.
- Given our preferences, we decide that the given difference in performance needs to be significant to accept an increase in model complexity. In scenario A, we choose the linear model; in scenario B we choose the random forest.

BUT FIRST: HOW TO MEASURE ML PERFORMANCE?

We can choose between:

- Holdout
- Repeated Holdout
- Cross-Validation
- Repeated Cross-Validation
- Bootstrap

But which ones is best?

BUT FIRST: HOW TO MEASURE ML PERFORMANCE?

We can choose between:

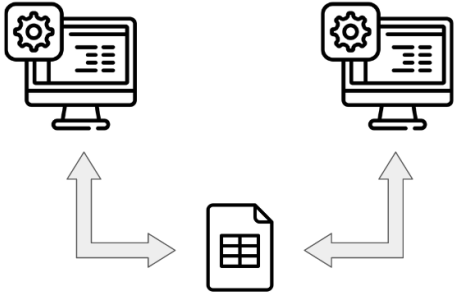
- Holdout
- Repeated Holdout
- Cross-Validation
- Repeated Cross-Validation
- Bootstrap

But which ones is best?
Generally: the more resources invested, the better we can select a model.

Scenario 1: Two Trained Models on 1 Data Set

FIRST BENCHMARKING SCENARIO

Two trained models on 1 data set



TWO-MATCHED-SAMPLES T-TEST

- Given a holdout test set $\mathcal{D}_{\text{test}}$, can we assume a difference in performance between $\hat{f}_1$ and $\hat{f}_2$ generalizes to other data?

  $H_0 : \text{GE}(\hat{f}_1, L, \mathcal{D}_{\text{test}}) = \text{GE}(\hat{f}_2, L, \mathcal{D}_{\text{test}})$ vs. $H_1 : \text{GE}(\hat{f}_1, L, \mathcal{D}_{\text{test}}) \neq \text{GE}(\hat{f}_2, L, \mathcal{D}_{\text{test}})$

- Test statistic:

$$T_{\text{t-test}} = \sqrt{n_{test}} \frac{\bar{d}}{\sigma_d} \sim t(n_{test} - 1), \text{ with } \sigma_d = \sqrt{\frac{1}{n_{\text{test}} - 1} \sum_{i=1}^{n_{test}} \left(d_i - \bar{d}\right)^2}$$

$$d_i = L(y^{(i)}, \hat{f}_1(\mathbf{x}^{(i)})) - L(y^{(i)}, \hat{f}_2(\mathbf{x}^{(i)})) \qquad \bar{d} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} d_i \qquad (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}_{\text{test}}$$

- The two-matched samples t-test (or paired t-test) has multiple requirements:
  1. **Loss values need to be iid samples**
  2. **(Pseudo-) normality of loss values:** Requires a minimum of ≈ 30 loss values or loss values following a normal distribution.
  3. **Equal variances:** Assumes equal variances of both populations.

TWO-MATCHED-SAMPLES T-TEST

- **Example:** We evaluate learners `classif.rpart` and `classif.ranger` on the binary classification task `kin8nm`.
- We use the Brier score as loss function:

id	truth	rpart_prob	ranger_prob	rpart_loss	ranger_loss	diff_loss
1	P	0.7823	0.7759	0.0474	0.0502	-0.0028
2	P	0.7823	0.9432	0.0474	0.0032	0.0442
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8192	P	0.4489	0.8476	0.3038	0.0232	0.2805

- The test statistic corresponds to:

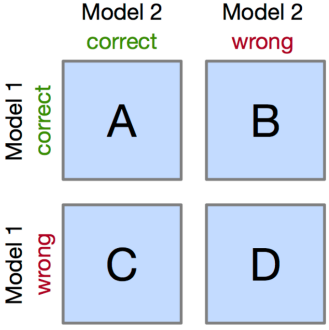
$$T_{\text{t-test}} = \sqrt{8192}\,\frac{0.0016}{0.178} = 79.523$$

- For $\alpha = 0.05$, the critical values to reject H_0 are -1.9637 and 1.9637.
- For the given significance level, there is enough evidence to reject H_0. We assume the is a difference in performance on this population of data.

