THE UNIVERSITY *of* EDINBURGH
informatics

Applied Machine Learning (AML)

Model Selection

Oisin Mac Aodha • Siddharth N.

Direct Comparison

Comparing Evaluation Measures

email
"send us your password"
"send us review"
"review your account"
"review us"
"send your password"
"send us your account"
⋮

Comparing Evaluation Measures

email	true
"send us your password"	+
"send us review"	−
"review your account"	−
"review us"	+
"send your password"	+
"send us your account"	+
⋮	

Acc
κ
F1-score
ROC AUC
⋮

Comparing Evaluation Measures

email	true	pred (A)
"send us your password"	+	+
"send us review"	−	+
"review your account"	−	−
"review us"	+	−
"send your password"	+	+
"send us your account"	+	+
⋮		

	Naive Bayes (A)
Acc	72.6%
κ	54.1%
F1-score	85.6%
ROC AUC	48.4%
⋮	⋮

Comparing Evaluation Measures

email	true	pred (A)	pred (B)
“send us your password”	+	+	+
“send us review”	–	+	–
“review your account”	–	–	+
“review us”	+	–	–
“send your password”	+	+	+
“send us your account”	+	+	–
⋮			

	Naive Bayes (A)	Logistic Regression (B)
Acc	72.6%	84.5%
κ	54.1%	66.2%
F1-score	85.6%	89.1%
ROC AUC	48.4%	55.7%
⋮	⋮	⋮

Comparing Evaluation Measures

email	true	pred (A)	pred (B)
"send us your password"	+	+	+
"send us review"	−	+	−
"review your account"	−	−	+
"review us"	+	−	−
"send your password"	+	+	+
"send us your account"	+	+	−
⋮			

	Naive Bayes (A)	Logistic Regression (B)
Acc	72.6%	84.5%
κ	54.1%	66.2%
F1-score	85.6%	89.1%
ROC AUC	48.4%	55.7%
⋮	⋮	⋮

Clearly, logistic regression (B) has higher scores than naive Bayes (A)!

Comparing Evaluation Measures

email	true	pred (A)	pred (B)
"send us your password"	+	+	+
"send us review"	−	+	−
"review your account"	−	−	+
"review us"	+	−	−
"send your password"	+	+	+
"send us your account"	+	+	−
⋮			

	Naive Bayes (A)	Logistic Regression (B)
Acc	72.6%	84.5%
κ	54.1%	66.2%
F1-score	85.6%	89.1%
ROC AUC	48.4%	55.7%
⋮	⋮	⋮

Clearly, logistic regression (B) has higher scores than naive Bayes (A)!

Should we choose B over A?

Comparing Evaluation Measures

email	true	pred (A)	pred (B)
"send us your password"	+	+	+
"send us review"	−	+	−
"review your account"	−	−	+
"review us"	+	−	−
"send your password"	+	+	+
"send us your account"	+	+	−
⋮			

	Naive Bayes (A)	Logistic Regression (B)
Acc	72.6%	84.5%
κ	54.1%	66.2%
F1-score	85.6%	89.1%
ROC AUC	48.4%	55.7%
⋮	⋮	⋮

Clearly, logistic regression (B) has higher scores than naive Bayes (A)!

Should we choose B over A? maybe?

Comparing Point Estimates

$$\mathcal{D} = \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}\} \qquad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset$$

Comparing Point Estimates

$$\mathcal{D} = \{\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}\} \qquad \mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset$$

	Naive Bayes (A)		Logistic Regression (B)
Acc	72.6%	<	84.5%
κ	54.1%	<	66.2%
F1-score	85.6%	<	89.1%
ROC AUC	48.4%	<	55.7%
⋮	⋮		⋮

Comparing Point Estimates

$$\mathcal{D} = \{\mathcal{D}'_{\text{train}}, \mathcal{D}'_{\text{test}}\} \qquad \mathcal{D}'_{\text{train}} \cap \mathcal{D}'_{\text{test}} = \emptyset$$

Comparing Point Estimates

$$\mathcal{D} = \{\mathcal{D}'_{\text{train}}, \mathcal{D}'_{\text{test}}\} \qquad \mathcal{D}'_{\text{train}} \cap \mathcal{D}'_{\text{test}} = \emptyset$$

	Naive Bayes (A)		Logistic Regression (B)
Acc	79.3%	$>$	78.1%
κ	61.9%	$>$	60.3%
F1-score	86.1%	$>$	82.4%
ROC AUC	50.1%	$<$	50.4%
⋮	⋮		⋮

