Stanford
MEDICINE

Department of Psychiatry
and Behavioral Sciences

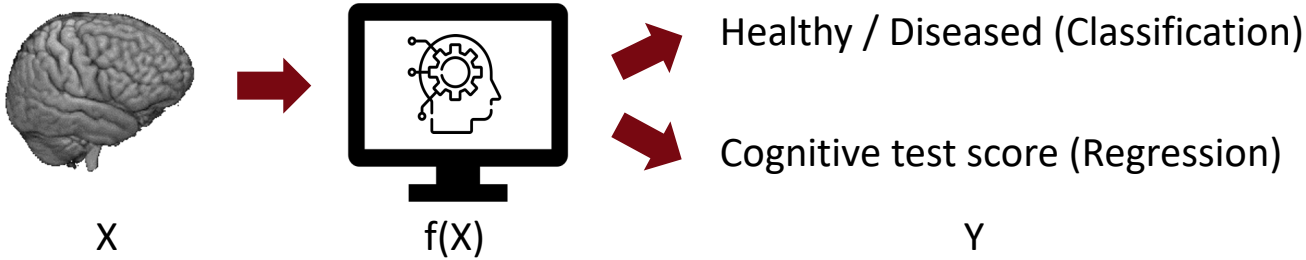
Machine Learning for Neuroimaging

Autumn 2023

Session 5 or 6 – 10/2023

Statistical analysis on ML models

Statistical Tests for Machine Learning



- Is classifier 1 significantly better than classifier 2?
- Does the disease impact the brain (structural or functional organization)?
  - Is the accuracy of the classifier significantly better than chance?
- Is the association between X and Y stronger in females than males?
  - Is the classifier significantly more accurate in females?
- Is the association between X and Y stronger than the association between X and Z

…

Statistical Tests for Machine Learning

Beat SOTA model by 0.6%

90.8±5% vs. 90.2±2%

Statistical Tests for Model Validation (Classification)

Healthy / Diseased (Classification)

Correct or wrong?

Ground-truth diagnosis label

Chi-Squared Test

Null Hypothesis: the ML prediction is no better than guessing (50%/50%)

X^2 follows a Chi-squared distribution of DOF = 1

Observed $X^2 = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50}$ → p=0.045

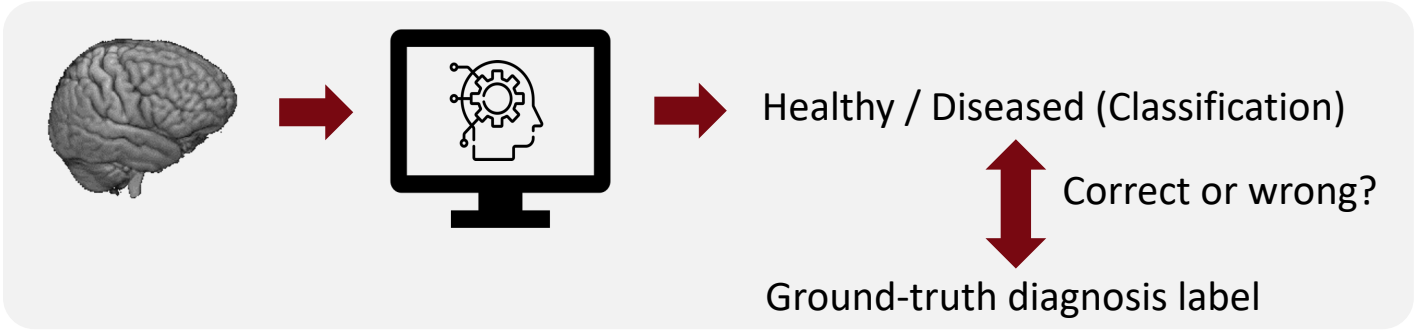
Prediction

Correct	Wrong	Total
60	40	100

Expected

Correct	Wrong	Total
50	50	100

Statistical Tests for Model Validation (Classification)



Chi-Squared Test

Null Hypothesis: the ML prediction is impartial to males and females

Review the example of handedness in males and females (X^2 follows Chi-squared distribution of DOF = 1)

Prediction

Cohorts	Correct	Wrong	Total
Male	75	35	110
Female	64	26	90
Total	139	61	200

Expected

Cohorts	Control	Disease
Control	76.45	33.35
Disease	62.55	27.45

Comparing two ML models on a cohort

Chi-Squared Test

Null Hypothesis: the two ML models are equally accurate (when tested on the same dataset)

Prediction

ML Models	Correct	Wrong	Total
Model 1	75	25	100
Model 2	64	36	100
Total	139	61	200

?

- Violating the independent samples assumption !
- Duplicating the 100 samples to 200 samples !
- Do NOT construct contingency matrix like this !

