Lecture Notes on

Pattern Recognition

Session 07: Model Evaluation

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

0

Road map

1

Model Evaluation

- What do we care?
- The Test Set method
- The Leave One Out method
- The K-fold Cross-Validation

Thanks to Prof Andrew Moore

2

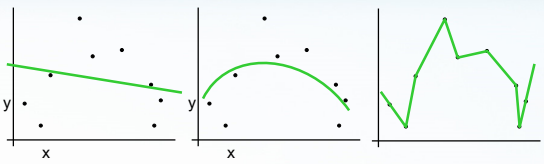
What do we care?

- Regular Performance of Algorithms
  - *Accuracy*
  - *Scope*
  - *Efficiency*
- Which ones we care?
  1. *Accuracy*
  2. *Efficiency*
  3. *Scope*

? Fitting Accuracy, or
? Prediction Accuracy

3

What kind of accuracy we want?

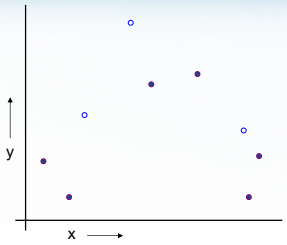


Why not choose the method with the best fit to the data?

"How well are you going to ***predict*** future data drawn from the same distribution?"

4

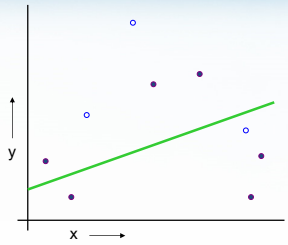
The test set method



1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set

5

The test set method

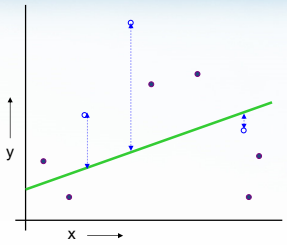


1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set

(Linear regression example)

6

The test set method

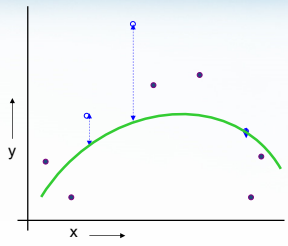


1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Linear regression example)
Mean Squared Error = 2.4

7

The test set method

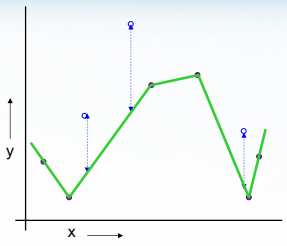


1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Quadratic regression example)
Mean Squared Error = 0.9

8

The test set method



1. Randomly choose 30% of the data to be in a test set
2. The remainder is a training set
3. Perform your regression on the training set
4. Estimate your future performance with the test set

(Join the dots example)
Mean Squared Error = 2.2

9

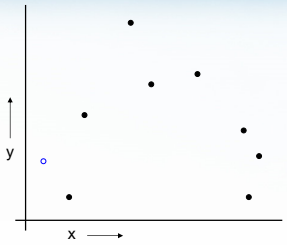
The test set method

Pros	Cons
• Very very simple • Can then simply choose the method with the best test-set score	• Waste data: we get an estimate of the best method to apply to 30% less data • If we don't have much data, our test-set might just be lucky or unlucky

We say the "test-set estimator of performance has high variance"