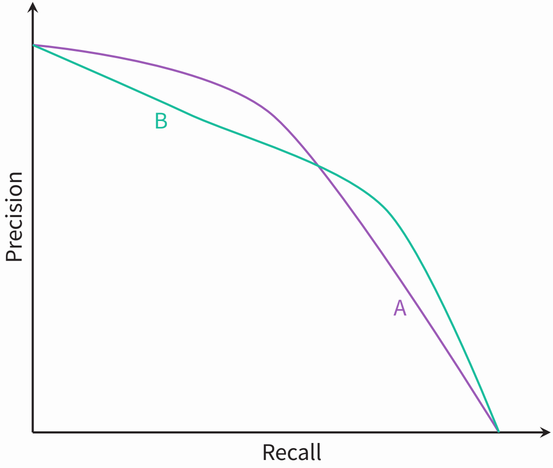
Comparing Point Estimates

$$\mathcal{D} = \{\mathcal{D}'_{\text{train}}, \mathcal{D}'_{\text{test}}\} \qquad \mathcal{D}'_{\text{train}} \cap \mathcal{D}'_{\text{test}} = \emptyset$$

	Naive Bayes (A)		Logistic Regression (B)
Acc	79.3%	>	78.1%
κ	61.9%	>	60.3%
F1-score	86.1%	>	82.4%
ROC AUC	50.1%	<	50.4%
⋮	⋮		⋮

Point estimates can be susceptible to many kinds of random effects!

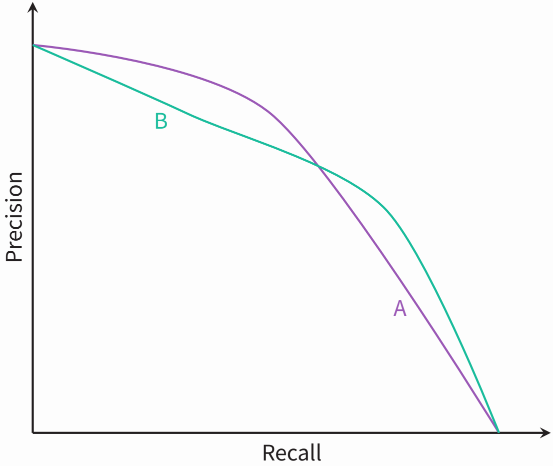
Comparison with Tradeoff



AUC of Precision-Recall

- Which model is better?

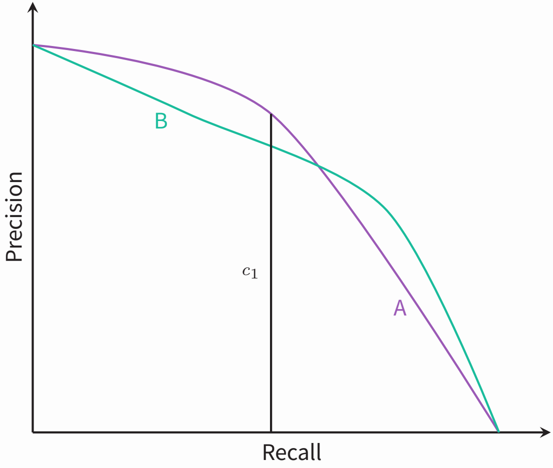
Comparison with Tradeoff



AUC of Precision-Recall

- Which model is better?
- Choice can depend on trade-off

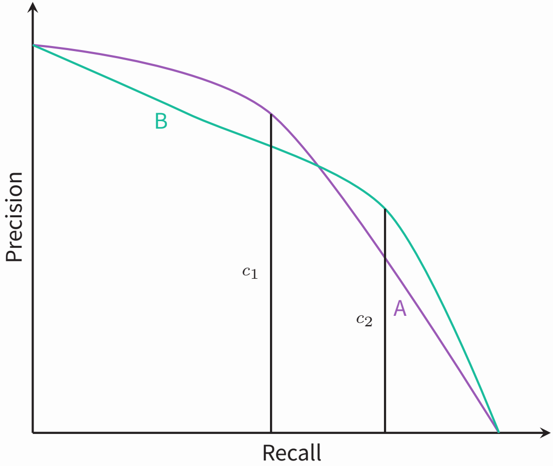
Comparison with Tradeoff



AUC of Precision-Recall

- Which model is better?
- Choice can depend on trade-off
  - lower recall, higher precision (c_1): $A > B$

