THE UNIVERSITY of EDINBURGH
informatics

Applied Machine Learning (AML)

Class Starting at 4:10pm

Oisin Mac Aodha • Siddharth N.

Applied Machine Learning

Week 7: Evaluation and Model Selection

This slides will be made available on the project website after the class.
This session will be recorded.

Overview

1) Outline your tasks this for week
2) Discussion of Week 6’s topics

Coursework Progress Report

- Progress reports due by 5pm, 29 Oct (Wed) (not assessed)
- Submit through form linked on website
  - Only one team member needs to submit
  - Use 2-digit 0-padded naming convention [e.g. 06.pdf, 23.pdf]
- (Optional) Feedback session 1-4pm, 31 Oct (Fri); AT 5.04
  - Primarily to check you are not miscalibrated
  - Will communicate schedule once we know signup

Coursework Final Submission

- Final submission due 12pm, 20 Nov (Thu)
  - Only report submission required here
- Supplementary materials deadline in the following week
  - Source for project (Readme + minimal structure)
  - Source for report
- Further details for both announced later

Coursework: Communication & Contribution

- Important to communicate with teammates — do not neglect / delay
- Regular contact ensures everyone is on the same page
- Individual grade based on final report grade **
    - Statement of contribution **required** (not counted for page limit)
    - Disparities will be taken into account for individual grades

- Also remember Guidance on use of Generative AI
    - **Must** explicitly acknowledge use and describe how it was used

Week 7: Your tasks for this week

1) Complete Lab 3
2) Watch videos for week 7
    - **Clustering** and **Non-Linear Dimensionality Reduction**
3) Ask questions on Piazza if stuck
4) Continue working on the coursework
5) Start Tutorial 3 which takes places next week - link in week 8

Evaluation

Which evaluation metrics are commonly used for evaluating the performance of a classification model when there is a class imbalance problem?

1. Accuracy and Error
2. Precision and Error
3. Precision and Recall
4. Error and Recall

Match the term to the formulation

1. Precision
2. Recall
3. False Positive Rate
4. True Positive Rate
5. False Negative Rate

$$A. \quad \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$B. \quad \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

$$C. \quad \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$D. \quad \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$$

Error Measures

- False Positive Rate (FPR) = $\dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
  - (False Alarm) % of ‘–’ misclassified as ‘+’
- False Negative Rate (FNR) = $\dfrac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}}$
  - (Miss) % of ‘+’ misclassified as ‘–’
- Recall / True Positive Rate (TPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
  - (1 - Miss) % of ‘+’ correctly predicted
- Precision / Positive Predictive Rate (PPR) = $\dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
  - % of ‘+’ out of all positive predictions

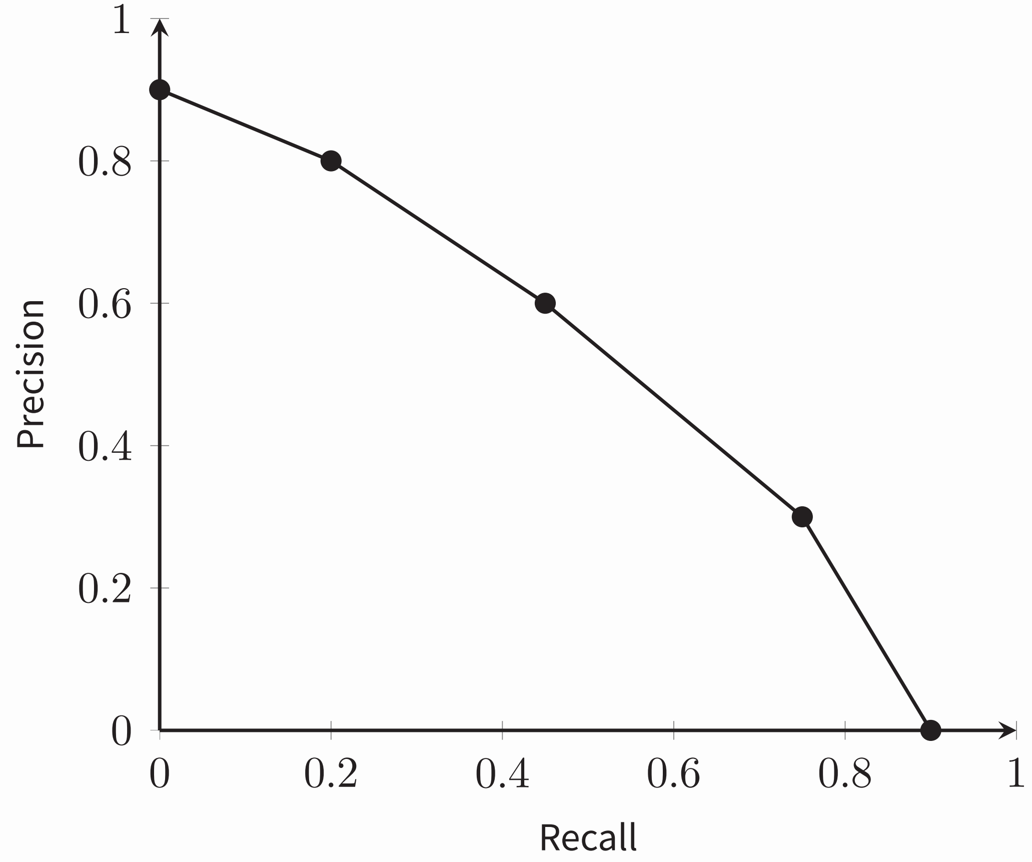
		Predicted	
		+	−
True	+	TP 200	FN 300
	−	FP 100	TN 400

