# Mid-term Exam

- Time: March 1 (Thursday) 2:00 pm-3:15 pm
- Location: 117 Debartolo
- Let me know early if you will be off campus: Coordinate with TA to find a date and TA will proctor you.

- Write down your answers/solutions on the blue book that we will give you.
- Return your exam paper after the exam.
- You can have a double-sided letter-size reference paper.

- You must bring a pen/pencil/writing tool.
- You had better bring a calculator.
- You are not allowed to use laptop/computer/cellphone!
- You are not allowed to bring text book.

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

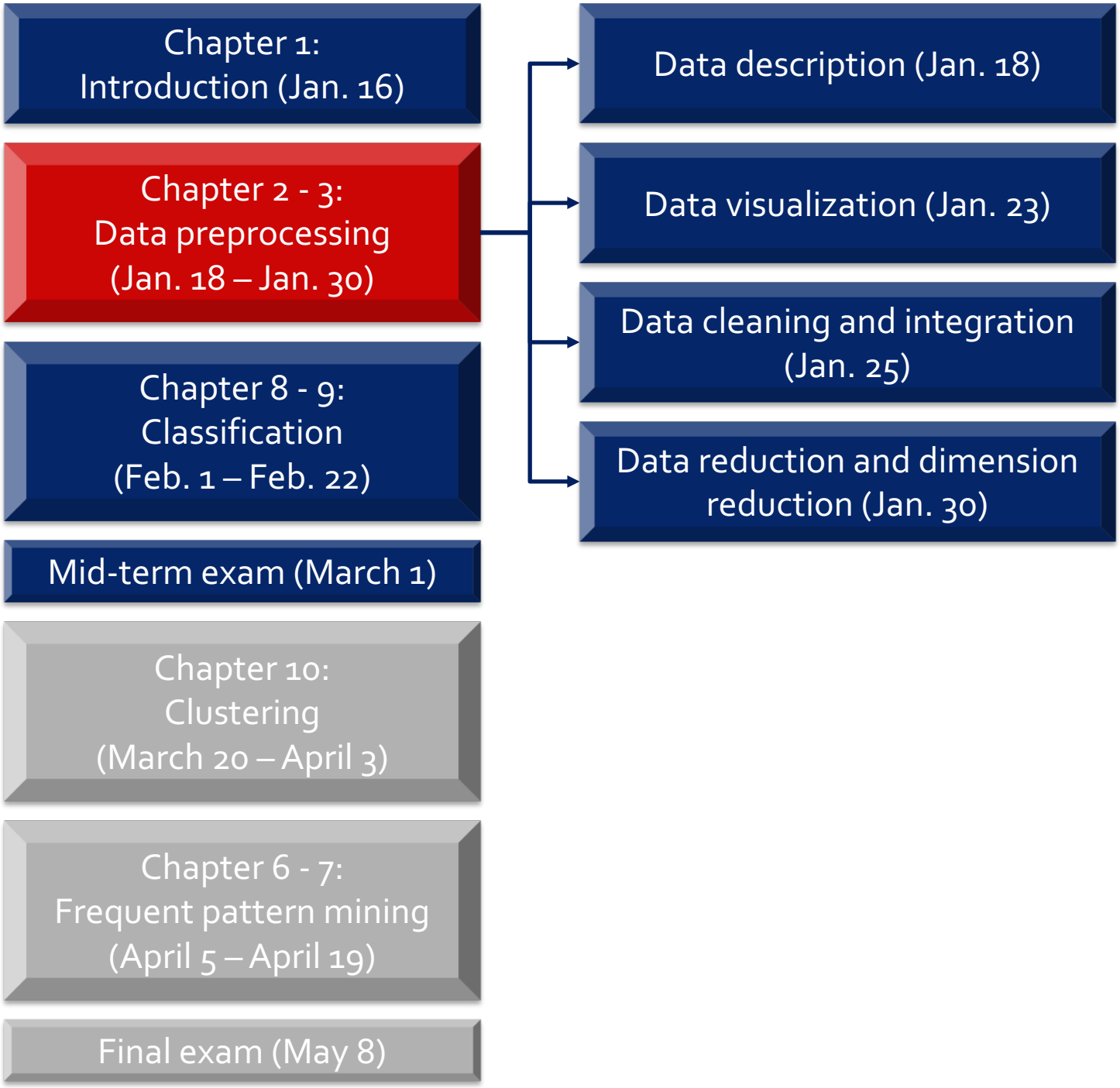
Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

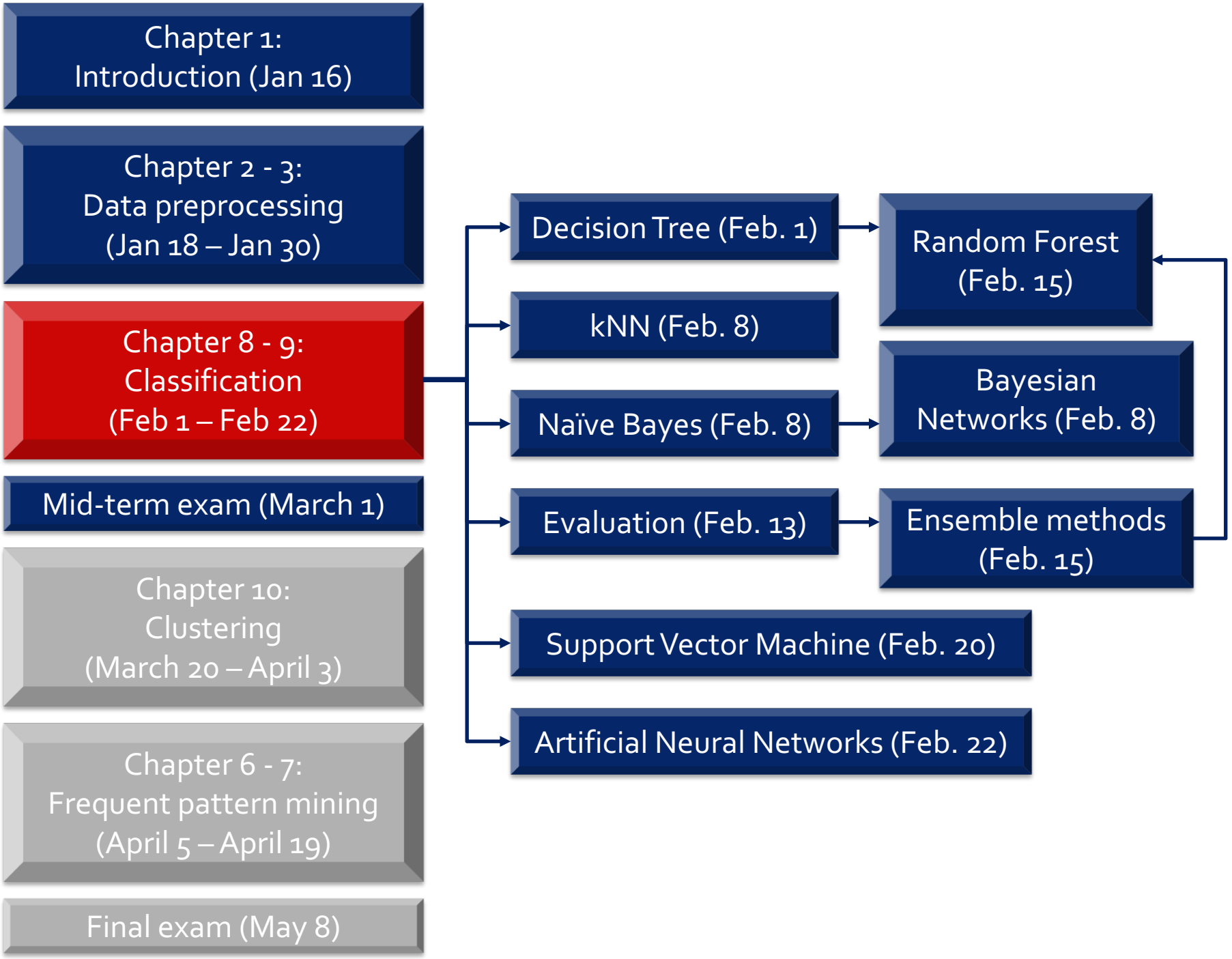
Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

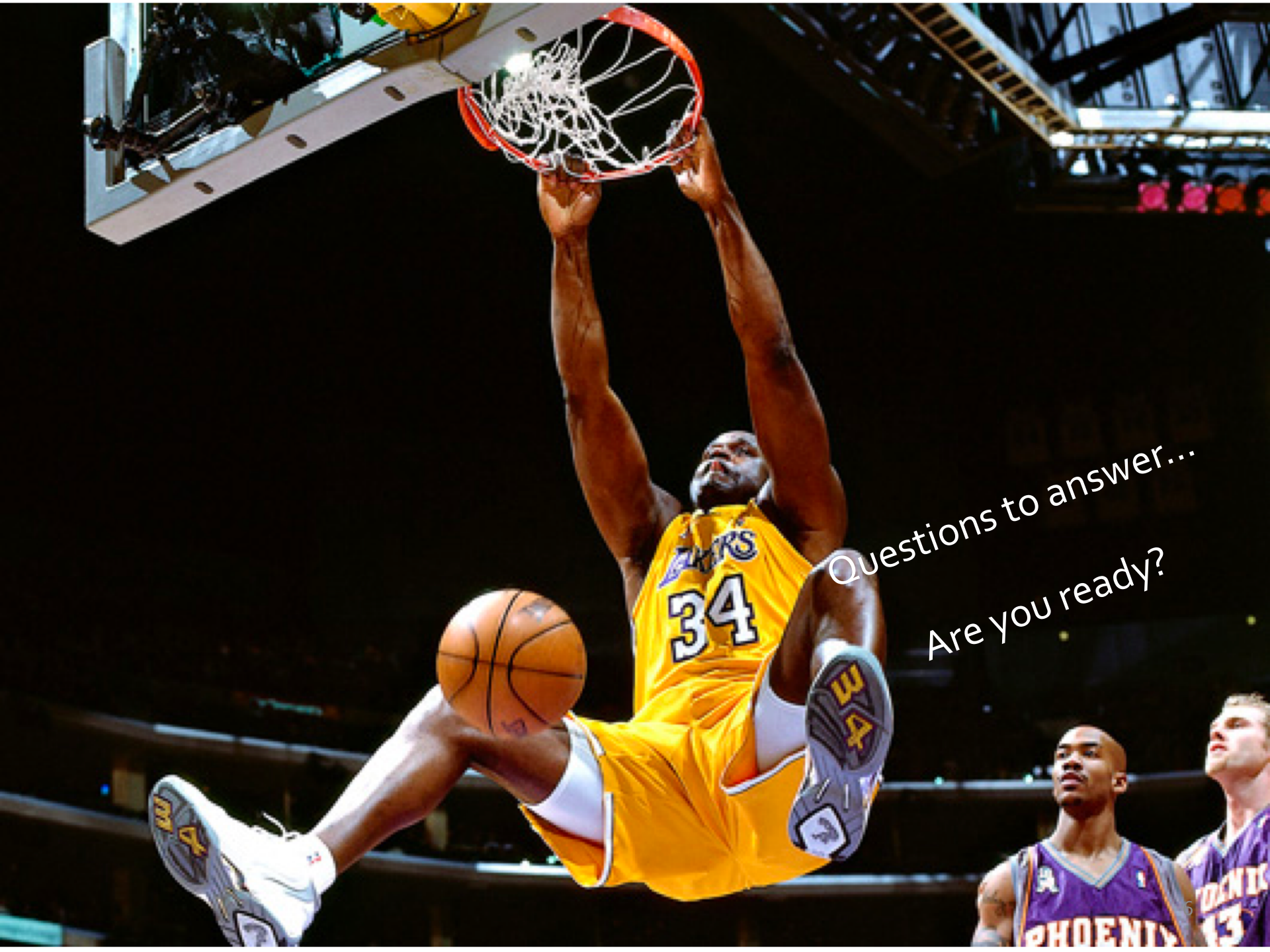
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)