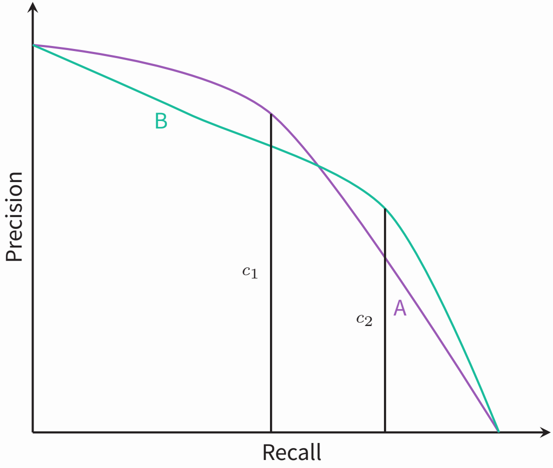
Comparison with Tradeoff



AUC of Precision-Recall

- Which model is better?
- Choice can depend on trade-off
  - lower recall, higher precision (c_1): $A > B$
  - lower precision, higher recall (c_2): $B > A$

Comparison with Tradeoff



AUC of Precision-Recall

- Which model is better?
- Choice can depend on trade-off
  - lower recall, higher precision (c_1): $A > B$
  - lower precision, higher recall (c_2): $B > A$
- Random effects (e.g. data split) can make comparison hard

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

$$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}, \{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}, \ldots, \{\mathcal{D}^K_{\text{train}}, \mathcal{D}^K_{\text{test}}\}$$

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

$$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}, \{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}, \ldots, \{\mathcal{D}^K_{\text{train}}, \mathcal{D}^K_{\text{test}}\}$$

A> B   A> B   …   B> A

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

$$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}, \{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}, \ldots, \{\mathcal{D}^K_{\text{train}}, \mathcal{D}^K_{\text{test}}\}$$

A> B A> B ... B> A

- Model (e.g. stochastic linear regression)

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

$$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}, \{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}, \ldots, \{\mathcal{D}^K_{\text{train}}, \mathcal{D}^K_{\text{test}}\}$$

A> B A> B … B> A

- Model (e.g. stochastic linear regression)

$$y_i = w_0 + w_1 x_i + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

Embracing Uncertainty

Variation in error

- Dataset partitioning (e.g. cross validation)

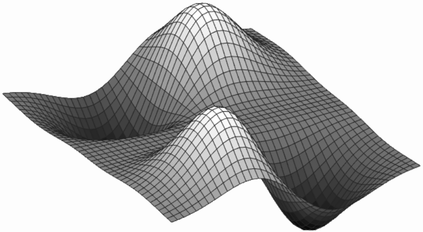
$$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}, \{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}, \ldots, \{\mathcal{D}^K_{\text{train}}, \mathcal{D}^K_{\text{test}}\}$$

A> B   A> B   …   B> A

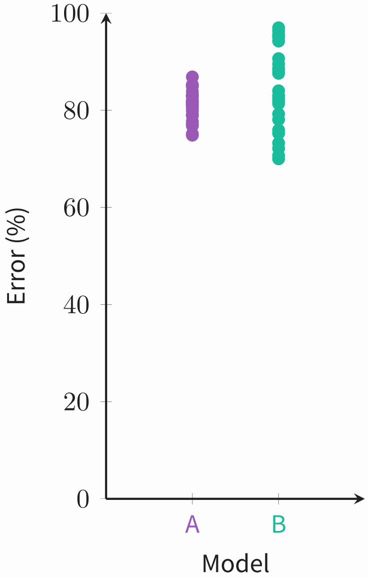
- Model (e.g. stochastic linear regression)

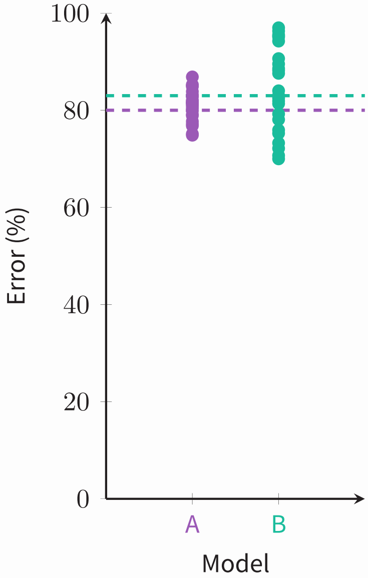
$$y_i = w_0 + w_1 x_i + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