Precision-Recall Curve

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

		Predicted +	Predicted −
True	+	TP 200	FN 300
	−	FP 100	TN 400

email	label	$p(y\|\boldsymbol{x})$
"send us your password"	+	0.92
"send us review"	−	0.80
"review your account"	−	0.72
"review us"	+	0.65
"send your password"	+	0.61
"send us your account"	+	0.43
⋮		



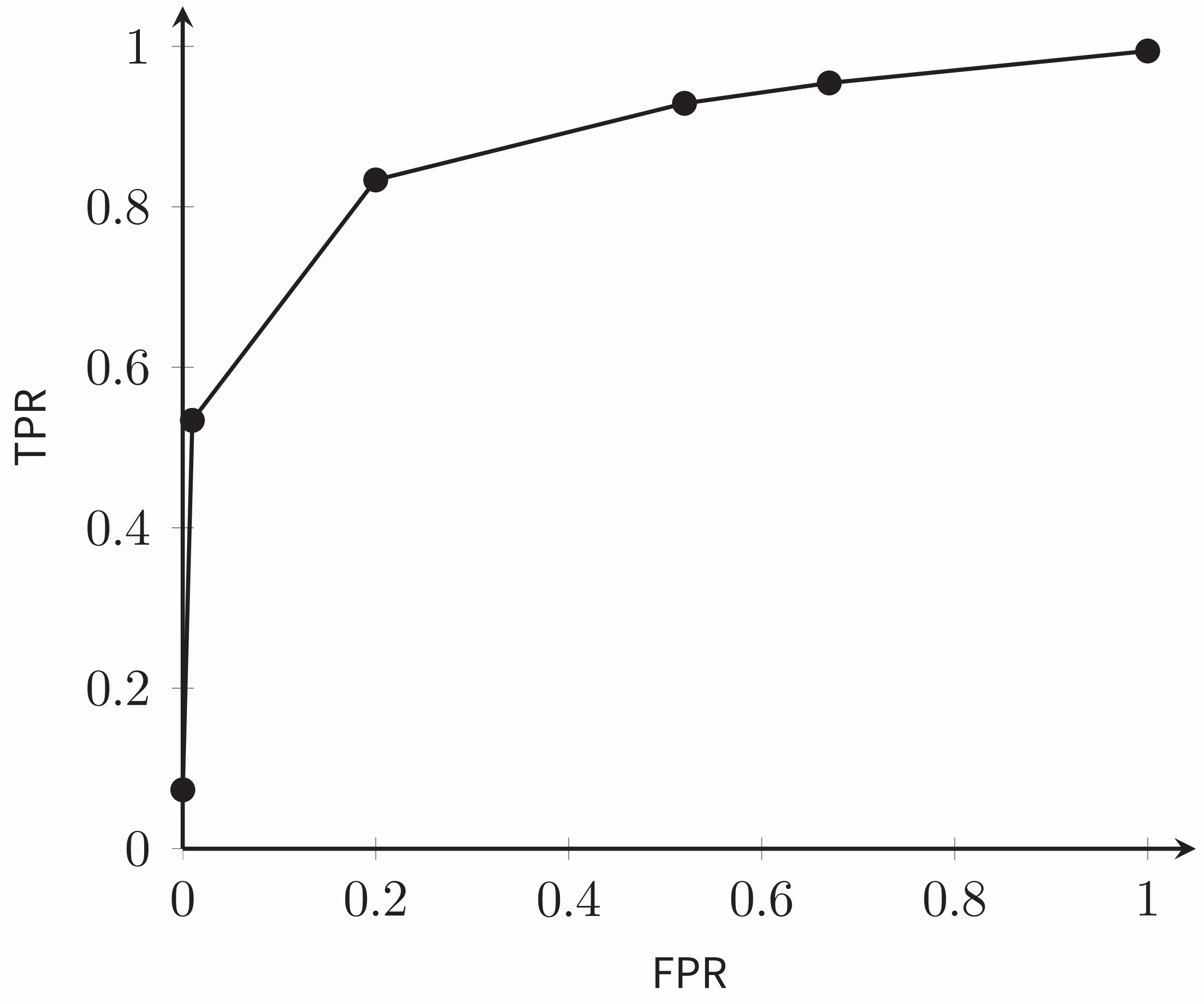
Receiver Operating Characteristic (ROC)

True Postive Rate (TPR) / Recall $= \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$ False Positive Rate (FPR) $= \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$