Chapter 1:
Introduction (Jan. 16)
Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)
Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)
Mid-term exam (March 1)
Chapter 10:
Clustering
(March 20 – April 3)
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)
Final exam (May 8)
Data description (Jan. 18)
Data visualization (Jan. 23)
Data cleaning and integration
(Jan. 25)
Data reduction and dimension
reduction (Jan. 30)


Chapter 1:
Introduction (Jan 16)
Chapter 2 - 3:
Data preprocessing
(Jan 18 – Jan 30)
Chapter 8 - 9:
Classification
(Feb 1 – Feb 22)
Mid-term exam (March 1)
Chapter 10:
Clustering
(March 20 – April 3)
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)
Final exam (May 8)
Decision Tree (Feb. 1)
kNN (Feb. 8)
Naïve Bayes (Feb. 8)
Evaluation (Feb. 13)
Support Vector Machine (Feb. 20)
Artificial Neural Networks (Feb. 22)
Random Forest
(Feb. 15)
Bayesian
Networks (Feb. 8)
Ensemble methods
(Feb. 15)


Review


Questions to answer…
Are you ready?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q1: Who is the instructor of ND-CSE Data Science Spring'18?

A) 

B) 

C) 

D) 

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q2: What are the **two** key features of data science research?

A)


B)


C)


D)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q3: What are the **four** Vs of "Big Data"?

A) **Volume**
SCALE OF DATA

B) **Variety**
DIFFERENT FORMS OF DATA

C) **Velocity**
ANALYSIS OF STREAMING DATA

D) **Veracity**
UNCERTAINTY OF DATA

E) **Vacancy**
INCOMPLETENESS OF DATA

F) **Vacation**
CONFERENCES OF BIG DATA

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

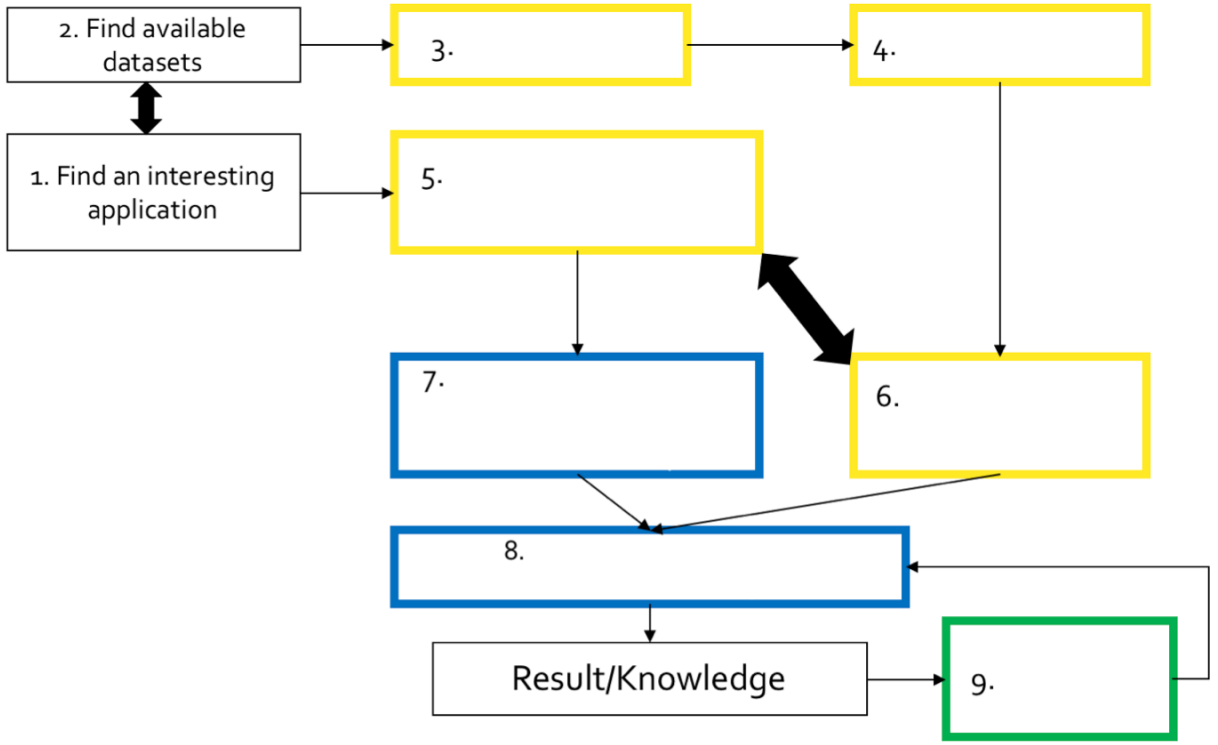
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q4: What are the **components** of data science research?

(Seven steps)



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

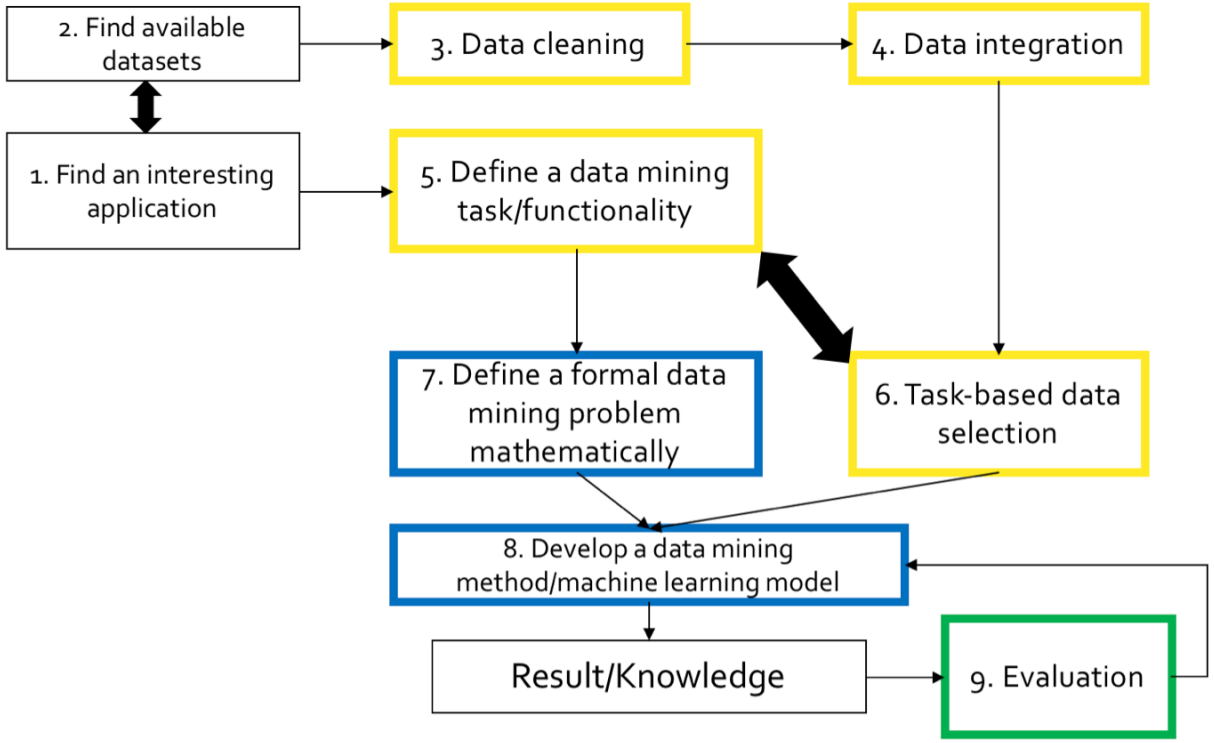
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q4: What are the **components** of data science research?

(Seven steps)



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q5: What are the **data mining functionalities**?

1)

2)

3)

4)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

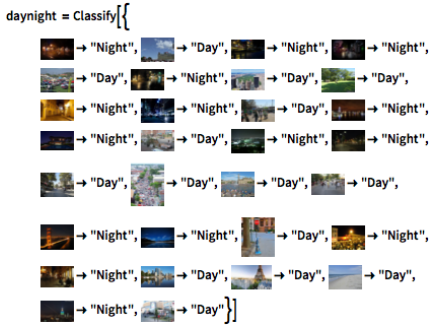
Final exam (May 8)

# Q5: What are the **data mining functionalities**?

Classification

Clustering

1)




2)

3)


4)

Pattern/association mining

Outlier detection

……

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q6: Suppose the task is football game result prediction. Link the concepts.

| | |
|---|---|
| 1) "Home/Away" | a) Data objects / instances / examples / tuples … |
| 2) Past games | b) Feature / attribute |
| 3) "Win/Lose" | c) Label |

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q6: Suppose the task is football game result prediction. Link the concepts.