10

Leave-One-Out Method

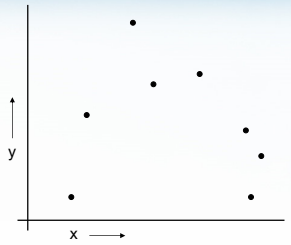


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record

11

Leave-One-Out Method

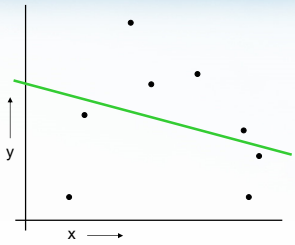


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset

12

Leave-One-Out Method

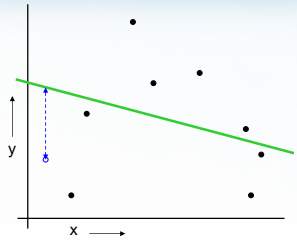


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints

13

Leave-One-Out Method

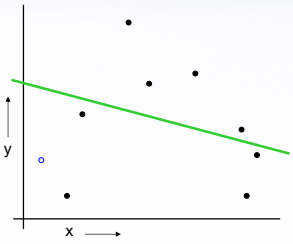


For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

14

Leave-One-Out Method



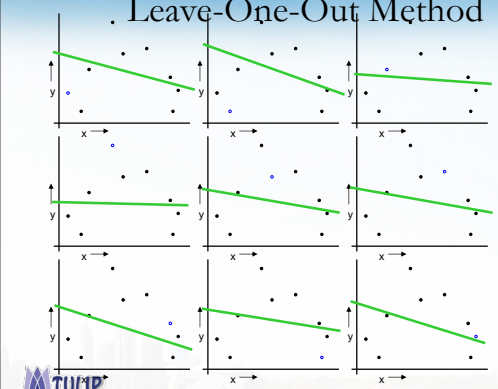
For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

15

Leave-One-Out Method



For k=1 to R

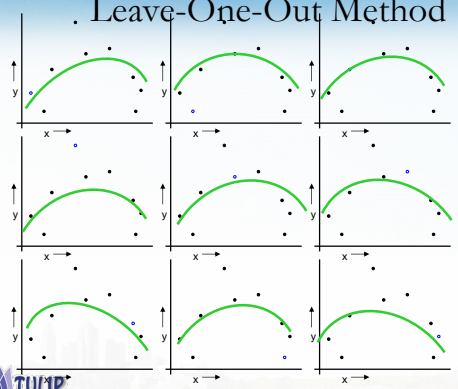
1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV} = 2.12$

16

Leave-One-Out Method



For k=1 to R

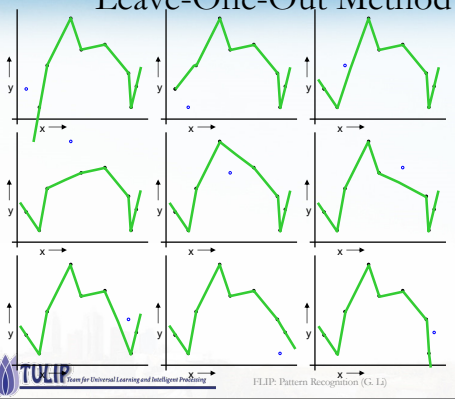
1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV}=0.962$

17

Leave-One-Out Method



For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV}=3.33$

18

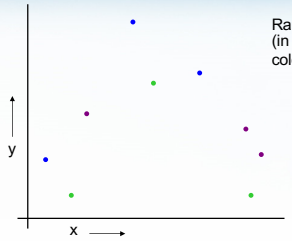
Which kind of Cross Validation?

	Cons	Pros
Test-set	Variance: unreliable estimate of future performance	Cheap
Leave-one-out	Expensive. Has some weird behavior	Doesn't waste data

..can we get the best of both worlds?

19

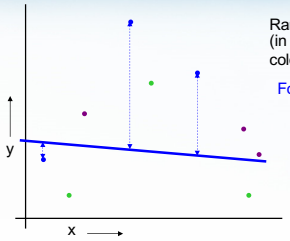
k-fold Cross Validation



Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

20

k-fold Cross Validation

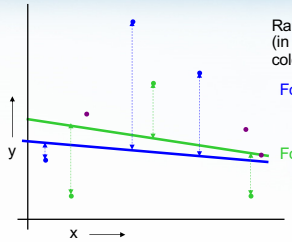


Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

21

k-fold Cross Validation



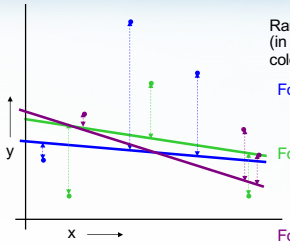
Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

22

k-fold Cross Validation



Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

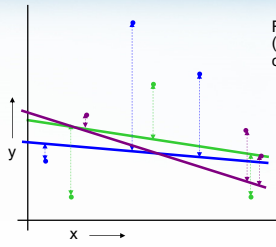
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the yellow partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

23

k-fold Cross Validation



Linear Regression $MSE_{3FOLD}=2.05$

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

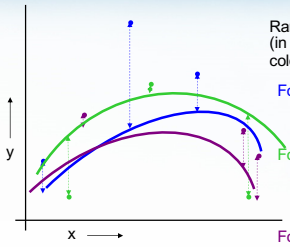
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the yellow partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

24

k-fold Cross Validation



Quadratic Regression $MSE_{3FOLD}=1.11$

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

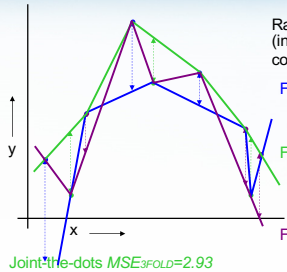
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the yellow partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

25

k-fold Cross Validation



Joint-the-dots $MSE_{3FOLD}=2.93$

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored blue Green and yellow)

For the blue partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the yellow partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

26

Which kind of Cross Validation?