AUC

- area under ROC curve
- larger area $\implies$ better model

True \ Predicted	+	−
+	TP 200	FN 300
−	FP 100	TN 400

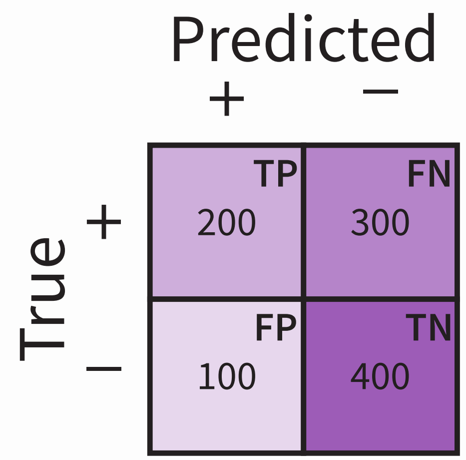


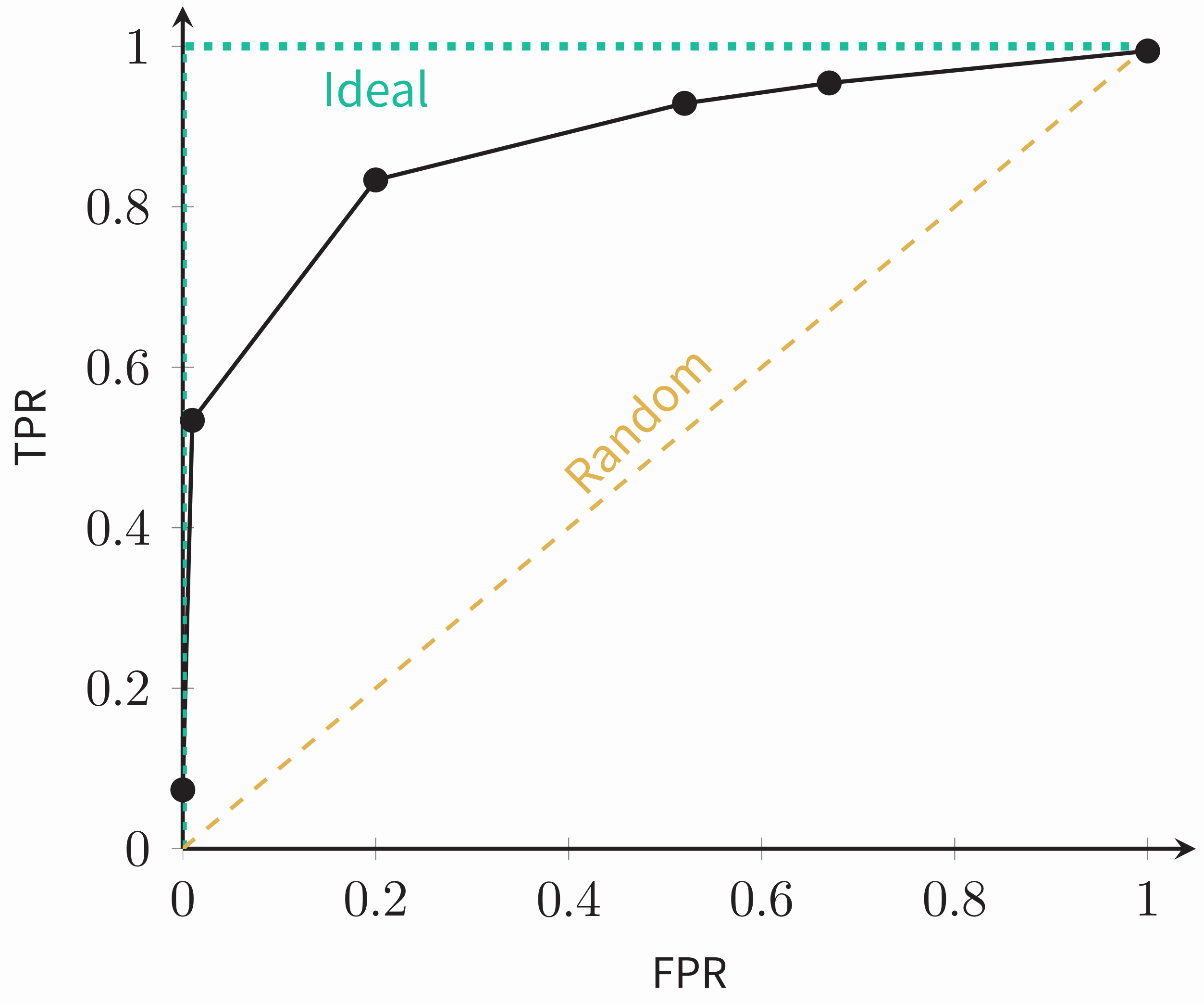
Receiver Operating Characteristic (ROC)

$$\text{True Postive Rate (TPR) / Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{False Positive Rate (FPR)} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