1) "Home/Away"

2) Past games

3) "Win/Lose"

a) Data objects / instances / examples / tuples …

b) Feature / attribute

c) Label

Answer:

1-b

2-a

3-c

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

## Q7: What are the attribute types?

1)

2)

3)

4)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

# Q7: What are the attribute types?

1)

**Nominal / Categorical**

{"red", "blue"}

2)

**Binary**

{true, false}

3)

**Ordinal**

{freshman, sophomore…}

4)

**Numeric**

#faculty
Math score

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q8: What are the statistical descriptions for central tendency?

1)

2)

3)

4)

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q8: What are the statistical descriptions for central tendency?

1) Mean

2) Median

3) Percentiles

4) Mode

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

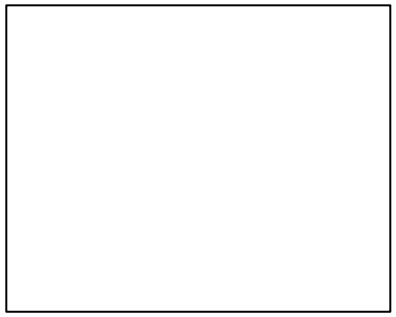
Chapter 10:
Clustering
(March 20 – April 3)

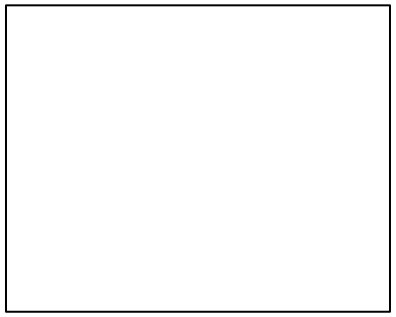
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

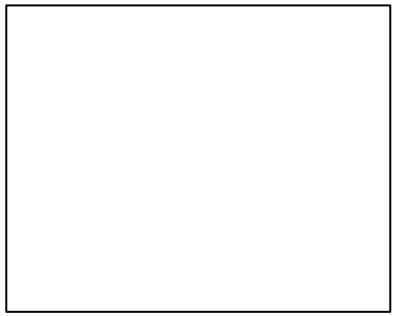
Final exam (May 8)

Data description (Jan. 18)

Q9: What are the statistical descriptions for outlier-ness?

1) 

2) 

3) 

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q9: What are the statistical descriptions for outlier-ness?

1) Variance

2) Standard deviation

3) Z score

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

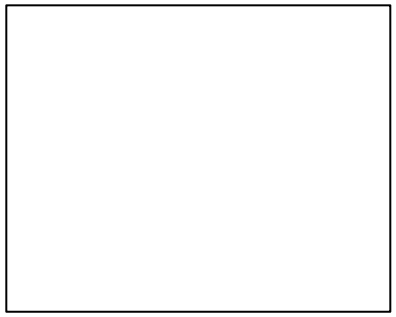
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

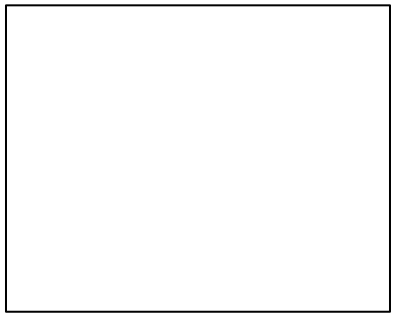
Data description (Jan. 18)

Q10: What are numeric feature normalization methods?

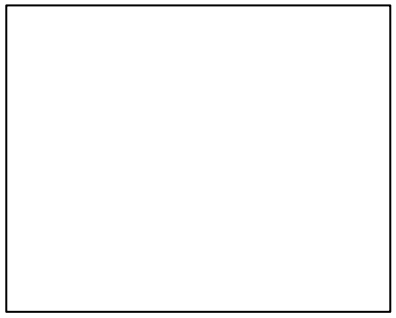
1)



2)



3)



.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data description (Jan. 18)

Q10: What are numeric feature normalization methods?

1) Min-max normalization

2) Z-score normalization

3) Decimal scaling normalization

.....

# HW1 Q1 (Data Description)

We sample $n = 9$ of them

| Student name | *Math* score | *Data Science* score |
|---|---|---|
| Giannis **A**ntetokounmpo | 82 | 84 |
| Kobe **B**ryant | 98 | 97 |
| Stephen **C**urry | 83 | 83 |
| Kevin **D**urant | 95 | 97 |
| Joel **E**mbiid | 76 | 87 |
| Markelle **F**ultz | 71 | 73 |
| Manu **G**inobili | 81 | 83 |
| James **H**arden | 85 | 87 |
| Brandon **I**ngram | 76 | 83 |

1. Calculate *mean*, *median*, and *mode* of *Data Science* scores.
2. Calculate *variance* and *standard deviation* of *Data Science* scores.

$$\mu = \frac{x_1+...+x_n}{n}.$$

$$v = \frac{(x_1-\mu)^2+...+(x_n-\mu)^2}{n-1}.$$

1. $mean = 86.$ $median = 84.$ $mode = 83.$
2. $variance = 55.5.$ $standard\_deviation = 7.45.$

We denote the $i$-th student's *Data Science* score as $x_i$ ($1 \leq i \leq n$), denote the *mean* of these $n$ scores as $\mu$, and denote the *variance* as $v$. Suppose we sample one more student "Michael **J**ordan" whose *Data Science* score is $x_{n+1}$. Now we denote the new mean (of the $n+1$ students' *Data Science* scores) as $\mu'$ and the new variance as $v'$. Please write down the function

$$\mu' = f(\mu, n, x_{n+1}) \tag{1}$$

and the function

$$v' = g(v, \mu, n, x_{n+1}) \tag{2}$$

to incrementally calculate $\mu'$ and $v'$. Note that none of $x_i$ ($1 \leq i \leq n$) or $\mu'$ is allowed to be used in the functions as input variable. You may assume $x_{n+1} = 100$ and use the given data points to verify if your functions are correct or not.

$\mu = \frac{x_1+...+x_n}{n}$. $v = \frac{(x_1-\mu)^2+...+(x_n-\mu)^2}{n-1}$.
$\mu' = \frac{x_1+...+x_n+x_{n+1}}{n+1} = \frac{n\mu+x_{n+1}}{n+1}$.

*Solution 1 on simplifying $v'$:*
$v = \frac{(x_1-\mu)^2+...+(x_n-\mu)^2}{n-1}$.
$v' = \frac{(x_1-\mu')^2+...+(x_n-\mu')^2+(x_{n+1}-\mu')^2}{n}$.
We have

$$\begin{aligned}
nv' - (n-1)v &= \{(x_1-\mu')^2 - (x_1-\mu)^2\} + ... + \{(x_n-\mu')^2 - (x_n-\mu)^2\} + (x_{n+1}-\mu')^2 \\
&= (2x_1 - \mu - \mu') \times (\mu - \mu') + ... + (2x_n - \mu - \mu') \times (\mu - \mu') + (x_{n+1} - \mu')^2 \\
&= \{2 \times (x_1 + ... + x_n) - n\mu - n\mu'\} \times (\mu - \mu') + (x_{n+1} - \mu')^2 \\
&= (2n\mu - n\mu - n\mu') \times (\mu - \mu') + (x_{n+1} - \mu')^2 \\
&= (n\mu - n\mu') \times (\mu - \mu') + (x_{n+1} - \mu')^2 \\
&= n(\mu - \mu')^2 + (x_{n+1} - \mu')^2
\end{aligned}$$

So $v' = v + (\mu - \mu')^2 + \frac{(x_{n+1}-\mu')^2 - v}{n} = \frac{n-1}{n}v + \frac{1}{n+1}(x_{n+1} - \mu)^2$.

*Solution 2 on simplifying $v'$:*

$$v = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1} = \frac{\sum_{i=1}^{n} {x_i}^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2}{n-1} = \frac{\sum_{i=1}^{n} {x_i}^2 - n\mu^2}{n-1}.$$

So,

$$v' = \frac{\sum_{i=1}^{n+1} {x_i}^2 - (n+1)\mu'^2}{n}.$$

Then we have

$$(n-1)v + n\mu^2 = \sum_{i=1}^{n} {x_i}^2 = nv' + (n+1)\mu'^2 - {x_{n+1}}^2.$$

Because $(n+1)\mu' = n\mu + x_{n+1}$, we have

$$\begin{aligned}(n^2-1)v + n(n+1)\mu^2 &= n(n+1)v' + (n+1)^2\mu'^2 - (n+1){x_{n+1}}^2 \\ &= n(n+1)v' + (n\mu + x_{n+1})^2 - (n+1){x_{n+1}}^2 \\ &= (n^2+n)v' + n^2\mu^2 + 2nx_{n+1}\mu - n{x_{n+1}}^2\end{aligned}$$

We have

$$\begin{aligned}v' &= \frac{(n+1)(n-1)v + n\mu^2 - 2nx_{n+1}\mu + n{x_{n+1}}^2}{n(n+1)} \\ &= \frac{n-1}{n}v + \frac{n(\mu - x_{n+1})^2}{n(n+1)} = \frac{n-1}{n}v + \frac{1}{n+1}(x_{n+1}-\mu)^2\end{aligned}$$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

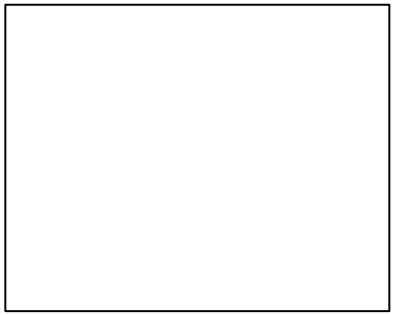
Chapter 10:
Clustering
(March 20 – April 3)

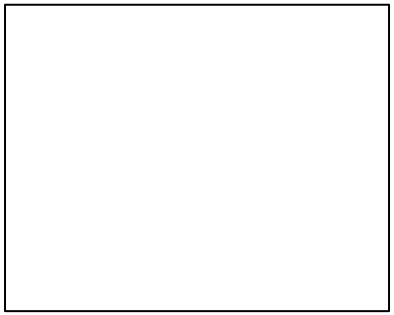
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

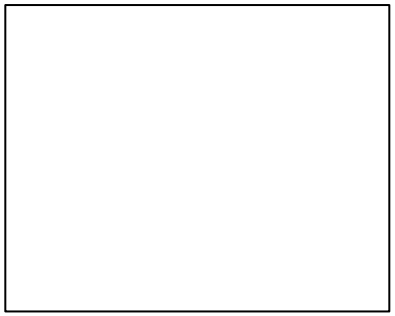
Final exam (May 8)

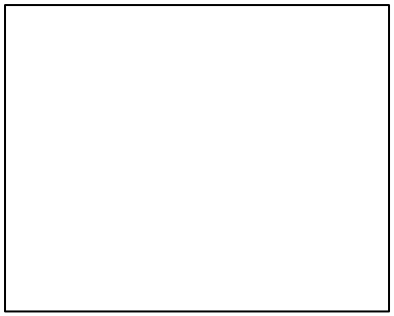
Data visualization (Jan. 23)

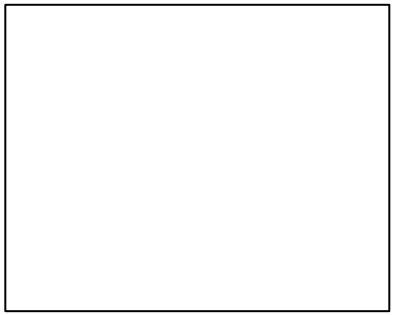
## Q11: What are the plots we learned?