MCNEMAR TEST FOR CLASSIFICATION

- McNemar test is non-parametric (no distributional assumptions).
- Caveat: Only works for classification tasks.
- H_0: $(A + B = A + C) \wedge (C + D = B + D) \Rightarrow H_0 : B = C$ vs. H_1: $B \neq C$
- Test statistic:

$$T_{\text{McNemar}} = \frac{(|B-C|-1)^2}{B+C} \sim \chi^2_1$$



- A: # obs. correctly classified by both.
- B: # obs. misclassified by model 2 but not by model 1.
- C: # obs. misclassified by model 1 but not by model 2.
- D: # obs. misclassified by both.

MCNEMAR TEST

Continuing previous example:

		ranger	
		correct	wrong
rpart	correct	5942	1
	wrong	2232	17

Calculate the test statistic:

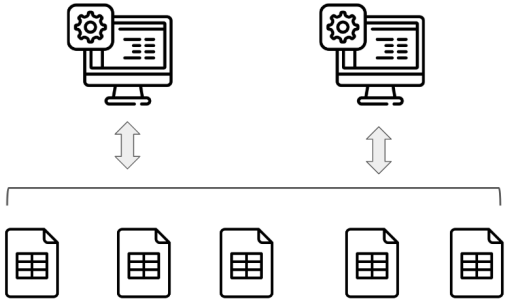
$$T_{\text{McNemar}} = \frac{(|1 - 2232| - 1)^2}{1 + 2232} \approx 2227$$

- For $\alpha = 0.5$, the critical value is 3.841.
- $T_{\text{McNemar}} > 3.841 \Rightarrow$ Reject H_0.
- We conclude that the tree and random forest do not have the same performance on this population of data.
- The t-test performs on observation-wise loss values, while the McNemar test performs on aggregate classification errors.

Scenario 2: Two Algorithms on Multiple Data Sets

SECOND BENCHMARKING SCENARIO

Two algorithms on multiple data sets



WILCOXON SIGNED RANK TEST

- Non-parametric and suitable for any performance measure.
- H_0: Algs. 1 and 2 have tied ranks in model performance.
  H_1: Algs. 1 and 2 have different ranks in performance.
- Consider M data sets with performance $\widehat{\mathrm{GE}}(\widehat{f}_k, \rho, \mathcal{D}_{\text{test},m})$ for model k and data set m. The difference d_m between GE estimates on the m-th data set is:

$$d_m = \widehat{\mathrm{GE}}(\widehat{f}_1, \rho, \mathcal{D}_{\text{test},m}) - \widehat{\mathrm{GE}}(\widehat{f}_2, \rho, \mathcal{D}_{\text{test},m})$$

- Let R^+ denote the rank sum of absolute differences $|d_m|$ for data sets where algorithm 1 outperforms algorithm 2 and vice versa for R^-. Ranks of $d_m = 0$ are split evenly among the sums (if there is an odd number of them, one is ignored):

$$R^+ = \sum_{d_m>0} \mathrm{rank}(|d_m|) + \frac{1}{2}\sum_{d_m=0} \mathrm{rank}(|d_m|) \qquad R^- = \sum_{d_m<0} \mathrm{rank}(|d_m|) + \frac{1}{2}\sum_{d_m=0} \mathrm{rank}(|d_m|)$$

- The test statistic does not follow a closed-form distribution, but probabilities can be derived from a distribution table:

$$T_{\text{Wilcoxon}} = \frac{\min(R^+, R^-) - \frac{1}{4}M(M+1)}{\sqrt{\frac{1}{24}M(M+1)(2M+1)}}$$

WILCOXON SIGNED RANK TEST

- A benchmark of rpart and ranger (with default hyperparameters) on 6 classification tasks, using the classification error as performance measure:

	task_id	rpart	ranger	diff	rank
1	pollen	0.4987	0.5005	-0.0018	2
2	threeOf9	0.1543	0.0117	0.1426	6
3	fri_c1_500_5	0.1900	0.1180	0.0720	4
4	space_ga	0.2240	0.1629	0.0611	3
5	kin8nm	0.2795	0.1608	0.1187	5
6	monks-problems-3	0.0108	0.0108	0.0000	1

$$R^+ = 2 \qquad R^- = 15 \qquad \min(R^+, R^-) = 2$$

$$T_{\text{Wilcoxon}} = \frac{2 - \frac{1}{4}6(6+1)}{\sqrt{\frac{1}{24}6(6+1)(12+1)}} = -1.78$$

- For $\alpha = 0.05$, the critical value (two-sided) is 1.
- $T_{\text{Wilcoxon}} < 1 \Rightarrow$ reject H_0!
- There is evidence that the two algorithms do not perform equally well.