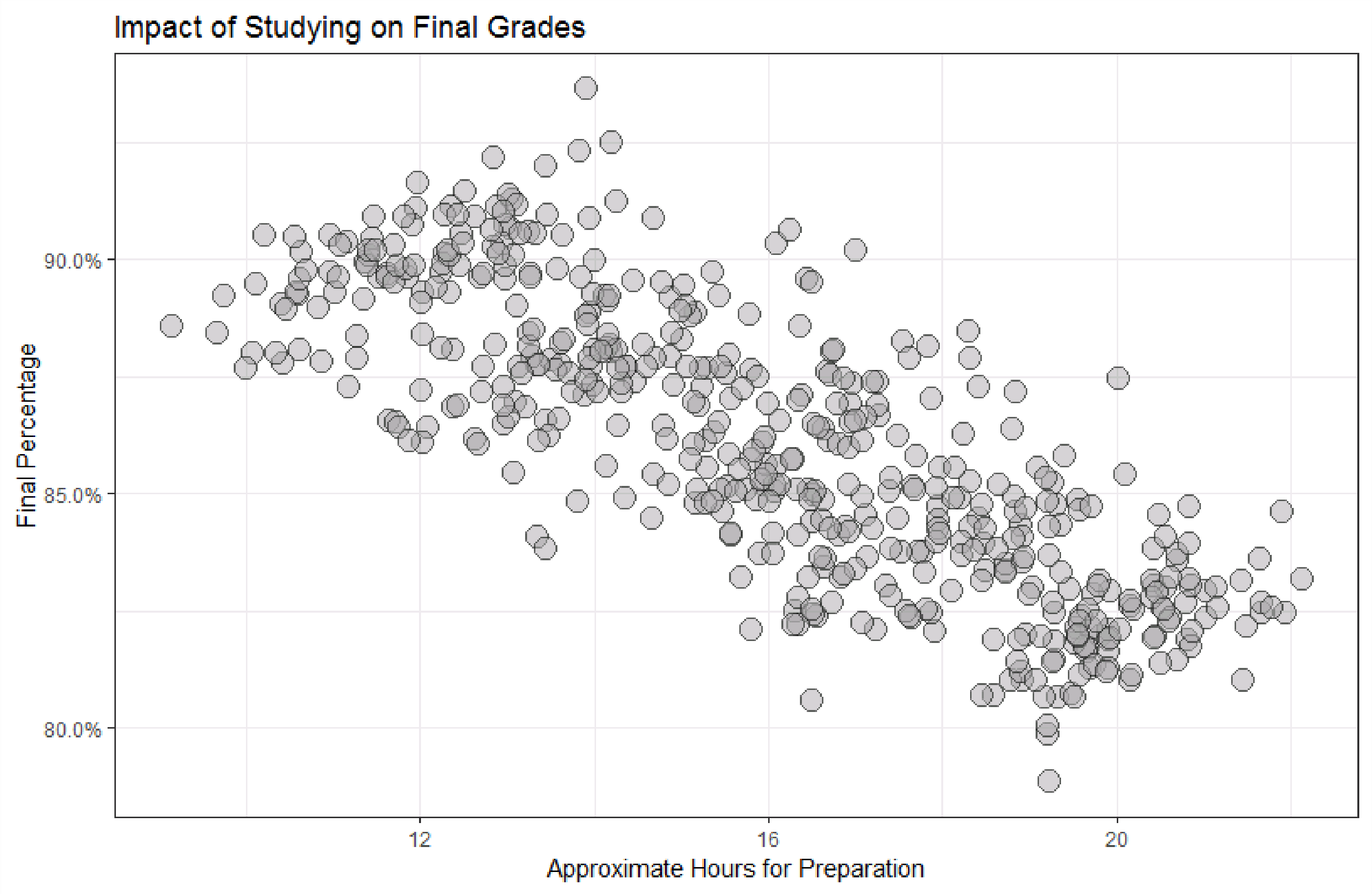
AUC

- area under ROC curve
- larger area $\implies$ better model

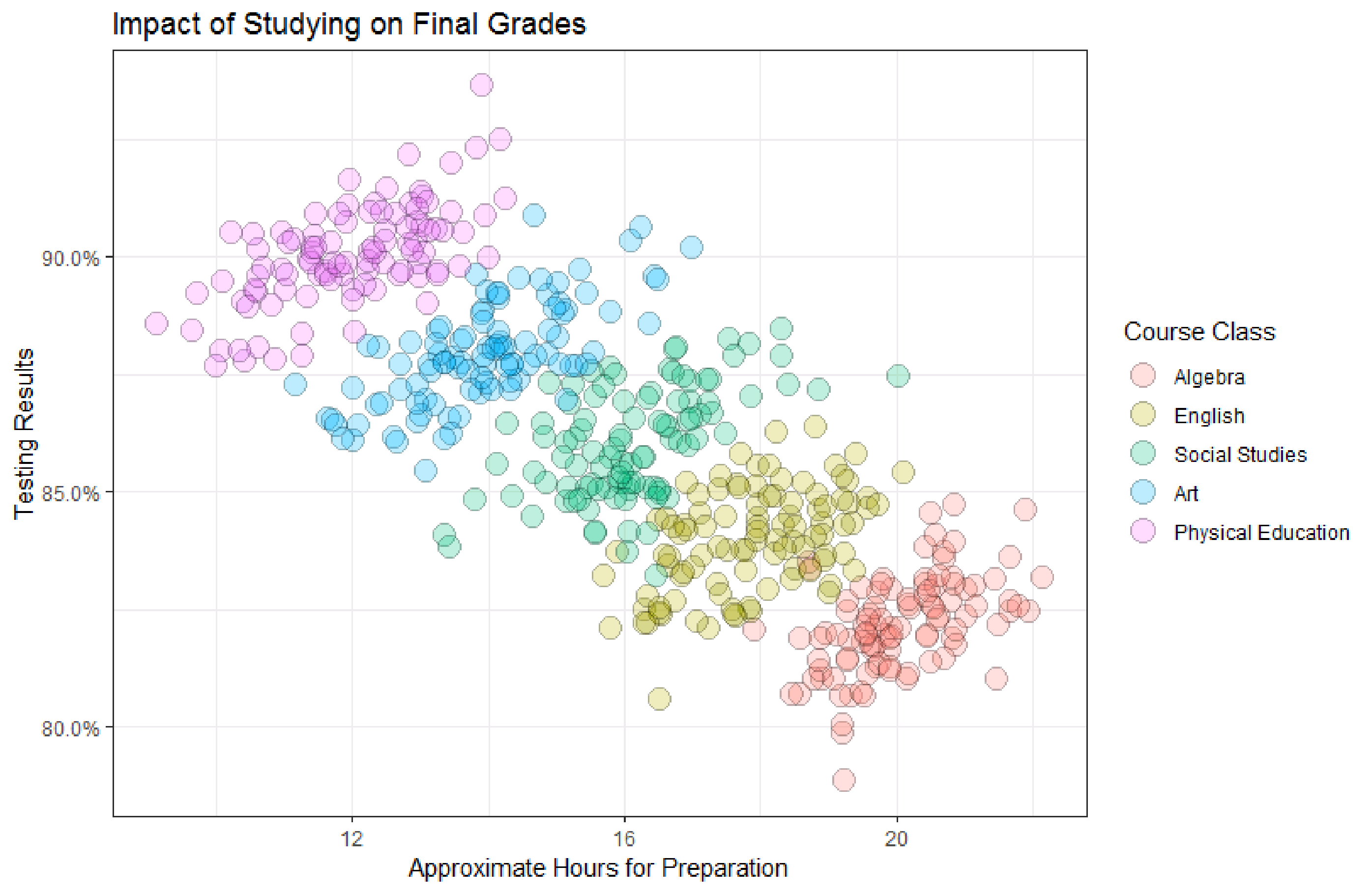




Simpson’s Paradox



Simpson’s Paradox



Model Selection

Preliminaries

Population vs. Sample statistics

Population: All the elements from a set

E.g. All leave-1-out splits of the dataset

Sample: Observations drawn from population