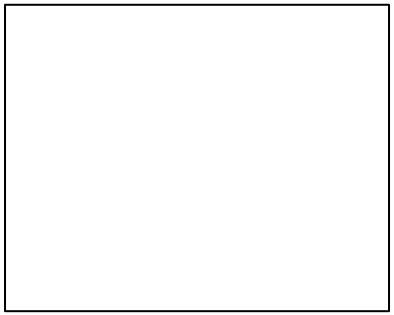
1) 

2) 

3) 

4) 

5) 

6) 

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

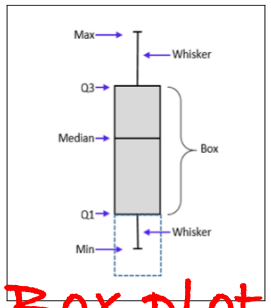
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

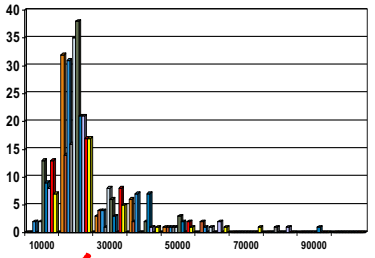
# Q11: What are the plots we learned?

1)



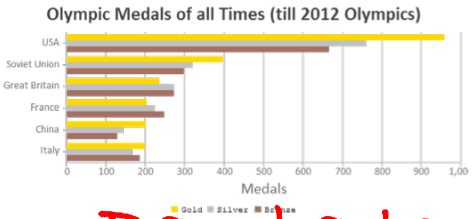

Box plot

2)



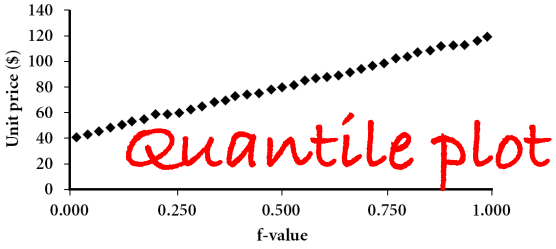

Histogram

3)



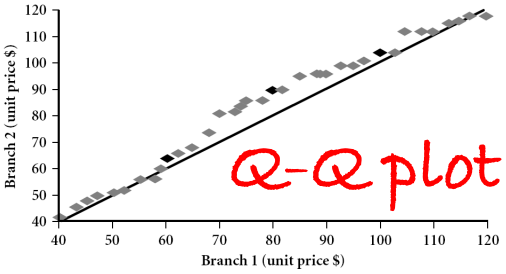

Bar charts

4)



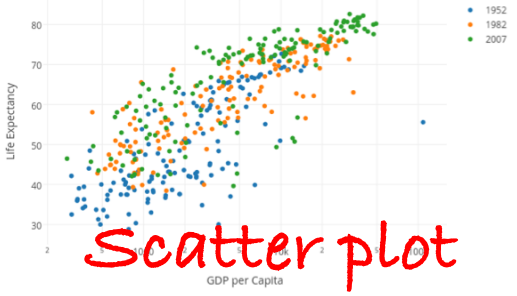

Quantile plot

5)




Q-Q plot

6)

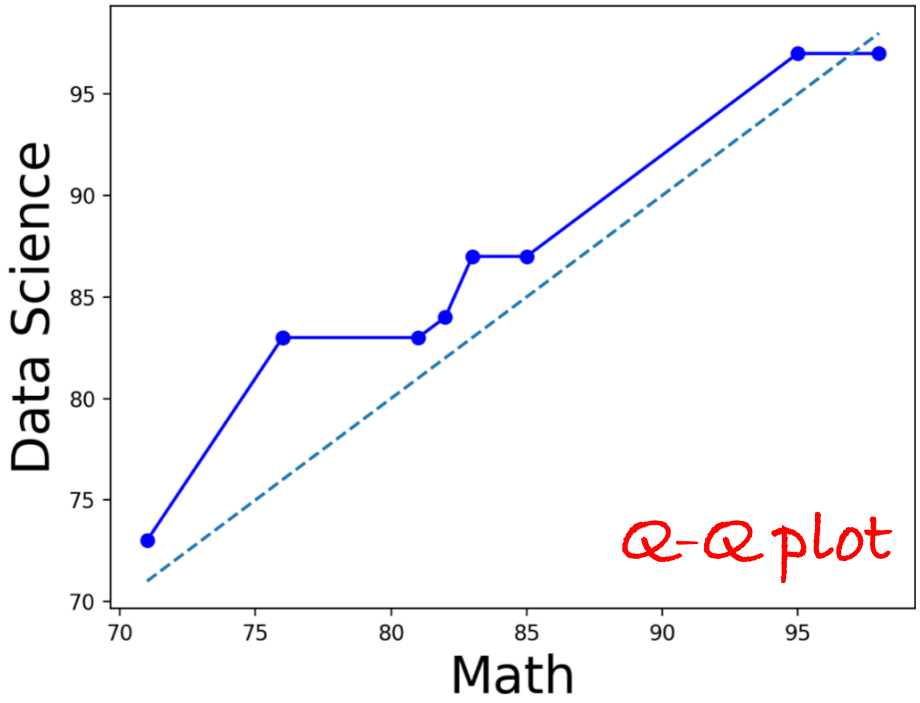


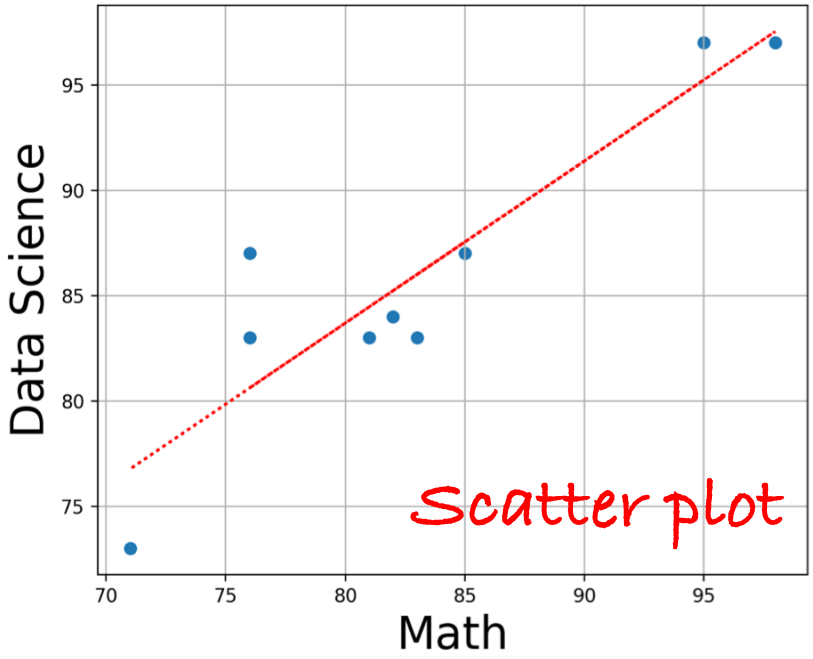

Scatter plot

.....

# HW1 Q2 (Data Visualization)

| Student name | *Math* score | *Data Science* score |
|---|---|---|
| Giannis Antetokounmpo | 82 | 84 |
| Kobe Bryant | 98 | 97 |
| Stephen Curry | 83 | 83 |
| Kevin Durant | 95 | 97 |
| Joel Embiid | 76 | 87 |
| Markelle Fultz | 71 | 73 |
| Manu Ginobili | 81 | 83 |
| James Harden | 85 | 87 |
| Brandon Ingram | 76 | 83 |



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

## Q12: Describe the following concepts.

1) Curse of Dimensionality

2) Data Sparsity

3) Data Resolution

4) Tag Cloud

5) Graphical Integrity

6) Lie Factor

.....

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

# Q13: What are the three special cases of Minkowski distance measures (L-*p* norm)?

$$d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

1) $L_1$ norm

2) $L_2$ norm

3) $L_\infty$ norm

Which is the biggest?
Which is the smallest?