McNemar's Test

- McNemar's test is applied to 2 × 2 contingency tables with matched pairs of subjects to determine whether the row and column marginal frequencies are equal.

Null Hypothesis: the two ML models are equally accurate on the same cohort

Null Hypothesis: marginal frequencies are equal.

Model 2

Model 1	Correct	Wrong	Total
Correct	60	5	65
Wrong	20	15	35
Total	80	20	100

Marginal frequencies

McNemar's Test

- McNemar's test is applied to 2 × 2 contingency tables with matched pairs of subjects to determine whether the row and column marginal frequencies are equal.

Null Hypothesis: the two ML models are equally accurate on the same cohort

Null Hypothesis: off-diagonal frequencies are equal.

Model 2

Model 1	Correct	Wrong	Total
Correct	60	5	65
Wrong	20	15	35
Total	80	20	100

	Test 2 positive	Test 2 negative
Test 1 positive	a	b
Test 1 negative	c	d

a	60
b	5
c	20
d	15
Test type	Standard McNemar's test
Test statistic χ^2	9
p-value	0.002699796

Pitfall of Class Imbalance

Model 1: Always predicting control

Prediction

Ground Truth	Control	Disease	Total
Control	80	0	80
Disease	20	0	20
Total	100	0	100

Model 2: Random guessing

Prediction

Ground Truth	Control	Disease	Total
Control	40	40	80
Disease	10	10	20
Total	50	50	100

McNemar's test $p<0.0001$, Model 1 is better ?

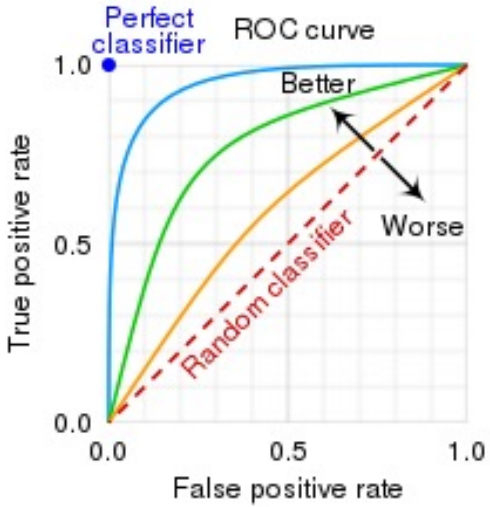
Model 2

Model 1	Correct	Wrong	Total
Correct	40	40	80
Wrong	10	10	20
Total	50	50	100

Both models are non-informative !

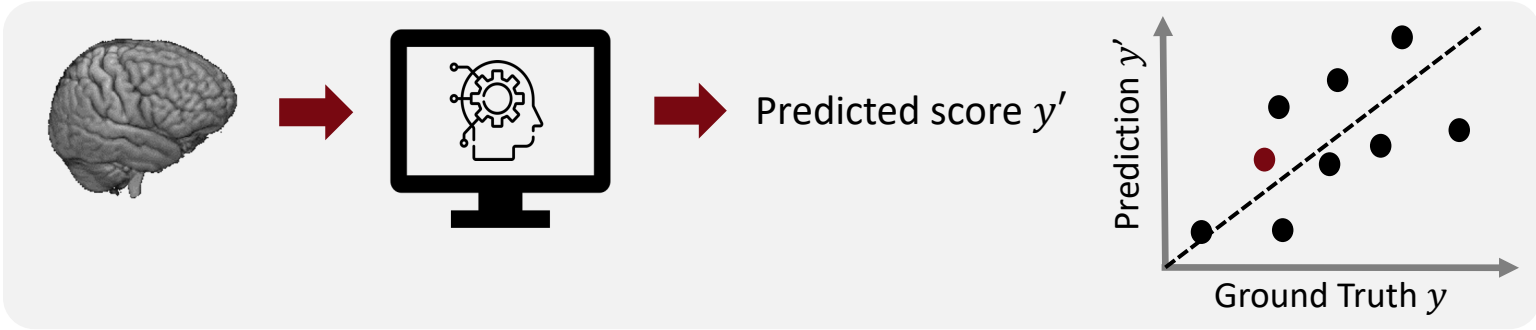
Balanced Accuracy and AUC

- Alternative metrics for binary classifiers
  - Balanced accuracy (Bacc): (true positive rate + true negative rate)/2
  - Area under the ROC curve (AUC)
- Testing Procedures
  - Hardin-Shumway test
    - Build probability distribution of Model-1 AUC (or Bacc) via bootstrapping
    - Compare Model-2 AUC (or Bacc) to the Model-1 distribution to derive *p*-value
  - DeLong's test
    - the Mann-Whitney *U*-statistics



Statistical Significance and Normalized Confusion Matrices, Photogramm. Eng. Remote Sens, 1997
Comparing areas under two or more correlated reciever operating characteristics curves: a nonparamentric approach. Biometrics, 1988