- Learning algorithm (e.g. SGD)
  - initialisation effects
  - local minima



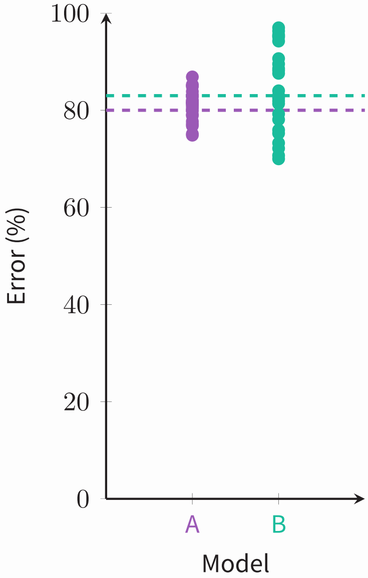
Comparing Distributions





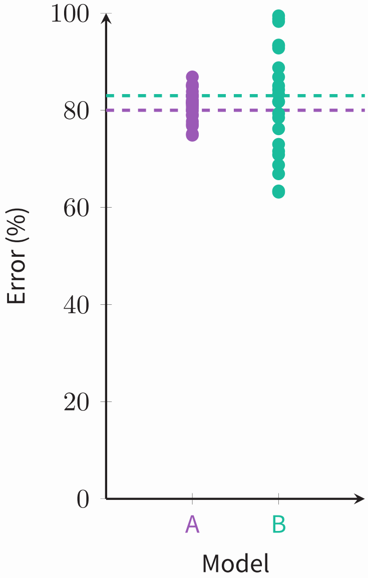
Comparing Distributions

- Compute the difference in *mean* error



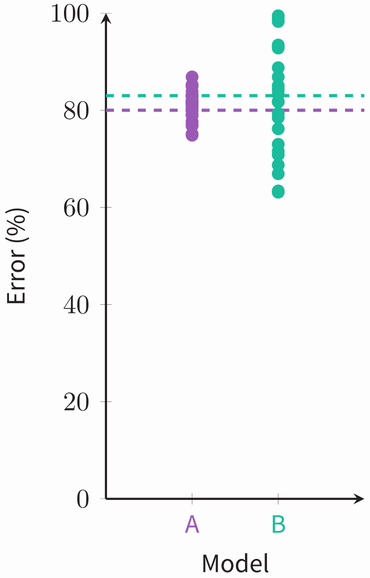
Comparing Distributions

- Compute the difference in *mean* error
  - what difference is enough to decide B> A?



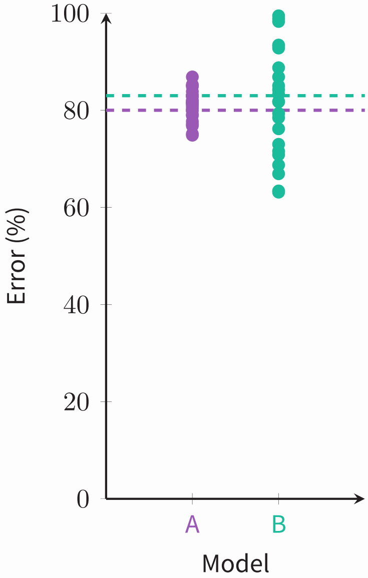
Comparing Distributions

- Compute the difference in *mean* error
  - what difference is enough to decide B> A?
  - does the spread / variance affect this choice?



Comparing Distributions

- Compute the difference in *mean* error
  - what difference is enough to decide B> A?
  - does the spread / variance affect this choice?
- Difficult to provide a general approach to say one model is “better” than another



Comparing Distributions

- Compute the difference in *mean* error
  - what difference is enough to decide B> A?
  - does the spread / variance affect this choice?
- Difficult to provide a general approach to say one model is "better" than another
- Weaker, but feasible, approach:

  How likely is it that the observed disparities are due to chance?

Statistical Tests

Preliminaries

Population vs. Sample statistics

Preliminaries

Population vs. Sample statistics

Population: All the elements from a set

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

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 $x_1, \ldots, x_N$

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

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

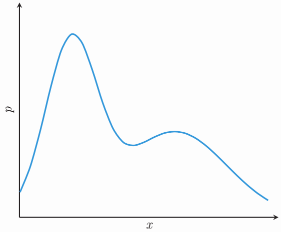
*Bessel's correction

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

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

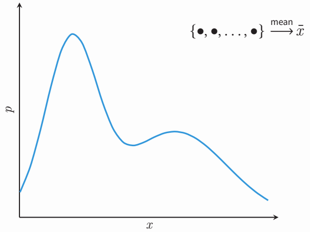


Preliminaries

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

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

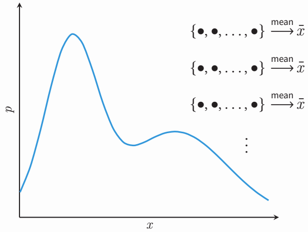


Preliminaries

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

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

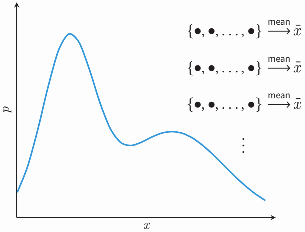


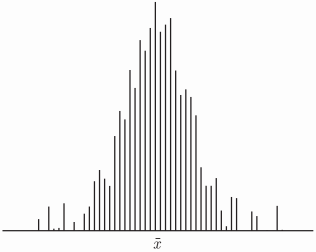
Preliminaries

Central Limit Theorem (CLT)