	Downside	Upside
Test-set	Variance: unreliable estimate of future performance	Cheap
Leave-one-out	Expensive. Has some weird behavior	Doesn't waste data
10-fold	Wastes 10% of the data. 10 times more expensive than test set	Only wastes 10%. Only 10 times more expensive instead of R times.
3-fold	Wastier than 10-fold. Expensivier than test set	Slightly better than test-set
R-fold	Identical to Leave-one-out	

27

CV-based Model Selection

- We're trying to decide which algorithm to use.
- We train each model and make a table…

i	f_i	TRAIN-ERR	10-FOLD-CV-ERR	Choice
1	f_1			
2	f_2			
3	f_3			?
4	f_4			
5	f_5			
6	f_6			

28

Other Model Selection Methods

- Model selection methods:
  1. *Cross-validation*
  2. *Occam's Razor Type: The smaller, the better*
     - *Minimum Description Length (MDL)*
     - *Minimum Messaging Length (MML)*
     - *AIC (Akaike Information Criterion)*
     - *BIC (Bayesian Information Criterion)*
     - *VC-dimension (Vapnik-Chervonenkis Dimension)*

Only directly applicable to choosing classifiers

29

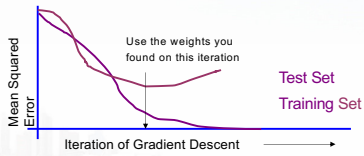
Cross-Validation for regression

- Choosing the number of hidden units in a neural net
- Feature selection
- Choosing a polynomial degree
- Choosing which regressor to use

30

Supervising Gradient Descent

- This is a weird but common use of ***Test-set validation***
  - Suppose you have a neural net with too many hidden units. It will overfit.
  - As gradient descent progresses, maintain a graph of `MSE-testset-error` vs. `Iteration`



31

Cross-validation for classification

- Instead of computing the **mean squared errors (MSE)** on a test set, you should compute various measurements …
  - *Error rate* (or its dual part *Accuracy*):
    - The total number of misclassifications on a test-set

32

Evaluation and performance

- Confusion Matrix
- Accuracy
- Model Comparison

33

Evaluation and performance

- How well is your classification algorithm?
  - Focus on the predictive capability of a model, rather than how fast it takes to classify or build models, scalability, etc.
- Confusion matrix
  - Detail classification result for each class.
- Accuracy
  - How well we can predict for each class

34

Evaluation

- Take 'Yes' as the positive class (class of interest)

True Class	Algorithm returns
Yes	Yes
Yes	Yes
Yes	No
Yes	Yes
No	No
No	No
No	Yes
No	No
No	Yes
Yes	Yes
Yes	Yes
No	No
No	No
Yes	Yes

Classified As →	A=Yes	B=No
A=Yes	6	1
B=No	2	5

		Predicted class		
		+	-	a: TP (true positive) b: FN (false negative) c: FP (false positive) d: TN (true negative)
Actual Class	+	a	b	
	-	c	d	

`sklearn.metrics.confusion_matrix`

35

Evaluation

- Most widely-used metric:

  `Accuracy = (a+d)/(a+b+c+d)`

- Consider a binary classification problem
  - # of Class "Positive" instances = 9990
  - # of Class "Negative" instances = 10
- A naïve model that predicts everything positive with accuracy 99.9%

		Predicted class	
		+	-
Actual Class	+	a	b
	-	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

36

Evaluation

- Alternative Measures

  `Precision = a/(a+c)`

  `Recall = a/(a+b)`

  `F-measure or F1 = 2a/(2a+b+c)`

- Other measures
  - ROC
  - AUC

		Predicted class	
		+	-
Actual Class	+	a	b
	-	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

37

Cost Matrix

- **`C(i,j)`**
  - Cost of misclassifying class **`i`** instance as class **`j`**

Cost matrix		Predicted class	
	C(i,j)	+	-
Actual Class	+	...	
	-	...	