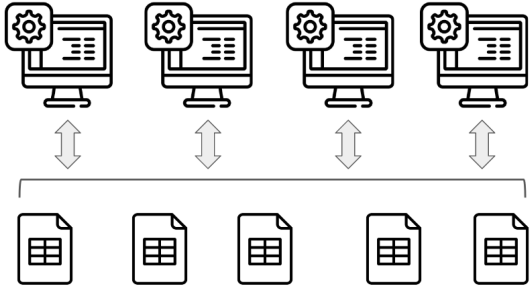
FURTHER NOTES

- Wilcoxon Signed Rank Test assumes commensurability of scores $\rightarrow$ use sign test if violated
- Researchers use the Wilcoxon Signed Rank Test to compare two algorithms on the same data, but violate independence assumption

Scenario 3: Multiple Algorithms on Multiple Data Sets

THIRD BENCHMARKING SCENARIO

Multiple algorithms on multiple data sets



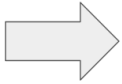
COMPARING MULTIPLE MODELS

- So far: comparing perf. of 2 models
- Now: comparing multiple models using omnibus test
- Omnibus hypotheses to be tested are:
  - H_0 : All algorithms are equivalent in their performance and hence their average ranks should be equal.
  - H_1 : The average rank for at least one algorithm is different.
- Omnibus tests are only useful to indicate that at least one model performs differently than the others.
  $\Rightarrow$ If yes, run post-hoc pairwise comparisons.

If null hypothesis rejected:

Omnibus test

Null hypothesis:
All models perform equally well.



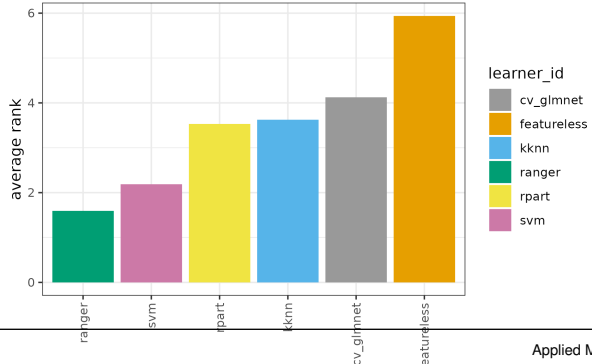
Post-hoc pairwise tests

Null hypothesis:
Two models perform equally well.

FRIEDMAN TEST

- Example of a benchmark of 6 mlr3 learners (with default hyperparameters) on 16 classification tasks, using the classification error as performance measure:

	task_id	featureless	cv_glmnet	rpart	ranger	kknn	svm
1	pollen	0.5148 (5.5)	0.5148 (5.5)	0.4987 (1)	0.5005 (2)	0.5107 (4)	0.5013 (3)
2	threeOf9	0.4648 (6)	0.2228 (5)	0.1543 (4)	0.0117 (2)	0.1231 (3)	0.0078 (1)
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
16	strikes	0.5312 (6)	0.4016 (5)	0.0993 (2)	0.0288 (1)	0.1184 (3)	0.2528 (4)



FRIEDMAN TEST

After ranking algorithms, we calculate the following quantities (where M denotes the number of data sets and K the number of algorithms):

- The overall mean rank:

$$\bar{R} = \frac{1}{mk}\sum_{k=1}^{K}\sum_{m=1}^{M} R_{mk} = \frac{1}{96}\sum_{k=1}^{6}\sum_{m=1}^{16} R_{mk} = 3.5$$