E.g. Some N splits of the dataset

If sample set is $x_1, \ldots, x_N$

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

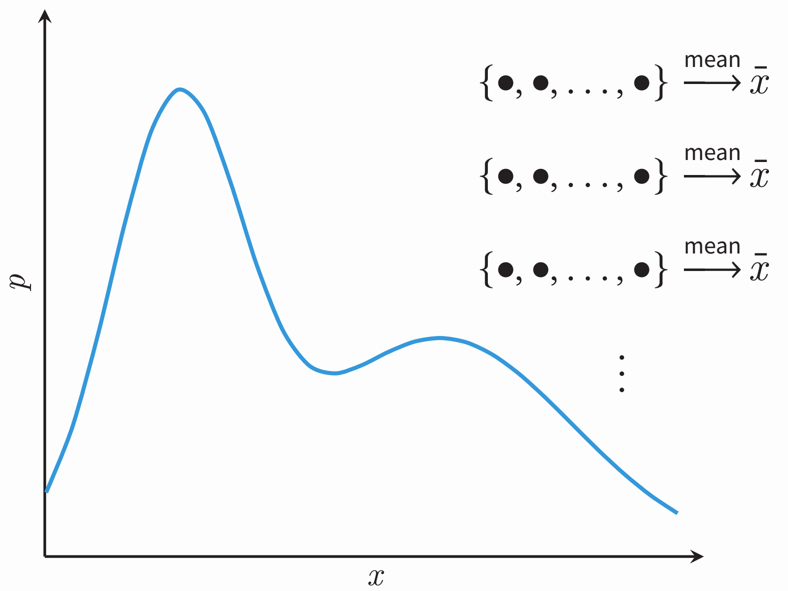
*Bessel's correction

Preliminaries

Central Limit Theorem (CLT)

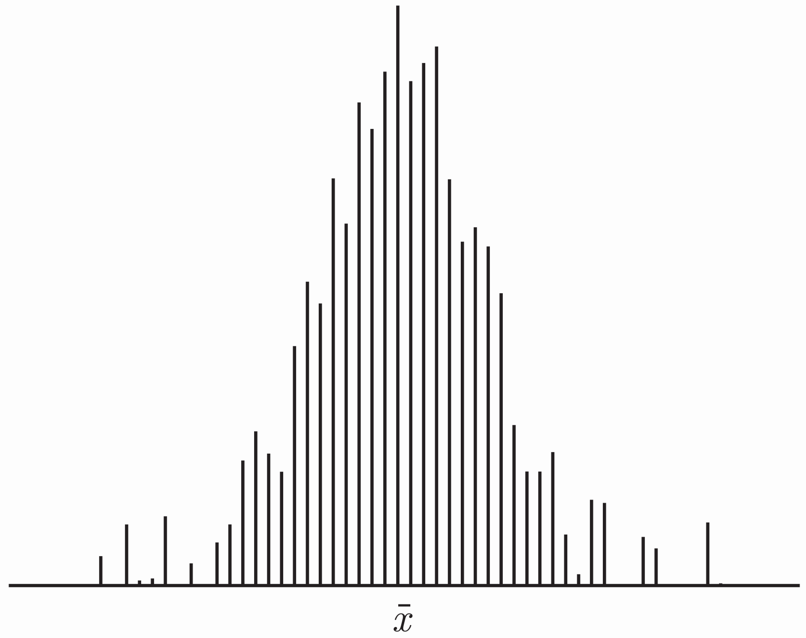
For a set of samples $x_1, \ldots, x_N, \ldots$ from a population with expected mean μ and finite variance σ^2