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

# Q13: What are the three special cases of Minkowski distance measures (L-*p* norm)?

$$d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

1) $L_1$ norm

Manhattan distance (biggest)

2) $L_2$ norm

Euclidean distance

3) $L_\infty$ norm

Supremum distance (smallest)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

Q14: Calculate similarities between these two phrases based on the words:

(1) "university of illinois at chicago"

(2) "university of illinois at urbana champaign"

Jaccard (phrase1, phrase2) =

Cosine (phrase1, phrase2) =

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

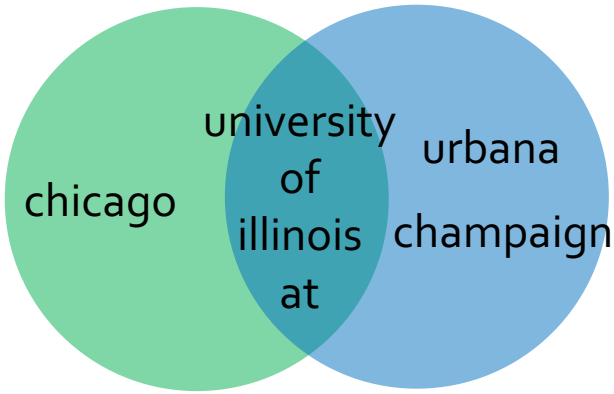
Data visualization (Jan. 23)

Q14: Calculate similarities between these two phrases based on the words:

(1) "university of illinois at chicago"

(2) "university of illinois at urbana champaign"

Jaccard(phrase1, phrase2) = 4/7 = 0.57




| phrase1 \ phrase2 | 1 | 0 |
|---|---|---|
| 1 | q=4 | r=2 |
| 0 | s=1 | t |

Jaccard = q/(q+r+s)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data visualization (Jan. 23)

Q14: Calculate similarities between these two phrases based on the words:

(1) "university of illinois at chicago"

(2) "university of illinois at urbana champaign"

| | u. | of | ill. | at | chi. | urb. | cham. | others | **norm** |
|---|---|---|---|---|---|---|---|---|---|
| p1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0… | sqrt(5) |
| p2 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0… | sqrt(6) |

vec1 = [1,1,1,1,1,0,0,0…]/sqrt(5)
vec1 = [1,1,1,1,0,1,1,0…]/sqrt(6)

Cosine (phrase1, phrase2) = vec1*vec2
= 4/sqrt(5)/sqrt(6) = 0.73

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

Q15: [Cleaning] *Missing/Incomplete data issue*.

Definition:

Why:

Types:

How to handle:

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

Q15: [Cleaning] *Missing/Incomplete data issue*.

| | |
|---|---|
| Definition: | Lacking attribute values<br>Lacking certain attributes … |
| Why: | Not always available<br>Equipment malfunction<br>Deleted due to inconsistency … |
| Types: | Missing Completely at Random<br>Missing at Random<br>Missing Not at Random |
| How to handle: | Manually fill the data<br>Automatically fill the data<br>• Global constant<br>• Filling with attribute mean<br>• Filling with attribute mean for all samples belonging to the same class<br>• Inference-based approaches |

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

## Q16: [Cleaning] *Noisy data issue*.

Definition:

Why:

How to handle:

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

# Q16: [Cleaning] *Noisy data issue*.

| | |
|---|---|
| Definition: | Containing noise, errors, or outliers |
| Why: | Faulty data collection instruments<br>Data transmission problems<br>Technology limitation… |
| How to handle: | • Binning: smooth by bin means, smooth by bin median, smooth by bin boundaries<br>• Regression: Smooth by fitting the data into regression functions<br>• Outlier detection<br>  • Clustering<br>  • Outlier-ness (Z-score) normalization<br>• Human-in-loop denoise |

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

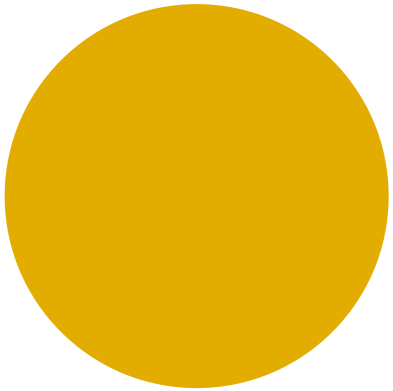
Final exam (May 8)

Data cleaning and integration
(Jan. 25)

## Q17: [Cleaning] *Inconsistency data issue*.

Definition:

Why:



How to handle:

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

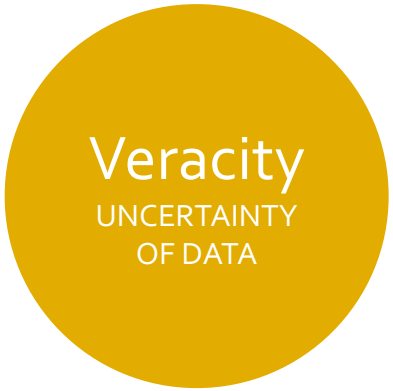
Final exam (May 8)

Data cleaning and integration
(Jan. 25)

# Q17: [Cleaning] *Inconsistency data issue*.

Definition: Containing discrepancies in codes or names …

Why:




How to handle:

- Truth finding

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

Q18: [**Integration**] *Data redundancy issue*.

Definition:

Why:

How to handle:

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data cleaning and integration
(Jan. 25)

Q18: [**Integration**] *Data redundancy issue*.

Definition: Redundant data occur often when integration of multiple databases

Why:

*Object identification*: The same attribute or object may have different names in different databases

*Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue, age

How to handle:

**Correlation analysis (categorical attr.)**

Chi-square test

$$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$

observed

expected

**Covariance analysis (numerical attr.)**

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data reduction and dimension reduction (Jan. 30)

Q19: Describe methods for numerosity reduction (reducing #instances)

Parametric methods

- Regression: Estimate/use the **parameters** of regression models to represent the data

Nonparametric methods

- Histograms: Use **bins** to represent the data
- Clustering: Use **clusters/groups** to represent the data
- Sampling (and stratified sampling): Use samples to represent the data

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data reduction and dimension reduction (Jan. 30)

Q20: Describe regression models.

Least Square Method

- Linear regression
- Non-linear regression
  - Logistic regression
  - Log-linear regression
  - Log-log regression: Power-law distribution

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Data reduction and dimension reduction (Jan. 30)

Q21: Describe methods for dimensionality reduction (reducing #features/attributes)

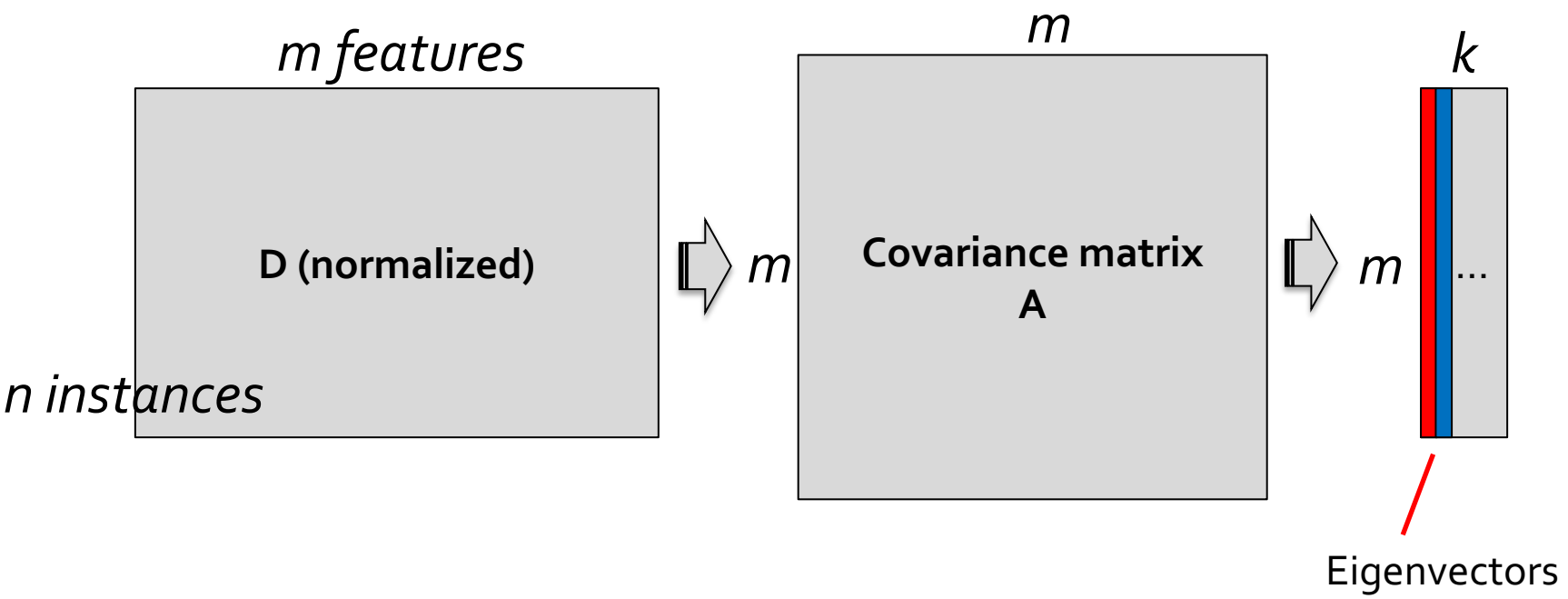
Feature selection

- Heuristic search

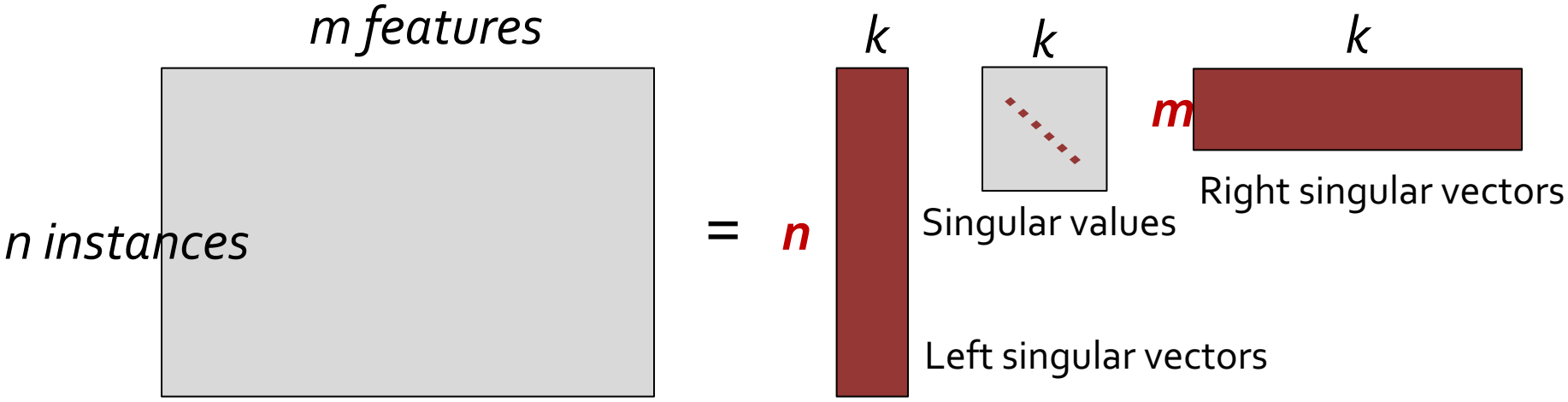
Feature extraction

- Principal Component Analysis (PCA)
  - Eigenvalue and Eigenvectors
- Singular Value Decomposition (SVD)
  - Singular value and singular vectors