- The total sum of squares:

$$SS_{Total} = M\sum_{k=1}^{K}(\bar{R}_{.k} - \bar{R})^2 = 16\sum_{k=1}^{6}(\bar{R}_{.k} - 3.427)^2 \approx 187.281$$

with the mean rank for the k-th algorithm $\bar{R}_{.k} = \frac{1}{M}\sum_{m=1}^{M} R_{mk} = \frac{1}{16}\sum_{m=1}^{16} R_{mk}$

- The error sum of squares:

$$SS_{Error} = \frac{1}{M(K-1)}\sum_{m=1}^{M}\sum_{k=1}^{K}(R_{mk} - \bar{R})^2 = \frac{1}{80}\sum_{m=1}^{16}\sum_{k=1}^{6}(R_{mk} - 3.427)^2 \approx 3.481$$

FRIEDMAN TEST

- The Friedman test statistic is calculated as:

$$T_{\text{Friedman}} = \frac{SS_{Total}}{SS_{Error}} \approx 52.872$$

- For sufficiently large M (> 15) and K (> 5):

$$T_{\text{Friedman}} \sim \chi^2_{K-1}$$

- The $(1 - \alpha)$ quantile of χ^2_5 corresponds to 11.075, leading us to reject H_0.

- We assume there is at least one algorithm performing better than the remaining algorithms. Which one(s) needs to be determined with pairwise post-hoc tests.

POST-HOC NEMENYI TEST

- The post-hoc Nemenyi test is used after an omnibus test (like the Friedman test) rejects the null hypothesis.
- Runs all $\frac{K(K-1)}{2}$ pairwise comparisons for M data sets and K algorithms.
- Calculates the average rank of k-th algorithm on all M data sets:

$$\bar{R}_{.k} = \frac{1}{M} \sum_{m=1}^{M} R_{m,k}$$

- $H_0 : \bar{R}_{.k_1} = \bar{R}_{.k_2}$ versus $H_1 : \bar{R}_{.k_1} \neq \bar{R}_{.k_2}$
- The critical difference in mean ranks to reject H_0 corresponds to:

$$|\bar{R}_{.k_1} - \bar{R}_{.k_2}| \geq \frac{q_{\alpha,\infty,K}}{\sqrt{2}} \sqrt{\frac{K(K+1)}{6M}}$$

where $q_{\alpha,\infty,K}$ is the studentized range statistic for significance level α, infinite degrees of freedom, and K samples for each group (here, a group corresponds to one data set).

POST-HOC NEMENYI TEST

Example (for pairwise comparison between rpart and ranger):

$$\bar{R}_{.rpart} = \frac{1}{6}\sum_{m=1}^{16} R_{m,\text{rpart}} = 3.53$$

$$\bar{R}_{.\text{ranger}} = \frac{1}{6}\sum_{m=1}^{16} R_{m,\text{ranger}} = 1.59$$

$$|\bar{R}_{.rpart} - \bar{R}_{.ranger}| = 1.94$$

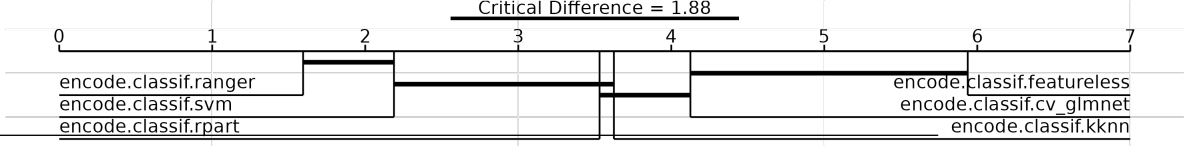
- For $\alpha = 0.05$ and $K = 6$, the critical difference in average ranks is:

$$|\bar{R}_{.\text{rpart}} - \bar{R}_{..\text{ranger}}| \geq \frac{4.02}{\sqrt{2}}\sqrt{\frac{6(6+1)}{6 \cdot 16}} = 1.88$$

- $|\bar{R}_{.rpart} - \bar{R}_{.ranger}| > 1.88 \Rightarrow$ Reject H_0!
- There is enough evidence to reject the claim that on more data sets, ranger and rpart would have the same average rank.