		Predicted class	
		+	-
Actual Class	+	a	b
	-	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

38

Cost Matrix: example

Cost matrix		Predicted class	
	C(i,j)	+	-
Actual Class	+	-1	100
	-	1	0

Model1		Predicted class	
		+	-
Actual Class	+	180	20
	-	60	240

accuracy = 82%
Cost = 1880

Model2		Predicted class	
		+	-
Actual Class	+	140	60
	-	10	290

accuracy = 86%
Cost = 5870

39

Cost-Sensitive Classification

- Traditional Classification
  - Given a query **`t`**, we will classify it as class **`i`** if

    $$i = \arg\max_j p(j \mid t)$$

  - For Binary Classification, classify it as "positive" if

    `p(+|t) > p(-|t)`

- Cost-Sensitive Classification
  - Classify **`t`** as j

    $$i = \arg\min_j \sum_k p(k \mid t) \times C(k, j)$$

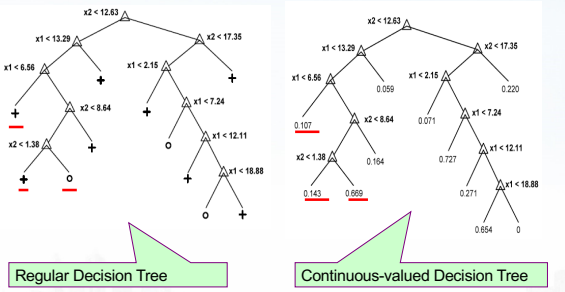
40

ROC: Receiver Operating Characteristic

- A graphical approach for displaying *trade-off* between *detection rate* and *false alarm rate*
  - Developed in 1950s for signal detection theory to analyze noisy signals
- To draw ROC curve, classifier must produce *continuous-valued output*
  - Outputs are used to *rank* test records, *from* the most likely record to be classified as a positive class *to* the most likely record to be classified as a negative class
  - *Performance of each classifier* is represented as *a point* on the ROC curve

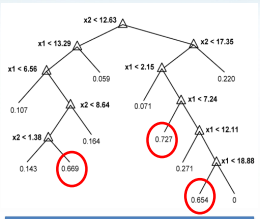
41

Continuous-valued outputs from a decision tree

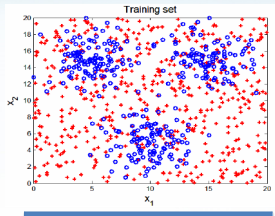


42

ROC Curve Example



a=0.3		Predicted class	
		o	+
Actual Class	o	645	209
	+	298	948

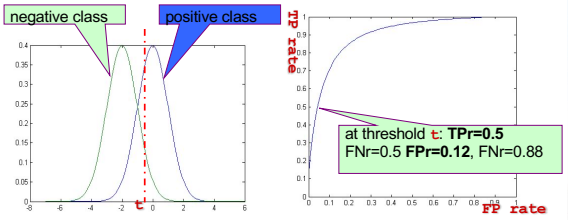


a=0.7		Predicted class	
		o	+
Actual Class	o	181	673
	+	78	1168

43

ROC Curve Example

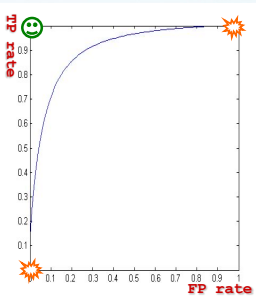
- 1-dimensional data set containing 2 classes
  - *Positive* vs. *negative*
  - Any points located at `x>t` is classified as positive



44

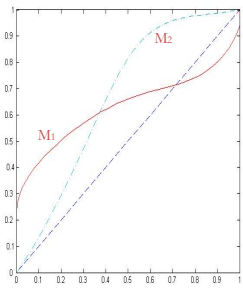
ROC Curve

- `(TPr, FPr)`
  - `TPR = TP/(TP+FN) = RECALL`
  - `FPR = FP/(FP+TN)`
  - `(0,0)`
    - Declare everything to be negative
  - `(1,1)`
    - Declare everything to be positive class
  - `(1,0)`
    - Ideal
- Diagonal line:
  - Random guessing
- Below diagonal line
  - Prediction is opposite of the true class



45

Using ROC for Model Comparison



- No Model consistently outperforms the other
  - M_1 is better for small FP Rate
  - M_2 is better for large FP Rate
- Area under the ROC curve
  - Ideal:
    - `Area = 1`
  - Random guess
    - `Area = 0.5`