# PCA


m features
n instances
D (normalized)
m
m
Covariance matrix
A
m
k
...
Eigenvectors

# SVD


m features
n instances
=
n
k
k
k
m
Singular values
Right singular vectors
Left singular vectors

# HW1 Q3 (Dimension Reduction)

Suppose the matrix **X** (size: $n \times n$) below is the adjacency matrix of student-student social graph: $X_{i,j}$ is "1" if the two students are the same ($i = j$) or connected; "0" if they are different ($i \neq j$) and not connected. We consider the $n = 9$ students as data objects (rows) and as features (columns) themselves.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

Use Python to call a Singular Value Decomposition (SVD) package and calculate *left singular vector* **U** (size: $n \times k$) and *singular values* $\lambda_i$ ($i = 1 \ldots k$) where the number of singular values $k$ is set as 2. The goal is to reduce the number of features from $n$ to $k$.

```
U, s, Vt = svds(A, k=2)
```

# Spectral Clustering using SVD

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

**U** $(9 \times 2)$ is:

The singular values are: **3** and **6**.

| | | |
|---|---|---|
| **A**ntetokounmpo | -9.57037059e-19 | 4.08248290e-01 |
| **B**ryant | 1.61417365e-17 | 4.08248290e-01 |
| **C**urry | -3.92117496e-17 | 4.08248290e-01 |
| **D**urant | 6.38328734e-18 | 4.08248290e-01 |
| **E**mbiid | -5.77350269e-01 | 2.54725816e-17 |
| **F**ultz | -5.77350269e-01 | 6.64582549e-17 |
| **G**inobili | -7.37926952e-17 | 4.08248290e-01 |
| **H**arden | 9.00696433e-18 | 4.08248290e-01 |
| **I**ngram | -5.77350269e-01 | 3.41581563e-17 |

$$\approx \begin{bmatrix} 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \\ -0.57 & 0 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \end{bmatrix}$$

# Review the Equation (k=2)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

$$\approx \begin{bmatrix} 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \\ -0.57 & 0 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & -0.57 & -0.57 & 0 & 0 & -0.57 \\ 0.41 & 0.41 & 0.41 & 0.41 & 0 & 0 & 0.41 & 0.41 & 0 \end{bmatrix}$$

# Start from k=1

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

$$\approx \begin{bmatrix} ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix} \begin{bmatrix} ? \end{bmatrix} \begin{bmatrix} ? & ? & ? & ? & ? & ? & ? & ? & ? \end{bmatrix}$$

# HITS Algorithm (Kleinberg 1999)

$$\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix} \Rightarrow \begin{bmatrix}0.33\\0.33\\0.33\\0.33\\0.33\\0.33\\0.33\\0.33\\0.33\end{bmatrix}$$

$$[0.384 \quad 0.384 \quad 0.384 \quad 0.384 \quad 0.192 \quad 0.192 \quad 0.384 \quad 0.384 \quad 0.192\ ]$$

$$[2 \quad 2 \quad 2 \quad 2 \quad 1 \quad 1 \quad 2 \quad 2 \quad 1\ ]$$

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

# HITS Algorithm (Kleinberg 1999)

$$[0.384 \quad 0.384 \quad 0.384 \quad 0.384 \quad 0.192 \quad 0.192 \quad 0.384 \quad 0.384 \quad 0.192\ ]$$

$$\begin{bmatrix} 0.402 \\ 0.402 \\ 0.402 \\ 0.402 \\ 0.1 \\ 0.1 \\ 0.402 \\ 0.402 \\ 0.1 \end{bmatrix}$$

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

# HITS Algorithm (Kleinberg 1999)

```
Iter 0 U= [0.33333333 0.33333333 0.33333333 0.33333333 0.33333333 0.33333333 0.33333333 0.33333333 0.33333333]
Iter 0 V= [0.38490018 0.38490018 0.38490018 0.38490018 0.19245009 0.19245009 0.38490018 0.38490018 0.19245009]
Iter 1 U= [0.40201513 0.40201513 0.40201513 0.40201513 0.10050378 0.10050378 0.40201513 0.40201513 0.10050378]
Iter 1 V= [0.40666285 0.40666285 0.40666285 0.40666285 0.05083286 0.05083286 0.40666285 0.40666285 0.05083286]
Iter 2 U= [0.40785019 0.40785019 0.40785019 0.40785019 0.02549064 0.02549064 0.40785019 0.40785019 0.02549064]
Iter 2 V= [0.40814866 0.40814866 0.40814866 0.40814866 0.01275465 0.01275465 0.40814866 0.40814866 0.01275465]
Iter 3 U= [0.40822338 0.40822338 0.40822338 0.40822338 0.00637849 0.00637849 0.40822338 0.40822338 0.00637849]
Iter 3 V= [0.40824206 0.40824206 0.40824206 0.40824206 0.00318939 0.00318939 0.40824206 0.40824206 0.00318939]
Iter 4 U= [0.40824673 0.40824673 0.40824673 0.40824673 0.00159471 0.00159471 0.40824673 0.40824673 0.00159471]
Iter 4 V= [0.4082479  0.4082479  0.4082479  0.4082479  0.00079736 0.00079736 0.4082479  0.4082479  0.00079736]
Iter 5 U= [4.08248193e-01 4.08248193e-01 4.08248193e-01 4.08248193e-01 3.98679876e-04 3.98679876e-04 4.08248193e-01 4.08248193e-01 3.98679876e-04]
Iter 5 V= [4.08248266e-01 4.08248266e-01 4.08248266e-01 4.08248266e-01 1.99339974e-04 1.99339974e-04 4.08248266e-01 4.08248266e-01 1.99339974e-04]
Iter 6 U= [4.08248284e-01 4.08248284e-01 4.08248284e-01 4.08248284e-01 9.96699913e-05 9.96699913e-05 4.08248284e-01 4.08248284e-01 9.96699913e-05]
Iter 6 V= [4.08248289e-01 4.08248289e-01 4.08248289e-01 4.08248289e-01 4.98349962e-05 4.98349962e-05 4.08248289e-01 4.08248289e-01 4.98349962e-05]
Iter 7 U= [4.08248290e-01 4.08248290e-01 4.08248290e-01 4.08248290e-01 2.49174982e-05 2.49174982e-05 4.08248290e-01 4.08248290e-01 2.49174982e-05]
Iter 7 V= [4.08248290e-01 4.08248290e-01 4.08248290e-01 4.08248290e-01 1.24587491e-05 1.24587491e-05 4.08248290e-01 4.08248290e-01 1.24587491e-05]
Iter 8 U= [4.08248290e-01 4.08248290e-01 4.08248290e-01 4.08248290e-01 6.22937455e-06 6.22937455e-06 4.08248290e-01 4.08248290e-01 6.22937455e-06]
Iter 8 V= [4.08248290e-01 4.08248290e-01 4.08248290e-01 4.08248290e-01 3.11468727e-06 3.11468727e-06 4.08248290e-01 4.08248290e-01 3.11468727e-06]
```

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Antetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Bryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Curry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Durant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Embiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| Fultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| Ginobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Harden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| Ingram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

$$\approx \begin{bmatrix} 0.41 \\ 0.41 \\ 0.41 \\ 0.41 \\ 0 \\ 0 \\ 0.41 \\ 0.41 \\ 0 \end{bmatrix} [6] \begin{bmatrix} 0.41 & 0.41 & 0.41 & 0.41 & 0 & 0 & 0.41 & 0.41 & 0 \end{bmatrix}$$

# Then?

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

$$\approx \begin{bmatrix} 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \\ -0.57 & 0 \\ 0 & 0.41 \\ 0 & 0.41 \\ -0.57 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & -0.57 & -0.57 & 0 & 0 & -0.57 \\ 0.41 & 0.41 & 0.41 & 0.41 & 0 & 0 & 0.41 & 0.41 & 0 \end{bmatrix}$$

# Update A as …

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| **A**ntetokounmpo | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **B**ryant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **C**urry | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **D**urant | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **E**mbiid | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **F**ultz | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| **G**inobili | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **H**arden | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| **I**ngram | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |

$$- \begin{bmatrix} 0.41 \\ 0.41 \\ 0.41 \\ 0.41 \\ 0 \\ 0 \\ 0.41 \\ 0.41 \\ 0 \end{bmatrix} [6] [0.41 \quad 0.41 \quad 0.41 \quad 0.41 \quad 0 \quad 0 \quad 0.41 \quad 0.41 \quad 0 ]$$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

# Q22: Classification vs clustering. Supervised learning vs unsupervised learning.

Supervised learning (**classification**)

- Supervision: The training data instances and their attributes/features are accompanied by labels indicating the class of the instances.
- **Predict labels** for testing data instances.

Unsupervised learning (**clustering**)

- The class **labels** of training data is **unknown**
- Given a set of attributes, with the aim of establishing the existence of classes or clusters.