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \quad \text{as } N \to \infty$$



Assume

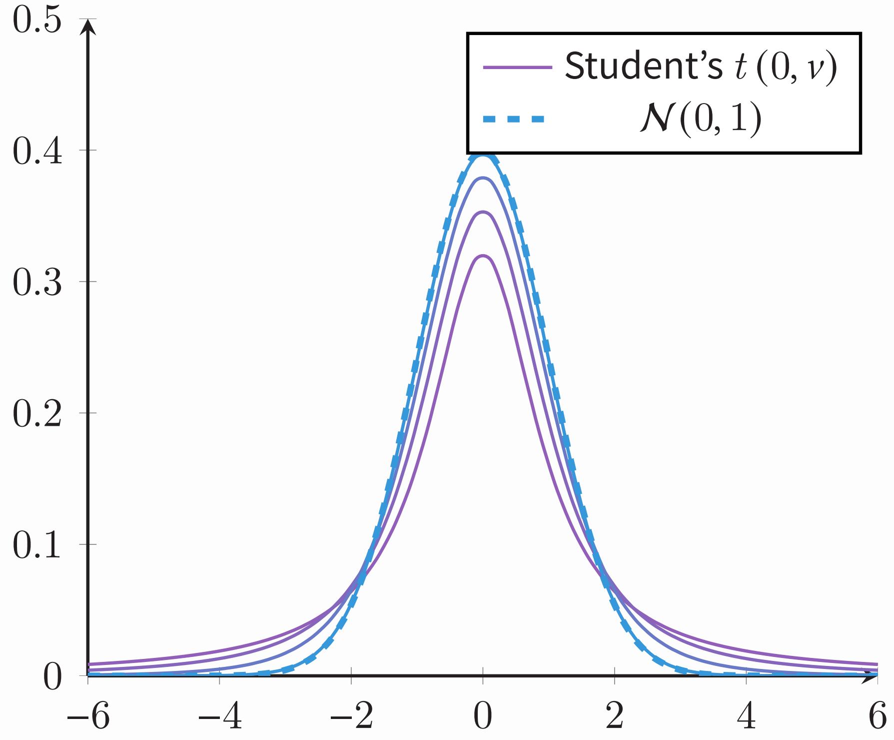
- population μ known
- population σ^2 known



Preliminaries

Student's-t distribution

- CLT: (weak) convergence to $\mathcal{N}(0,1)$ as $N \longrightarrow \infty$
- for smaller N, not Gaussian!



Assume

- population μ known
- population σ^2 *unknown*
- estimate sample variance $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x}_N)^2$

$$f(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

$$t = \frac{\overline{x} - \mu}{s/\sqrt{N}}, \qquad \nu = N - 1$$

Hypothesis Testing

- Formally examine two opposing conjectures (hypothesis): H_0 and H_1
- Mutually exclusive and exhaustive: $H_0 = \texttt{True} \implies H_1 = \texttt{False}$
- Analyse data to determine which is `True` and which is `False`

True \ Decision (Retain)	H_0	H_1
H_0	✓	Type II
H_1	Type I	✓

Null Hypothesis: H_0

- States the assumption to be tested
- Begin with assumption that $H_0 = \texttt{True}$
- Always evaluates (partial) equality ($=, \leq, \geq$)