46

How to Construct an ROC Curve

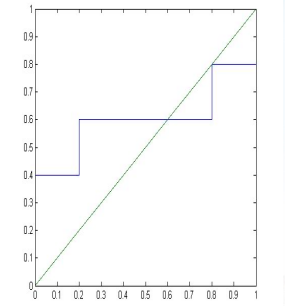
Instance	P(+|A)	True
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use classifier that produces continuous valued output for each test instance `P(+|A)`
- Sort the instances according to `P(+|A)` in decreasing order
- Apply threshold at each unique value of `P(+|A)`
- Count the number of `TP, FP, TN, FN` at each threshold
- `TPR = TP/(TP+FN)`
- `FPR = FP/(FP+TN)`

47

How to Construct an ROC Curve

Class	Thrd	TP	FP	TN	FN	TPR	FPR
+	0.25	5	5	0	0	1	1
-	0.43	4	5	0	1	0.8	1
+	0.53	4	4	1	1	0.8	0.8
-	0.76	3	4	1	2	0.6	0.8
-	0.85	3	3	2	2	0.6	0.6
-	0.85	3	2	3	2	0.6	0.4
+	0.85	3	1	4	2	0.6	0.2
-	0.87	2	1	4	3	0.4	0.2
+	0.93	2	0	5	3	0.4	0
+	0.95	1	0	5	4	0.2	0
	1.00	0	0	5	5	0	0



48

Model Comparison

- Task 1
  - If we have two models M_1 and M_2 tested on ***different data sets***, how can we tell the relative performance of these two models?
- Task 2
  - If we have two models M_1 and M_2 tested on ***same data sets***, how can we tell the relative performance of these two models?

49

Model Comparison

- Given two models:

  M_1: accuracy = 85%, tested on 30 instances

  M_2: accuracy = 75%, tested on 5000 instances

- Can we say M_1 is better than M_2?
  - How much ***confidence*** can we place on accuracy of M_1 and M_2?
  - Can the ***difference*** in performance measure be explained as a result of ***random fluctuations*** in the test set?

50

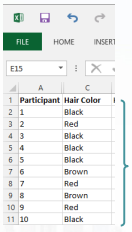
Confidence Interval

- Measure of Locations
- Measure of Dispersion
- Confidence Interval
- Population Parameter Comparison

51

Measure of location (Center Tendency)

- The **mode**
  - Label or value with the highest frequency



Participant	Hair Color
1	Black
2	Red
3	Black
4	Black
5	Black
6	Brown
7	Red
8	Brown
9	Red
10	Black

Hair Color	Frequency	Normalized Frequency (%)
Black	5	50
Red	3	30
Brown	2	20

52

Measure of location (Center Tendency)

- The **median**
  - If we order the data, the median is the middle value.

Sex	Seatbelt
Male	Rarely
Female	Mostimes
Female	Always
Male	Always
Female	Always
Female	Always
Female	Sometimes
Female	Always
Female	Mostimes
Female	Rarely
Female	Mostimes
Female	Always
Female	Mostimes
Male	Always
Male	Mostimes

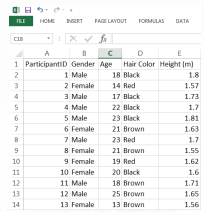
Seatbelt	Frequency
Never	0
Rarely	2
Sometimes	1
Mostimes	5
Always	7

R, R, S, M, M, M, M, M, A, A, A, A, A, A

53

Measure of location (Center Tendency)

- The **mean**
  - $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, all data used and sensitive to outliers



$$\frac{18+14+17+22+23+21+23+21+19+20+18+25+13}{13} = 19.5$$

54

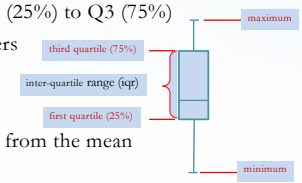
Measure of dispersion (Spread)

- The second type of measure is the measure of spread or dispersion.
  - This tells us how much the data spread out.
  - Equivalently, the degree of variation in your data.