CRITICAL DIFFERENCE PLOT

- The critical difference (CD) plot summarizes all pairwise tests in a single graph.
- The CD is the pairwise difference in mean ranks required to reject H_0, thus indicating a significant difference in performance.
- We can see the ordered mean ranks of all algorithms on a single line. If the distance between two algorithms exceeds the CD, the left algorithm significantly outperforms the right one.
- **Continuing example:** The difference in mean ranks between ranger and the SVM does not exceed the CD; the difference between ranger and rpart does.



POST-HOC BONFERRONI-DUNN TEST

- Instead of pairwise comparisons, we now compare all algorithms with one baseline algorithm (i.e., $K - 1$ comparisons).
- The Bonferroni-Dunn test uses a similar test statistic which is standard normally distributed under H_0:

$$T_{\text{Bonferroni-Dunn}} = \frac{\bar{R}_{.k} - \bar{R}_{.baseline}}{\sqrt{\frac{K(K+1)}{6M}}}$$

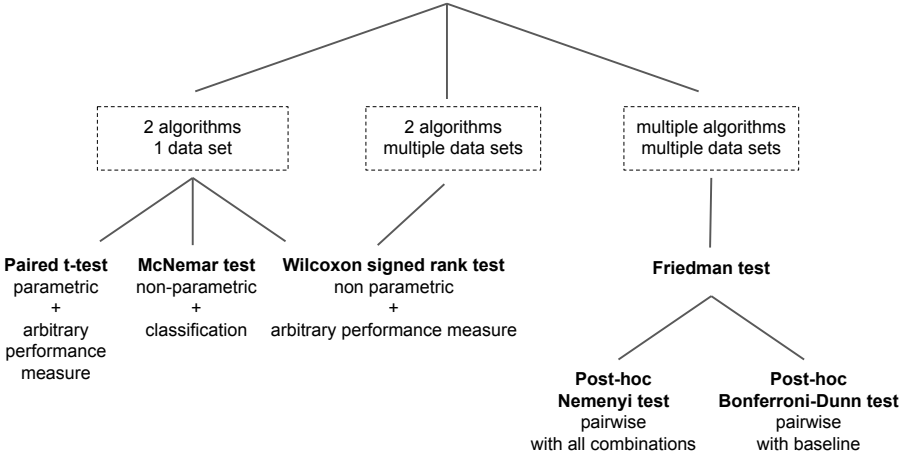
- It uses the Bonferroni correction for multiple testing, so the significance level α is adjusted downward via the number of comparisons:

$$\alpha_{\text{corrected}} = \frac{\alpha}{K - 1}$$

- For instance, $\alpha = 0.05$ is adjusted to $\alpha_{\text{corrected}} = \frac{0.05}{5} = 0.01$.
- Power is greater for Bonferroni-Dunn than for Nemenyi due to only comparing to a control classifier and not between all algorithms.

Conclusion

OVERVIEW OF DIFFERENT HYPOTHESIS TESTS



GUIDELINES ON HYPOTHESIS TESTS

- There is no golden standard for using hypothesis testing in performance evaluation. Often, tests are based on questionable or unverifiable assumptions or conclusions.
- The Wilcoxon and Friedman tests have been demonstrated to be most useful. They partially fulfil commensurability, do not assume normal distributions or homogeneity of variance, and can be applied to any evaluation metric (e.g., accuracy, error ratios).
- "Tests provide certain reassurance about the validity and non-randomness of the published results. For that to be true, they should be performed correctly and the resulting conclusions should be drawn cautiously. On the other hand, statistical tests should not be the deciding factor for or against publishing the work. Other merits of the proposed algorithm that are beyond the grasp of statistical testing should also be considered and possibly even favoured over pure improvements in predictive power." (Japkowiecz)

GUIDELINES ON MODEL SELECTION

- Know what you are optimizing $\rightarrow$ be suspicious of huge improvements
- Understand the difference between comparing trained models and machine learning algorithms
- Your machine learning metric is only a proxy for your business metric $\rightarrow$ gains might not materialize in real-world improvement
- Non-i.i.d. data require special treatment
- Multi-objective model selection and tests are still an open area of research