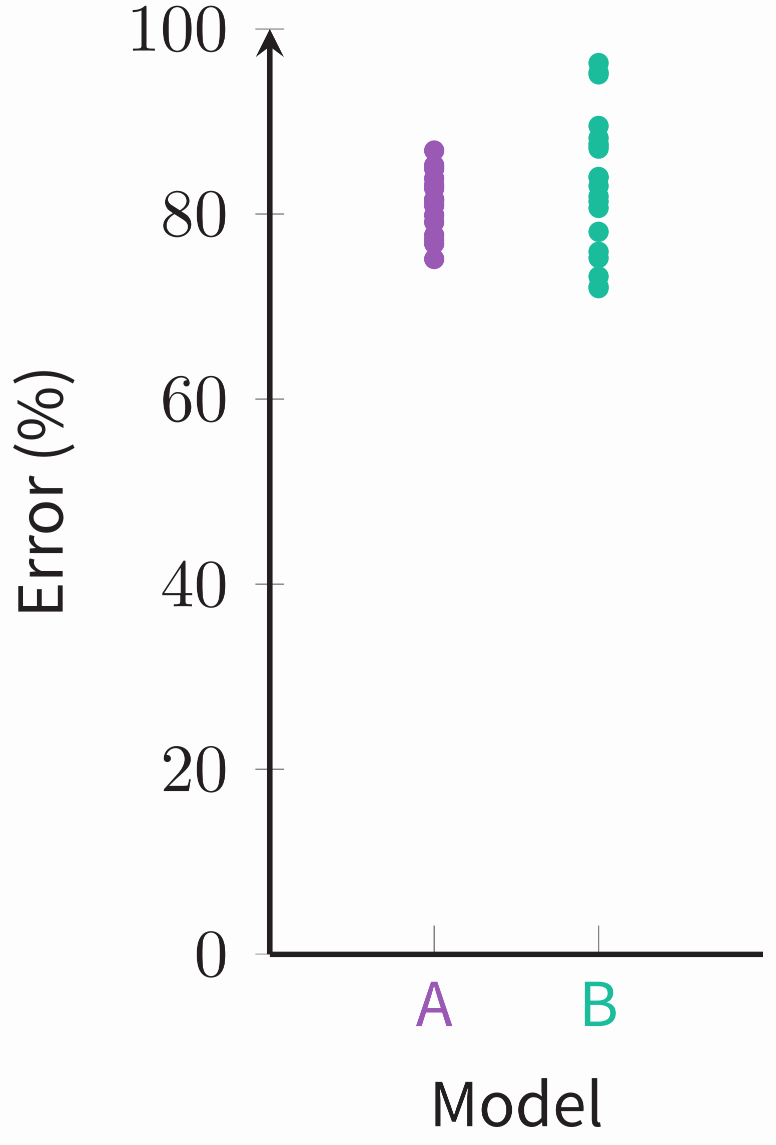
Alternative Hypothesis: H_1

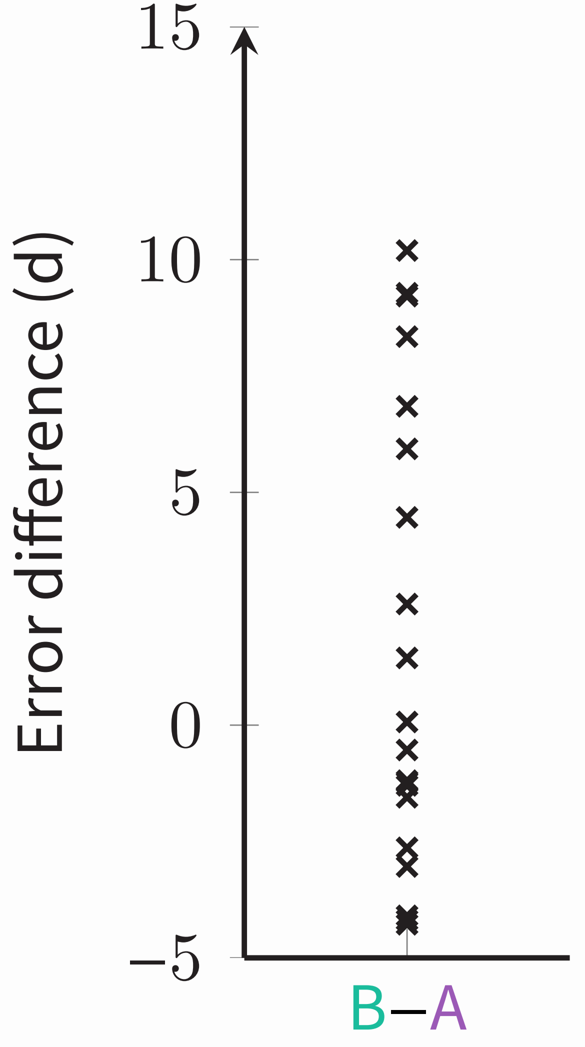
- States the assumption believed to be `True`
- Evaluate if evidence supports assumption
- Always evaluates (strict) *in*equality ($\neq, >, <$)