55

Measure of dispersion (Spread)

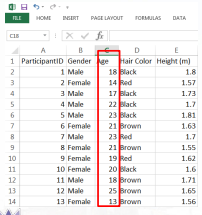
- The ***range***
  - The distance from minimum to maximum.
- The ***interquartile range*** (iqr)
  - The distance from Q1 (25%) to Q3 (75%)
  - Less sensitive to outliers
- ***Standard deviation***
  - Use all data
  - Measure the deviation from the mean



56

Standard deviation

- Let the data be $x_1, x_2, \ldots, x_n$
  - Standard deviation (std): $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$



$$\mu = \frac{18+14+17+22+23+21+23+21+19+20+18+25+13}{13} = 19.5$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \approx 3.39$$

57

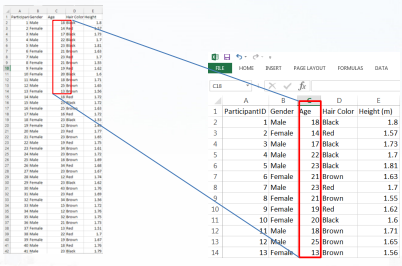
Population Parameter and Sample Statistic

- **Population**:
  - The entire collection of measurements that we would have if we could measure the whole population.
  - **Population parameter**: a fixed statistics associated with a population.
    - In the context of confidence interval, it is unknown and to be estimated (e.g., the mean).
- **Sample**:
  - the set of measurements that we obtain.
  - **Sample size**: denoted by n, is the number of measurements in the sample.
  - **Sample statistic**: a summary number computed from the sample (e.g., the mean)

58

Confidence interval

- Dataset is in fact just a sample from a true population.

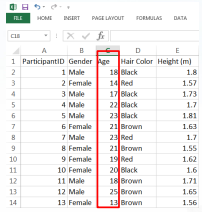


…. [go on forever]

59

Confidence interval

- Dataset is in fact just a sample from a true population.
  - We estimate the mean from this sample, $\mu \approx 19.5$.
  - How close is this sample mean to the true population mean?
  - Confidence interval quantifies this question.

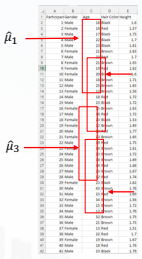


…. [go on forever]

60

Confidence interval

- Dataset is in fact just a sample from a true population.



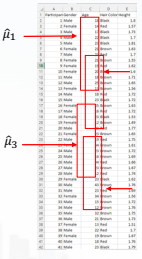
- If we repeatedly extract different junks of data (sample), each time we will obtain a different sample mean.
- These are amazing results:
  - The mean of all possible sample mean is the same the as population mean.
  - Sample means are normally distributed with the mean is the population mean.
- Therefore if we know the standard deviation from the collection of the samples, we can tell with a certain confidence about the true mean.

…. [go on forever]

61

Confidence interval

- Dataset is in fact just a sample from a true population.



- Central Limit Theorem

Let $X_1, X_2, \cdots$ be a sequence of *independent identically distributed* (i.i.d.) random variables with common mean μ and variance σ^2, define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}, \text{ which with } E(Z_n) = 0 \text{ and } var(Z_n) = 1$$

Then the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

In the sense that $\lim_{n\to\infty} P(Z_n \le z) = \Phi(z)$, for every z.

Intuition. Look at three variants of the sum of random variables

- $S_n = X_1 + \cdots + X_n$, with variance $n\sigma^2$
- $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$, with variance σ^2/n. *Converges in probability to E(X)* (WLLN)
- $\frac{S_n}{n}$ with constant variance σ^2.

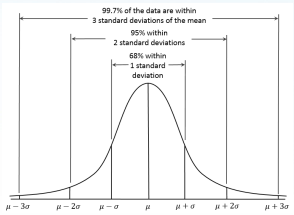
- **Standard Error** (of the mean)
  - If a variable has sample standard deviation σ then the standard error of the mean of a sample of size n is $\frac{\sigma}{\sqrt{n}}$

…. [go on forever]

62

Normal distribution

$$\frac{1}{2\pi\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



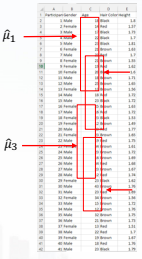
- 95% of area (probability) lies in $\mu \pm 1.96\sigma$
- N% of area (probability) lies in $\mu \pm z_N\sigma$

N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

63

Confidence interval

- Dataset is in fact just a sample from a true population.