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

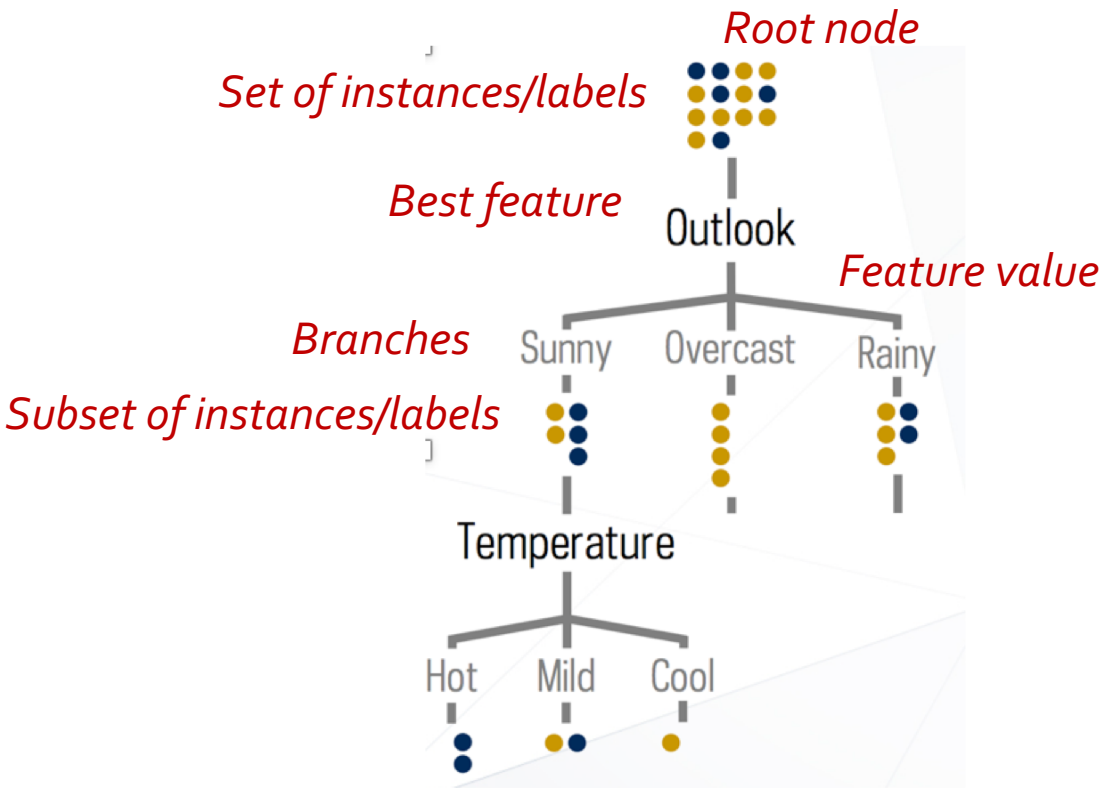
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Decision Tree (Feb. 1)

Q23: Describe decision tree.




- *Top-down*
- *Recursive*
- *Divide-and-conquer*
- *Feature selection*
- *Termination conditions*

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Decision Tree (Feb. 1)

Q24: Describe the following concepts and models.

ID3

- Information Gain
  - Entropy

C4.5

- Grain Ratio
  - SplitInfo

CART

- Gini Index

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

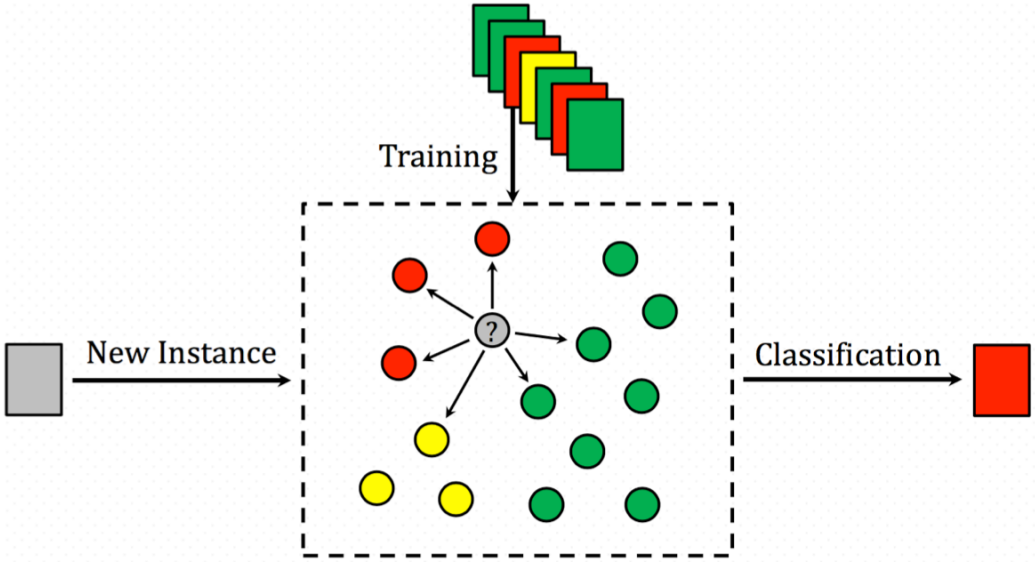
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

kNN (Feb. 8)

Q25: Describe the k-Nearest Neighbor classifier.



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Naïve Bayes (Feb. 8)

Q26: Describe Naïve Bayes model.

P(**X**|H): Likelihood

P(H): Prior Probability

P(**X**): Evidence

P(H|**X**): Posteriori Probability

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Naïve Bayes (Feb. 8)
Bayesian Networks (Feb. 8)

Q27: Describe the following concepts and models.

Zero-Probability Laplacian Correction

"Naïve": Assume conditional independence

Causal analysis?

Bayesian Networks

Directed Acyclic Graph (DAG)
Conditional Probability Tables (CPTs)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Evaluation (Feb. 13)

Q28: Describe the following validation settings

Hold-out validation

- k times: mean and standard derivation

Cross validation methods

- k-fold cross validation
- Leave-one-out cross validation

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Evaluation (Feb. 13)

Q29: Describe the following evaluation metrics.

- Confusion matrix (TP, FP, FN, TN)
- Accuracy, Error rate
- Sensitivity, Specificity
- Precision, Recall, F measure, G measure
- ROC curves, Area Under the Curve (AUC), Precision-Recall Curve
- Precision@K, Average precision
- Mean absolute error (MAE), Root mean squared error (RMSE)
- Ranking-based measures (Kendall's tau, Spearman's rho)

# HW2

- ID3: Information Gain
- C4.5: Gain Ratio
- Naïve Bayes

| ID | Date | Opponent | Is Home or Away? | Is Opponent in AP Top 25 at Preseason? | Media | Label: Win/Lose |
|---|---|---|---|---|---|---|
| 1 | 9/5/15 | Texas | Home | Out | 1-NBC | Win |
| 2 | 9/12/15 | Virginia | Away | Out | 4-ABC | Win |
| 3 | 9/19/15 | Georgia Tech | Home | In | 1-NBC | Win |
| 4 | 9/26/15 | UMass | Home | Out | 1-NBC | Win |
| 5 | 10/3/15 | Clemson | Away | In | 4-ABC | Lose |
| 6 | 10/10/15 | Navy | Home | Out | 1-NBC | Win |
| 7 | 10/17/15 | USC | Home | In | 1-NBC | Win |
| 8 | 10/31/15 | Temple | Away | Out | 4-ABC | Win |
| 9 | 11/7/15 | PITT | Away | Out | 4-ABC | Win |
| 10 | 11/14/15 | Wake Forest | Home | Out | 1-NBC | Win |
| 11 | 11/21/15 | Boston College | Away | Out | 1-NBC | Win |
| 12 | 11/28/15 | Stanford | Away | In | 3-FOX | Lose |
| 13 | 9/4/16 | Texas | Away | Out | 4-ABC | Lose |
| 14 | 9/10/16 | Nevada | Home | Out | 1-NBC | Win |
| 15 | 9/17/16 | Michigan State | Home | Out | 1-NBC | Lose |
| 16 | 9/24/16 | Duke | Home | Out | 1-NBC | Lose |
| 17 | 10/1/16 | Syracuse | Home | Out | 2-ESPN | Win |
| 18 | 10/8/16 | North Carolina State | Away | Out | 4-ABC | Lose |
| 19 | 10/15/16 | Stanford | Home | In | 1-NBC | Lose |
| 20 | 10/29/16 | Miami Florida | Home | Out | 1-NBC | Win |
| 21 | 11/5/16 | Navy | Home | Out | 5-CBS | Lose |
| 22 | 11/12/16 | Army | Home | Out | 1-NBC | Win |
| 23 | 11/19/16 | Virginia Tech | Home | In | 1-NBC | Lose |
| 24 | 11/26/16 | USC | Away | In | 4-ABC | Lose |
| 25 | 9/2/17 | Temple | Home | Out | 1-NBC | Win |
| 26 | 9/9/17 | Georgia | Home | In | 1-NBC | Lose |
| 27 | 9/16/17 | Boston College | Away | Out | 2-ESPN | Win |
| 28 | 9/23/17 | Michigan State | Away | Out | 3-FOX | Win |
| 29 | 9/30/17 | Miami Ohio | Home | Out | 1-NBC | Win |
| 30 | 10/7/17 | North Carolina | Away | Out | 4-ABC | Win |
| 31 | 10/21/17 | USC | Home | In | 1-NBC | Win |
| 32 | 10/28/17 | North Carolina State | Home | Out | 1-NBC | Win |
| 33 | 11/4/17 | Wake Forest | Home | Out | 1-NBC | Win |
| 34 | 11/11/17 | Miami Florida | Away | In | 4-ABC | Lose |
| 35 | 11/18/17 | Navy | Home | Out | 1-NBC | Win |
| 36 | 11/25/17 | Stanford | Away | In | 4-ABC | Lose |

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Ensemble methods (Feb. 15)

Q30: Describe the following ensemble methods and concepts.

Bagging

- Sampling with replacement
- Random Forest (bagged decision trees)

Boosting

- AdaBoost
  - Iteratively learn the classifiers
  - Weights to instances (misclassified)
  - Weights to classifiers (when voting)

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

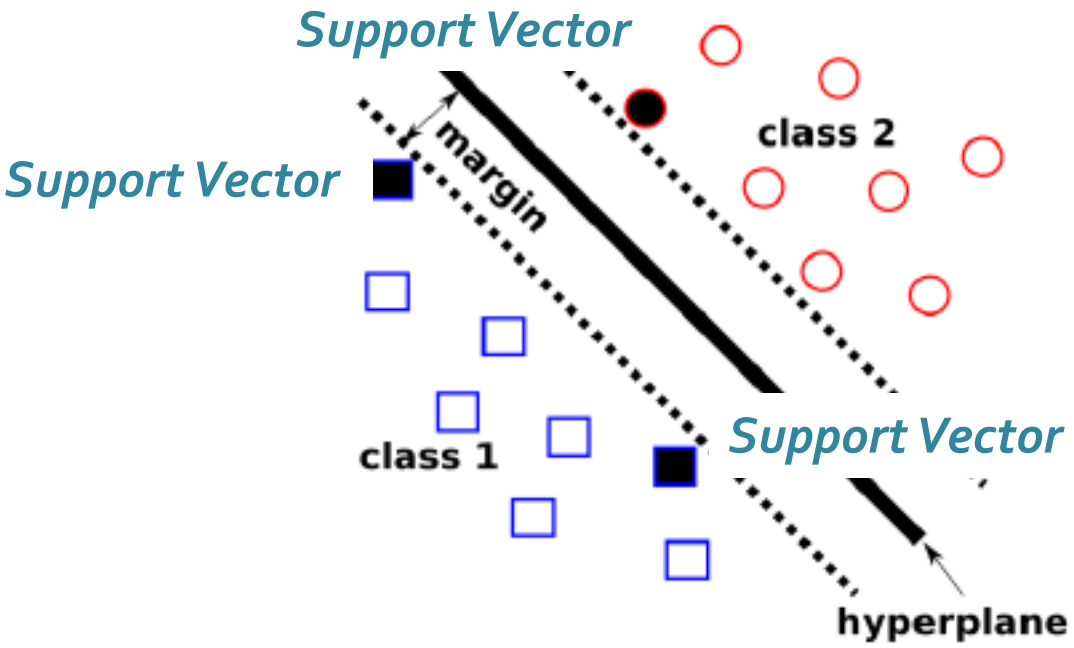
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Support Vector Machines
(Feb. 20)

Q31: Describe the following concepts.