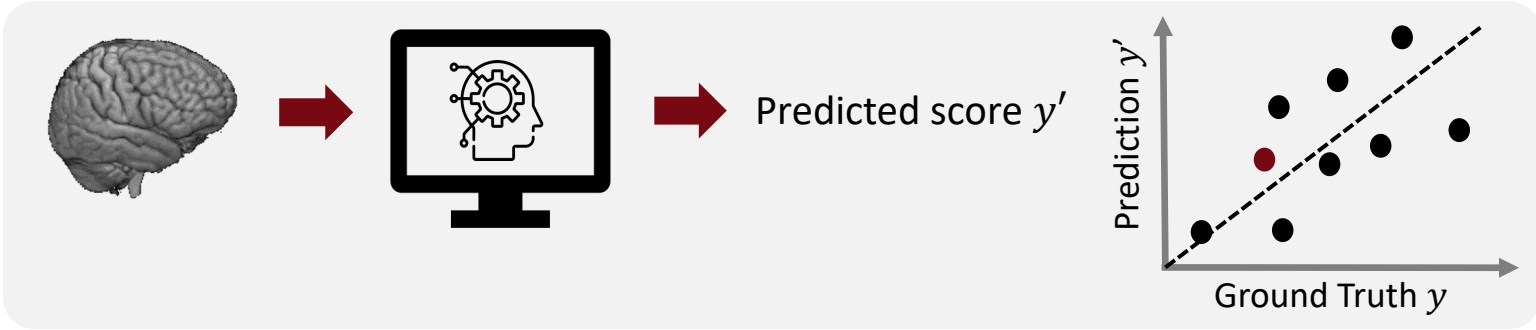
Statistical Tests for Model Validation (Regression)



Null Hypothesis: the ML prediction is no better than guessing

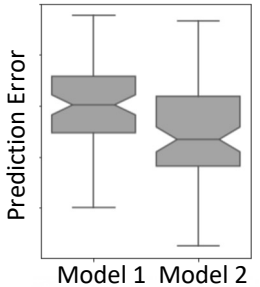
Test for Pearson's correlation between y and y'

Statistical Tests for Model Validation (Regression)



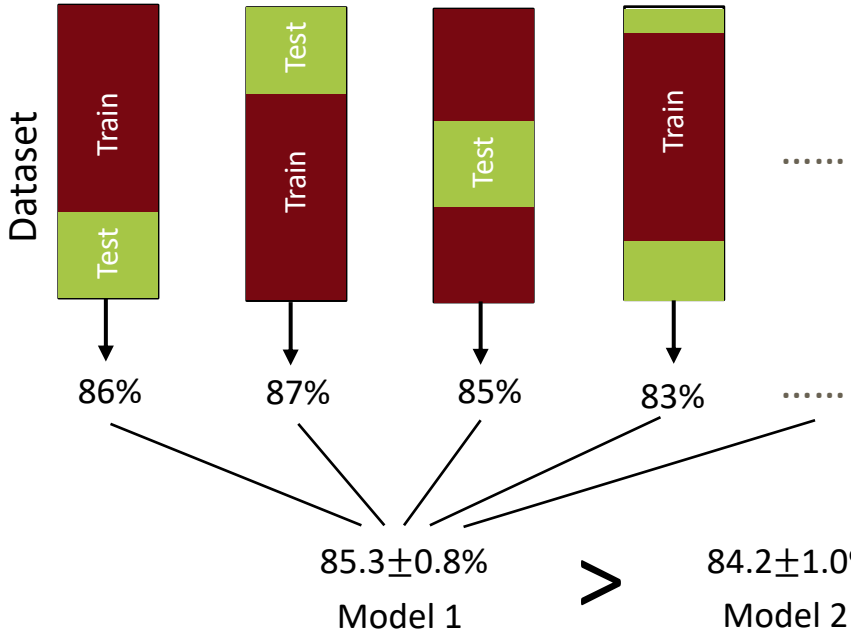
Null Hypothesis: Model 1 prediction is no better than Model 2 prediction

- Absolute error of a sample $\mathrm{err_i} = |\mathrm{y_i'} - y_i|, \mathrm{i} = 1, \ldots, \mathrm{N}$
  - Errors associated with N samples by Model-1 $\{\mathrm{err_i^1} | \mathrm{i} = 1, \ldots, \mathrm{N}\}$
  - Errors associated with N samples by Model-2 $\{\mathrm{err_i^2} | \mathrm{i} = 1, \ldots, \mathrm{N}\}$
- Compare $\{\mathrm{err_i^1}\}$ and $\{\mathrm{err_i^2}\}$ by paired t-test



Pitfall of Cross-Validation

- Accuracy scores from K-fold or Monte-Carlo Simulation



- Scores are not independent
- Arbitrarily inflated sample size

Significantly better by *t*-test ?