- How can we do this?
- In practice, we only have one sample
- **Standard error** of the mean:

$$s.e.(X) = s/\sqrt{n}$$

sample size; sample standard deviation

With 95% confidence:
Population Mean
= Sample Mean ± 1.96 x Standard Error

…. [go on forever]

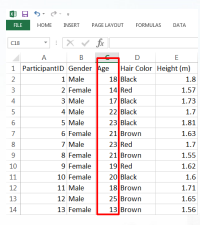
64

Confidence interval

- Dataset is in fact just a sample from a true population.
  - We estimate the mean from this sample, $\mu \approx 19.5$.
  - Standard error of the mean:

$$s.e.(X) = \frac{3.39}{\sqrt{13}} = 0.94$$

With 95% confidence:
population mean
= 19.5 ± 1.96 x 0.94 = 19.5 ± 1.84



65

Model Comparison

- Given two models:
  - M_1: accuracy = 85%, tested on 30 instances
  - M_2: accuracy = 75%, tested on 5000 instances
- Which one is better?
  - Calculate the confidence interval for each of them
    - M_1: $[0.85-1.96se_1, 0.85+1.96se_1]$
    - M_2: $[0.75-1.96se_2, 0.75+1.96se_2]$
    - where se_1 and se_2 are calculated based on their std.

66

Prediction Accuracy as Bernoulli Trials

- Prediction can be regarded as a ***Bernoulli trial***
  - A ***Bernoulli trial*** has
    - 2 possible outcomes (***Correct*** or ***Wrong***)
    - The results from different trials are independent
  - **Binomial Probability Model** for ***Bernoulli Trials***
    - A ***fixed*** number of **Bernoulli trials** are performed, it is often written as **Binom(N, p)**
    - The number of "success" out of n trials, follows the Binomial distribution, i.e., **X~Binom(N, p)**

69

Prediction Accuracy as Bernoulli Trials

Binomial	ML Accuracy Evaluation
Success rate P	P: The accuracy
Repeat time N	N: # Sample Size
#Success X	X: # Correct Classification
X~ Binom(N,P)	X ~ Binom(N,P)

70

Prediction Accuracy as Bernoulli Trials

- Estimate the accuracy of classifier
  - Given X (# of correct prediction), and N (sample size), find the upper and lower bounds of P (true accuracy of the classifier)
  - The accuracy X/N is only an approximate of the unknown P
  - How can we narrow down the range of possible P?
- Solution
  - For Binomial distribution, the variance is $p(1-p)$, hence the standard error of the mean is $\sqrt{p(1-p)/n}$

71

Model Comparison

- Given two models:
  - M_1: accuracy = 85%, tested on 30 instances
  - M_2: accuracy = 75%, tested on 5000 instances

$$p \pm 1.96\times\sqrt{p(1-p)/n}$$

- Which one is better?
  - Calculate the confidence interval for each of them
    - M_1: [0.72, 0.98]
    - M_2: [0.74, 0.76]

72

Confidence interval

- Summary
  - The mean measures the central tendency is extremely useful statistics.
  - However, in practice we only obtain a subset of data from the true population.
  - Confidence interval allows us to make precise statement about the true population mean from the sample.
  - With 95% confidence: $p \pm 1.96\times\sqrt{p(1-p)/n}$

    Population Mean = Sample Mean $\pm$ 1.96 x Standard Error
    - where standard error: $s.e.(X) = s/\sqrt{n}$, with s is the sample standard deviation and n is the number of data points.

73

Population Parameter Comparison

- Compare the difference
- Compare on the same set

74

Comparing population parameters

- With confidence interval we can estimate a *single* population parameter from the sample.
- In data science and statistical science, a common setting is the comparison of two population parameters:
  - E.g., mean heights of men and women are the same, are the birth weights for children born in hospital and at home the same, etc.

75

Comparing population parameters

- Are the birth weights for children born in hospital and at home the same?
  - How do apply the confidence interval for a single population to estimate the difference?
  - If 95% confidence interval for the difference between two population parameters include zero, the 95% confidence that there is no difference in the two parameter values. Otherwise, 95% confidence that there is a *statistically significant* difference in the mean.

Infant	Birthweight (g)	
	Hospital	Home
1	3710	3810
2	3650	3865
3	4490	4578
4	3421	3522
5	3399	3400
6	4094	4156
7	4006	4200
8	3287	3265
9	3594	3599
10	4206	4215
11	3508	3697
12	4010	4209
13	3896	3911
14	3800	3943
15	2860	3000
16	3798	3802
17	3666	3654
18	4200	4289
19	3615	3732
20	3193	3098
21	2994	3105
22	3266	3455
23	3400	3507
24	4090	4103
25	3303	3456
26	3447	3538
27	3388	3400
28	3613	3715
29	3541	3566
30	3886	4000

76

Comparing population parameters

- Are the birth weights for children born in hospital and at home the same?