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

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



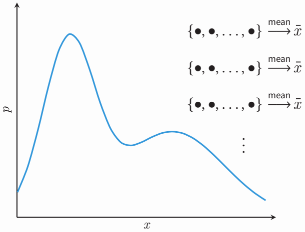


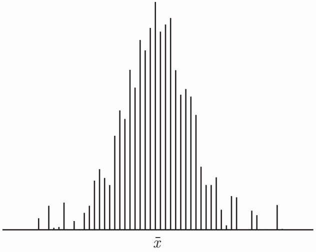
Preliminaries

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

Preliminaries

Student's-t distribution

- CLT: (weak) convergence to $\mathcal{N}(0, 1)$ as $N \rightarrow \infty$

Preliminaries

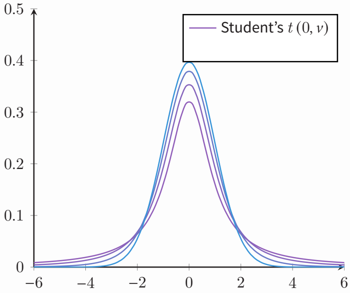
Student's-t distribution

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

Preliminaries

Student's-t distribution

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

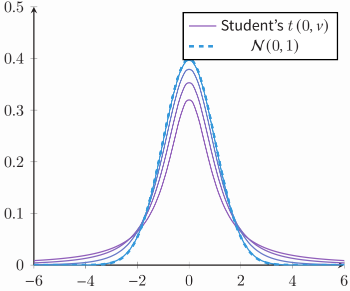


$$f(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Preliminaries

Student's-t distribution

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



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

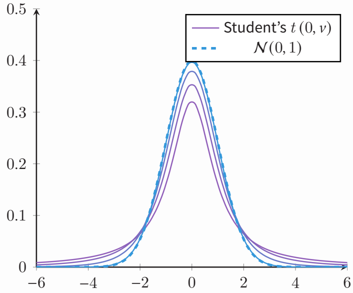
Preliminaries

Student's-t distribution

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

Assume

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



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

Preliminaries

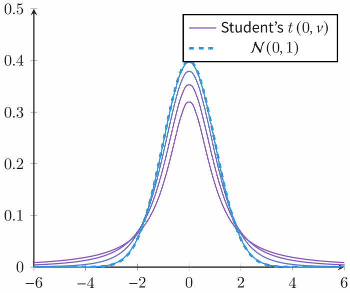
Student's-t distribution

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

Assume

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

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



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

Statistical Testing: A Sketch

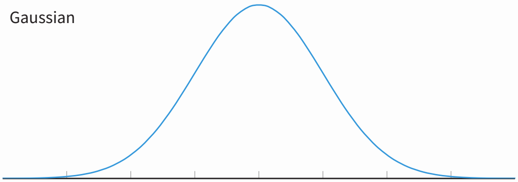
- Examine the *mean* of a set of samples

  e.g. difference in classification errors

Statistical Testing: A Sketch

- Examine the *mean* of a set of samples

  e.g. difference in classification errors
- Why? — tendency towards Gaussian

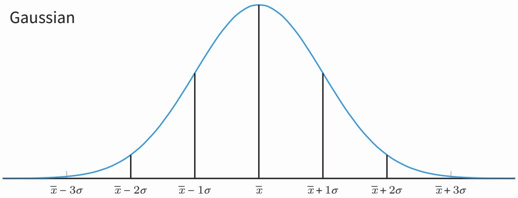


Statistical Testing: A Sketch

- Examine the *mean* of a set of samples

  e.g. difference in classification errors
- Why? — tendency towards Gaussian
- For some assumptions about the *population*: mean, variance (?)

  How likely is this observed sample mean value to have arisen by chance?

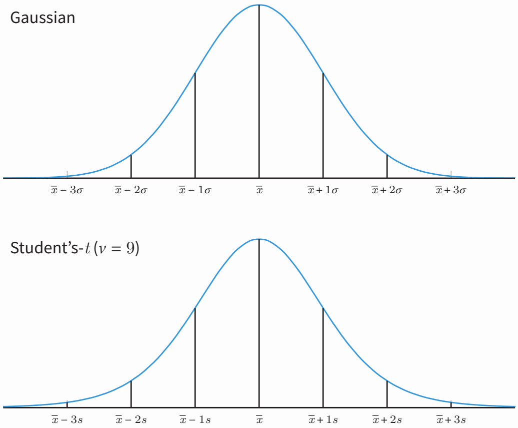


Statistical Testing: A Sketch

- Examine the *mean* of a set of samples

  e.g. difference in classification errors
- Why? — tendency towards Gaussian
- For some assumptions about the *population*: mean, variance (?)