Example: Hypothesis Testing for Models

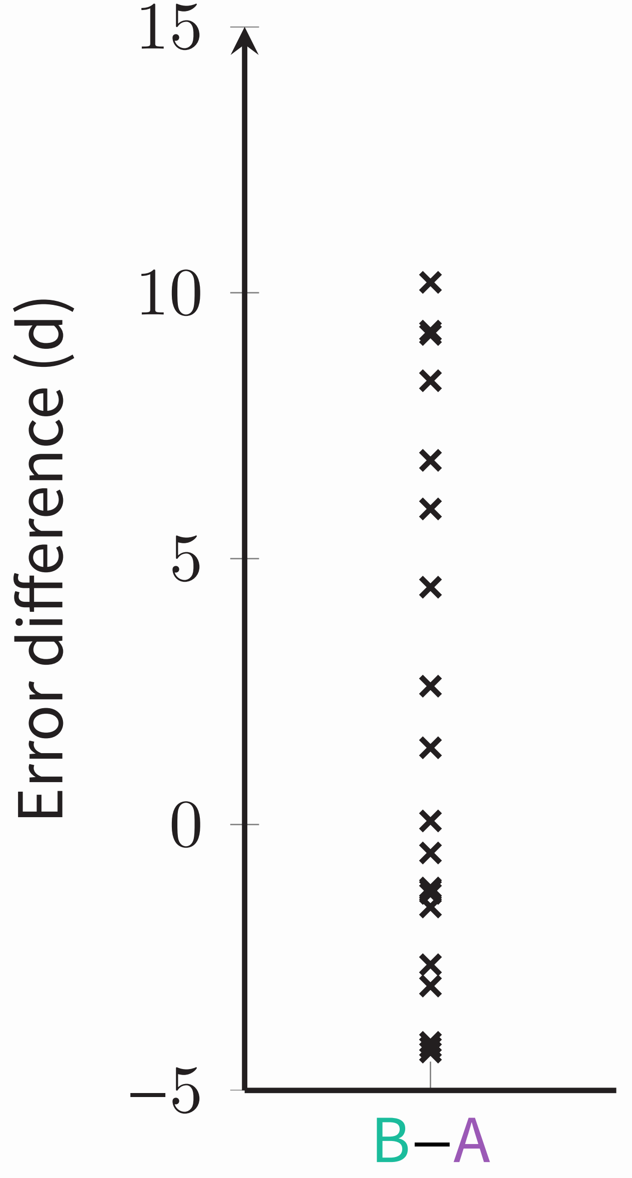
Generating Variation

Data Split	A	B	d
$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}$	ℓ_1^A	ℓ_1^B	$\ell_1^B - \ell_1^A$
$\{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}$	ℓ_2^A	ℓ_2^B	$\ell_2^B - \ell_2^A$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{\mathcal{D}^N_{\text{train}}, \mathcal{D}^N_{\text{test}}\}$	ℓ_N^A	ℓ_N^B	$\ell_N^B - \ell_N^A$





Example: Hypothesis Testing for Models



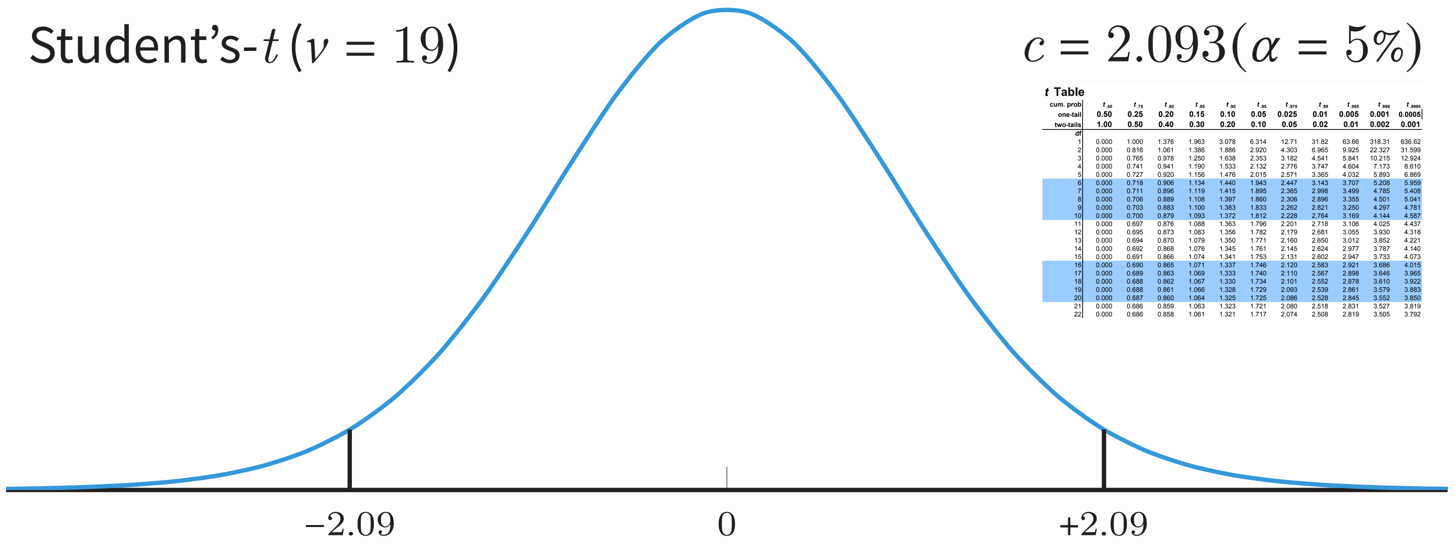
Hypotheses

$H_0 : \mu^d = 0$ $\alpha = 5\%$ (significance)

$H_1 : \mu^d \neq 0$ $N = 20$

Student's-$t\,(\nu = 19)$

$c = 2.093(\alpha = 5\%)$



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792

Hypothesis Testing: Caveats

- Rejecting H_0 does not imply 100% sure H_0 is `False`
- Failing to reject H_0 does not imply H_0 is `True`
- Confidence level $(\alpha = 0.05)$ is from convention; not always best
- Statistical significance does not imply practical *relevance*
  - Rejecting $H_0 : \mu^d = 0$ only tells us that $\mu^d \neq 0$ but not how big or important the difference is
  - **Remedy:** Report confidence interval (CI)

$$\bar{d} \pm c|_\alpha \cdot \frac{s}{\sqrt{N}}$$

which, for our example would be

$$2.53 \pm 2.093 \cdot \frac{5.27}{\sqrt{20}} \qquad (\alpha = 0.05,\, c|_{0.05} = 2.093)$$

$$2.53 \pm 2.47$$