Mean difference: −82.1667
Std. error difference: 98.4359
95% confidence interval: [−114.8742, 279.2076]

  - This interval traps **0**, hence 95% we are confident that the two means are **not** different.!

Infant	Birthweight (g)	
	Hospital	Home
1	3710	3810
2	3650	3865
3	4490	4578
4	3421	3522
5	3399	3400
6	4094	4156
7	4006	4200
8	3287	3265
9	3594	3599
10	4206	4215
11	3508	3697
12	4010	4209
13	3896	3911
14	3800	3943
15	2860	3000
16	3798	3802
17	3666	3654
18	4200	4289
19	3615	3732
20	3193	3098
21	2994	3105
22	3266	3455
23	3400	3507
24	4090	4103
25	3303	3456
26	3447	3538
27	3388	3400
28	3613	3715
29	3541	3566
30	3886	4000

77

Confidence Interval for Accuracy Comparison

- Given two models,
  - M_1: acc = e_1, tested on N_1 instances data set D_1
  - M_2: acc = e_2, tested on N_2 instances data set D_2
  - Assume D_1 and D_2 are independent

- If N_1 and N_2 are sufficiently large, then

$$e_1 \sim N(\mu_1, \delta_1^2)$$
$$e_2 \sim N(\mu_2, \delta_2^2)$$

With $\delta_i^2 = {e_i(1-e_i)}/{N_i}$

78

Confidence Interval for Accuracy Comparison

- Given two models,
  - M_1: acc = e_1, tested on N_1 instances data set D_1
  - M_2: acc = e_2, tested on N_2 instances data set D_2
  - Assume D_1 and D_2 are independent

- The distribution of $d = |e_1 - e_2|$ is Gaussian

$$d \sim N(\mu, \frac{e_1(1-e_1)}{N_1} + \frac{e_2(1-e_2)}{N_2})$$

79

Confidence Interval for Accuracy Comparison

- Given two models:
  - M_1: accuracy = 85%, tested on 30 instances
  - M_2: accuracy = 75%, tested on 5000 instances
- We have $\mu = 0.1$ (2-sided test)

$$\text{and } \delta^2 = \frac{0.15\times0.85}{30} + \frac{0.25\times0.75}{5000} = 0.0043$$

- At 95% confidence level, $d \in [-0.028, 0.228]$

Interval contains 0, so the difference is not statistically significant.

80

Comparing on the same sets

- Paired performance comparison
  - If two compared models are applied to the same test sets $D_1, D_2, \ldots, D_k$ (e.g., via K-fold cross validation)
  - Each learning algorithm may produce K models
    - L_1 may produce $M_{11}, M_{12}, \ldots, M_{1k}$
    - L_2 may produce $M_{21}, M_{22}, \ldots, M_{2k}$
  - For each set, compute $d_j = e_{1j} - e_{2j}$
    - d_j has mean d_t and variance σ_t
    - Estimate the confidence interval

$$\hat{\sigma}_t^2 = \frac{\sum_{j=1}^{k}(d_j - \bar{d})}{k(k-1)}$$

$$d_t = d \pm t_{1-\alpha,k-1}\hat{\sigma}_t$$

81

Comparing on the same sets

- 30-fold cross validation
  - Average difference = 0.05
  - Standard deviation of difference = 0.002
- At 95% confidence level, t =2.04

d_j
$= 0.05 \pm 2.04*0.002$
$= 0.05 \pm 0.00408$

Interval does not span the value 0, so the difference is statistically significant.

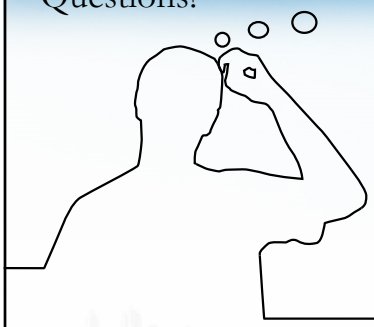
82

Are two population parameters different?

- What we learned was a two-sample t-test
  - Data from both groups are metric.
  - Distributions for each group must be reasonably normal.
  - Two standard deviations should be approximately the same.
- Many hypothesis testing exists:
  - Two-sample t-test
  - Matched-pairs t-test
  - Mann-Whitney test
  - Kruskal-Wallis Test
  - Wilcoxon test
  - Chi-squared test
  - McNemar's test, ….
  - It's best to have a reference!!!, e.g., see

83

Questions?



84