  How likely is this observed sample mean value to have arisen by chance?



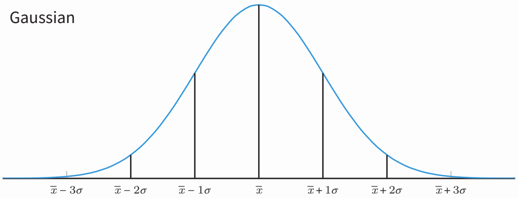
Statistical Testing: A Sketch

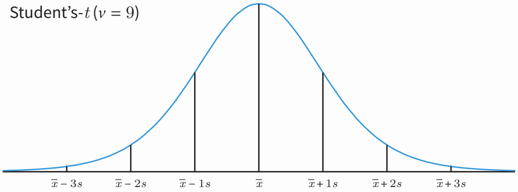
- Examine the *mean* of a set of samples

  e.g. difference in classification errors
- Why? — tendency towards Gaussian
- For some assumptions about the *population*: mean, variance (?)

  How likely is this observed sample mean value to have arisen by chance?

A common framework to evaluate chance occurrence.





Statistical Tests

Hypothesis Testing

Hypothesis Testing

Hypothesis Testing

- Formally examine two opposing conjectures (hypothesis): H_0 and H_1

Hypothesis Testing

- Formally examine two opposing conjectures (hypothesis): H_0 and H_1

Null Hypothesis: H_0

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

Hypothesis Testing

- Formally examine two opposing conjectures (hypothesis): H_0 and H_1

Null Hypothesis: H_0

- States the assumption to be tested
- Begin with assumption that $H_0 =$ `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 ($\neq, >, <$)

Hypothesis Testing

- Formally examine two opposing conjectures (hypothesis): H_0 and H_1
- Mutually exclusive and exhaustive: $H_0 = \texttt{True} \implies H_1 = \texttt{False}$

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 $(\neq, >, <)$

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`

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 ($\neq, >, <$)

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 $(\neq, >, <)$

Hypothesis Testing: Variants

- Test type
  - z-test: Gaussian distribution
  - t-test: Student's t distribution

Hypothesis Testing: Variants

- Test type

  z-test: Gaussian distribution

  t-test: Student's t distribution

- One or Two sided

  One: $H_0 : \mu^A - \mu^B \leq 0 \quad H_1 : \mu^A - \mu^B > 0$ (directional)

  Two: $H_0 : \mu^A - \mu^B = 0 \quad H_1 : \mu^A - \mu^B \neq 0$ (not directional)

Hypothesis Testing: Variants

- Test type

  z-test: Gaussian distribution

  t-test: Student's t distribution

- One or Two sided

  One: $H_0 : \mu^A - \mu^B \leq 0 \quad H_1 : \mu^A - \mu^B > 0$ (directional)

  Two: $H_0 : \mu^A - \mu^B = 0 \quad H_1 : \mu^A - \mu^B \neq 0$ (not directional)

- Test Statistic

  One-Sample: compare sample to population with known characteristics

  Two-Sample: compare two samples; typically experiment vs. control (e.g. vaccines)

  Paired: one-sample test on *difference* between samples

Hypothesis Testing: Variants

- Test type

  z-test: Gaussian distribution

  t-test: Student's t distribution

- One or Two sided

  One: $H_0 : \mu^A - \mu^B \leq 0 \quad H_1 : \mu^A - \mu^B > 0$ (directional)

  Two: $H_0 : \mu^A - \mu^B = 0 \quad H_1 : \mu^A - \mu^B \neq 0$ (not directional)

- Test Statistic

  One-Sample: compare sample to population with known characteristics

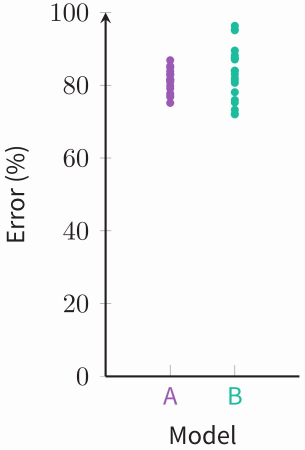
  Two-Sample: compare two samples; typically experiment vs. control (e.g. vaccines)

  Paired: one-sample test on *difference* between samples

Example: Hypothesis Testing for Models

Generating Variation

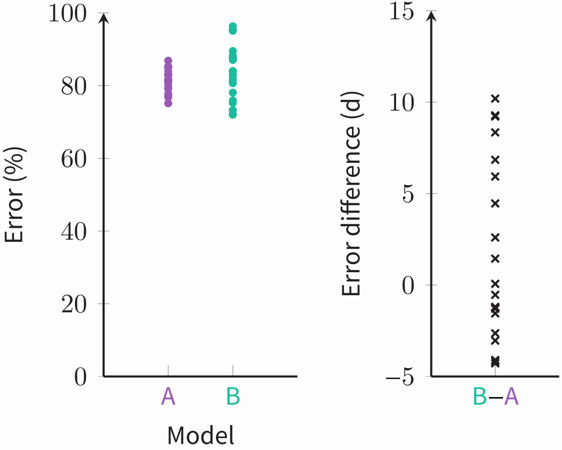
Data Split	A	B
$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}$	ℓ^A_1	ℓ^B_1
$\{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}$	ℓ^A_2	ℓ^B_2
$\vdots$	$\vdots$	$\vdots$
$\{\mathcal{D}^N_{\text{train}}, \mathcal{D}^N_{\text{test}}\}$	ℓ^A_N	ℓ^B_N



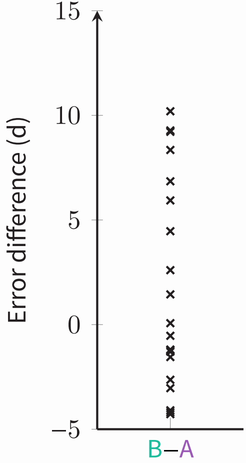
Example: Hypothesis Testing for Models

Generating Variation

Data Split	A	B	d
$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}$	ℓ^A_1	ℓ^B_1	$\ell^B_1 - \ell^A_1$
$\{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}$	ℓ^A_2	ℓ^B_2	$\ell^B_2 - \ell^A_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\{\mathcal{D}^N_{\text{train}}, \mathcal{D}^N_{\text{test}}\}$	ℓ^A_N	ℓ^B_N	$\ell^B_N - \ell^A_N$



Example: Hypothesis Testing for Models

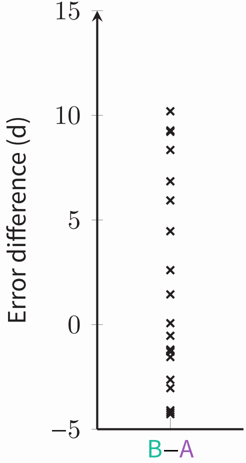


Hypotheses

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

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

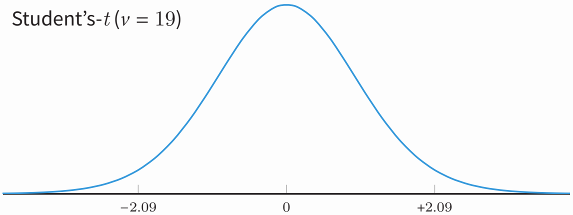
Example: Hypothesis Testing for Models



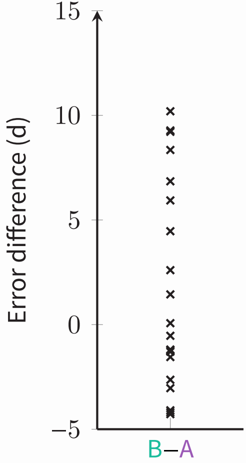
Hypotheses

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

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



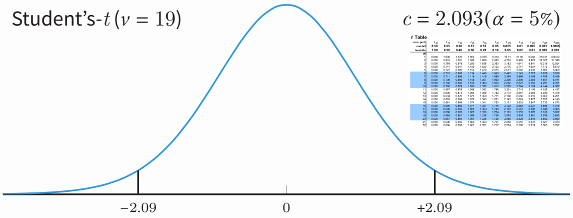
Example: Hypothesis Testing for Models



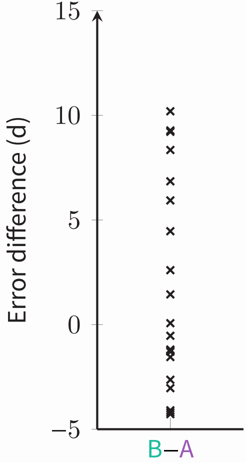
Hypotheses

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

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



Example: Hypothesis Testing for Models



d
−0.61
6.44
−4.83
7.43
−3.04
−4.75
6.47
9.62
9.70
2.51
−3.18
8.23
10.24
−4.59
6.41
2.08
−2.49
2.72
−0.24
2.56

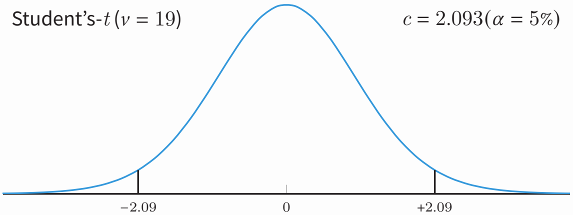
Hypotheses

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

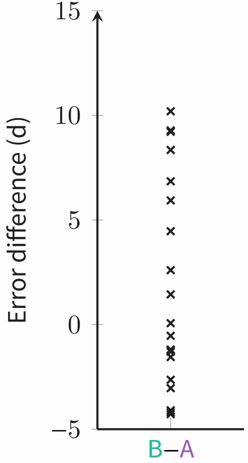
$H_1 : \mu^d \neq 0$ $\qquad N = 20$

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i = 2.53 \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2 = 27.78 \qquad t = \frac{\bar{d} - 0}{s/\sqrt{N}} = 2.14$$

Student's-t $(\nu = 19)$ $\qquad c = 2.093 (\alpha = 5\%)$



Example: Hypothesis Testing for Models



d
−0.61
6.44
−4.83
7.43
−3.04
−4.75
6.47
9.62
9.70
2.51
−3.18
8.23
10.24
−4.59
6.41
2.08
−2.49
2.72
−0.24
2.56

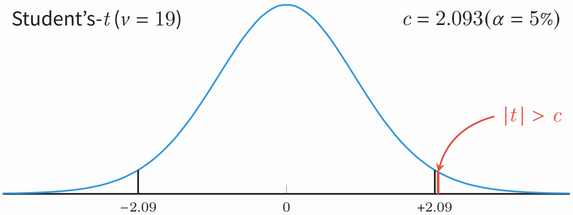
Hypotheses

$$H_0 : \mu^d = 0 \qquad \alpha = 5\% \text{ (significance)}$$

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

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i = 2.53 \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2 = 27.78 \qquad t = \frac{\bar{d} - 0}{s/\sqrt{N}} = 2.14$$

Student's-t ($\nu = 19$) $\qquad c = 2.093(\alpha = 5\%)$



Hypothesis Testing: Caveats

- Rejecting H_0 does not imply 100% sure H_0 is `False`

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`

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

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*

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

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)

$$\bar{d} \pm c|_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

which, for our example would be

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

$$2.53 \pm 2.47$$

Cross Validation for Variation: Caveat

- Recall that CLT requires the samples to be independent

Cross Validation for Variation: Caveat

- Recall that CLT requires the samples to be independent
- Simple cross-validation can violate that independence (overlap in $\mathcal{D}_{\text{train}}$!)

Data Split	A	B	d
$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}$	ℓ^A_1	ℓ^B_1	$\ell^B_1 - \ell^A_1$
$\{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}$	ℓ^A_2	ℓ^B_2	$\ell^B_2 - \ell^A_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Cross Validation for Variation: Caveat

- Recall that CLT requires the samples to be independent
- Simple cross-validation can violate that independence (overlap in $\mathcal{D}_{\text{train}}$!)

Data Split	A	B	d
$\{\mathcal{D}^1_{\text{train}}, \mathcal{D}^1_{\text{test}}\}$	ℓ^A_1	ℓ^B_1	$\ell^B_1 - \ell^A_1$
$\{\mathcal{D}^2_{\text{train}}, \mathcal{D}^2_{\text{test}}\}$	ℓ^A_2	ℓ^B_2	$\ell^B_2 - \ell^A_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

- **Solutions:**
  - 5x2 Cross Validation [1]
  - Adjust standard deviation to account for imbalance [2]
  - …and many more (ANOVA, Non-parametric tests, etc.)!

1. T. G. Dietterich, Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, 1998
2. C. Nadeau & Y. Bengio, Inference for the Generalization Error, 2003

Summary

Key

Being able to compare models and experiments is both a science and an art!

Most important aspect is to think what sources of variability affects results, and how large their effects are likely to be.

Summary

Key

Being able to compare models and experiments is both a science and an art!

Most important aspect is to think what sources of variability affects results, and how large their effects are likely to be.

- Some measures incorporate context; use it! (P-R, ROC)
- For when statistical tests are required (not always!)
  - ensure your assumptions on the model / data are clearly stated
  - ensure assumptions of the test are met
- Performance on error measures not all—speed, use of resources, and ease of implementation can, and should, affect preference!