1. **Hyperplane**
2. **Supper Vector**
3. **Margin**
4. **SVMs**
5. **Maximize Margin Width**
6. **Non-linear SVMs: Kernel Function**



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

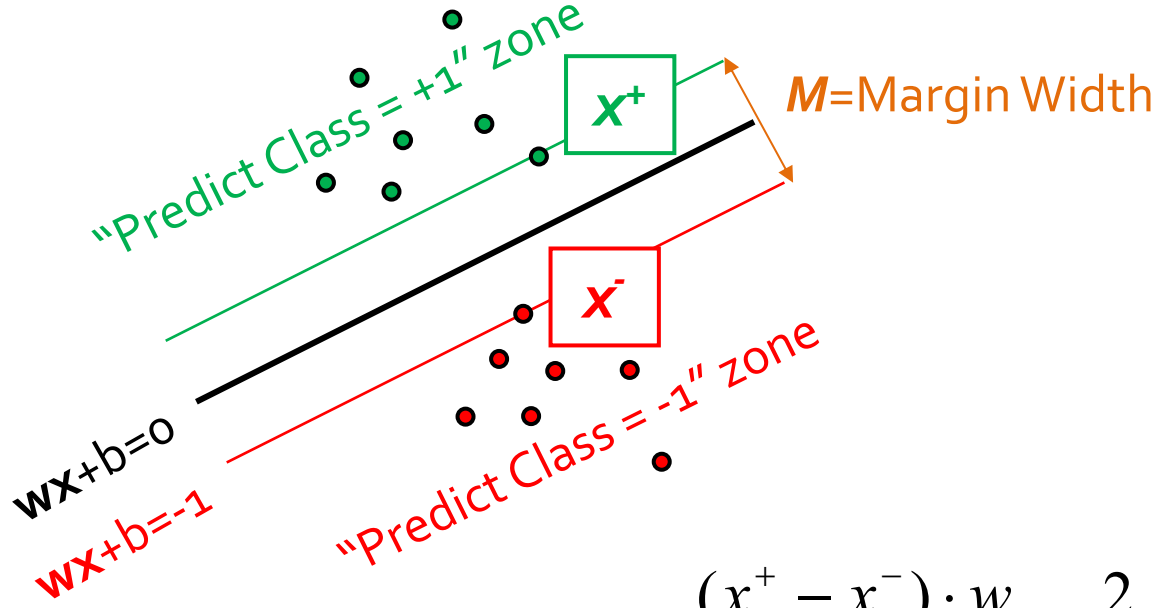
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Support Vector Machines
(Feb. 20)

Q31: Describe the following concepts.

1. **Hyperplane**
2. **Supper Vector**
3. **Margin**
4. **SVMs**
5. **Maximize Margin Width**
6. **Non-linear SVMs: Kernel Function**




$$max \quad M = \frac{(x^+ - x^-) \cdot w}{|w|} = \frac{2}{|w|}$$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Support Vector Machines
(Feb. 20)

Q31: Describe the following concepts.

1. **Hyperplane**
2. **Supper Vector**
3. **Margin**
4. **SVMs**
5. **Maximize Margin Width**
6. **Non-linear SVMs: Kernel Function**

$$\min \ \Phi(w) = \frac{1}{2} w^t w$$

$$\text{s.t.} \ \ y_i(wx_i + b) \geq 1 \ \text{ for all i}$$

**It's a Constrained (Convex) Quadratic Optimization Problem!**

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

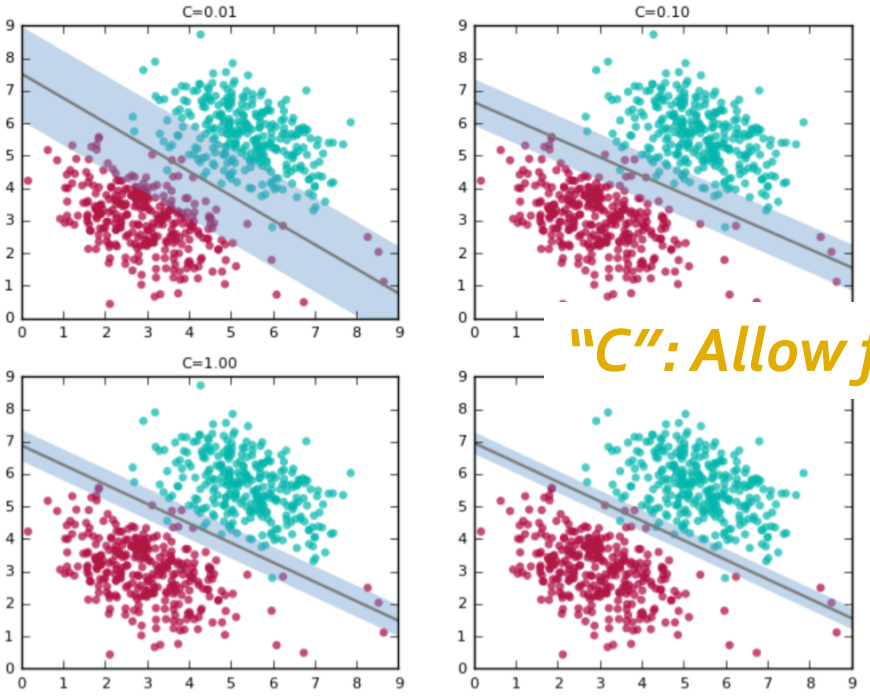
Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Support Vector Machines
(Feb. 20)

Q31: Describe the following concepts.

1. **Hyperplane**
2. **Supper Vector**
3. **Margin**
4. **SVMs**
5. **Maximize Margin Width**
6. **Non-linear SVMs: Kernel Function**




*"C": Allow for Errors*

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

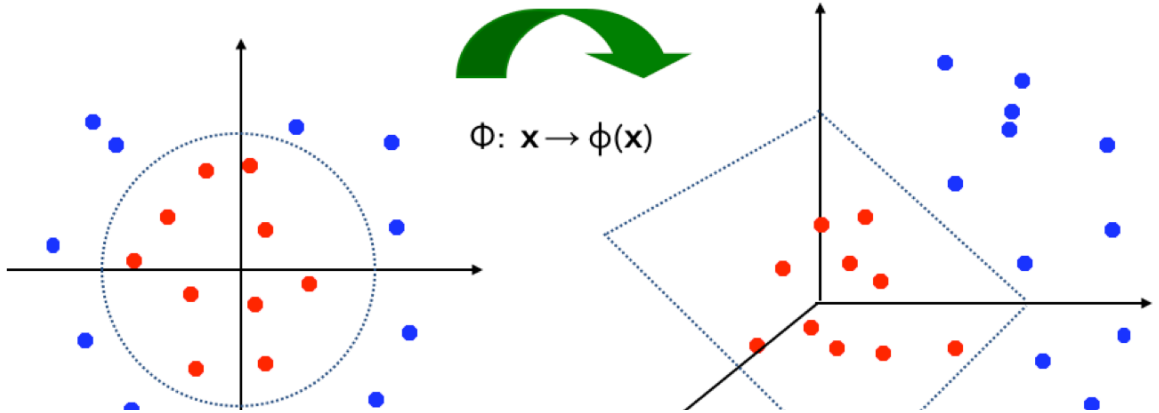
Final exam (May 8)

Support Vector Machines
(Feb. 20)

Q31: Describe the following concepts.

1. **Hyperplane**
2. **Supper Vector**
3. **Margin**
4. **SVMs**
5. **Maximize Margin Width**
6. **Non-linear SVMs: Kernel Function**

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$




Polynomial kernel of degree $h$ : $K(X_i, X_j) = (X_i \cdot X_j + 1)^h$

Gaussian radial basis function kernel : $K(X_i, X_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$

Sigmoid kernel : $K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta)$

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

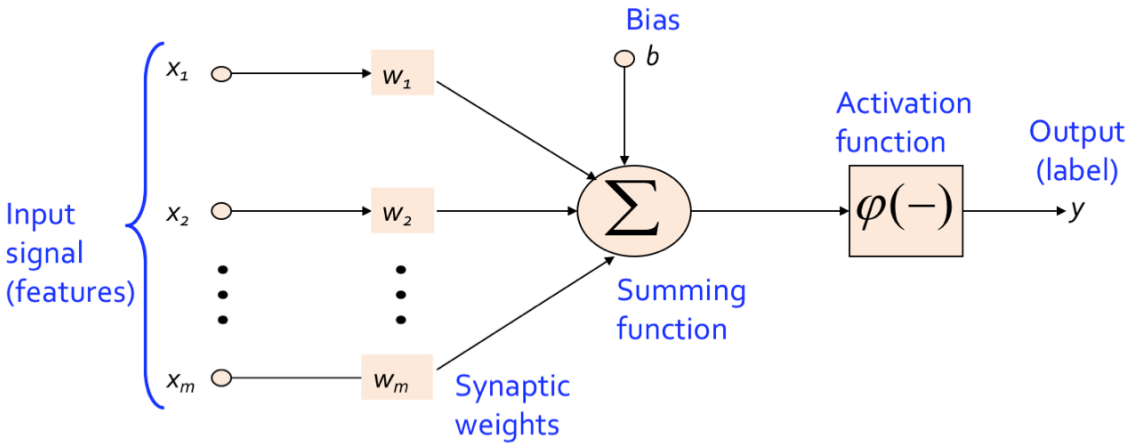
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Artificial Neural Networks
(Feb. 22)

Q32: Describe the following concepts.




Activation functions

- Logistic function
- Sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}$$

- Hyperbolic tangent function

Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

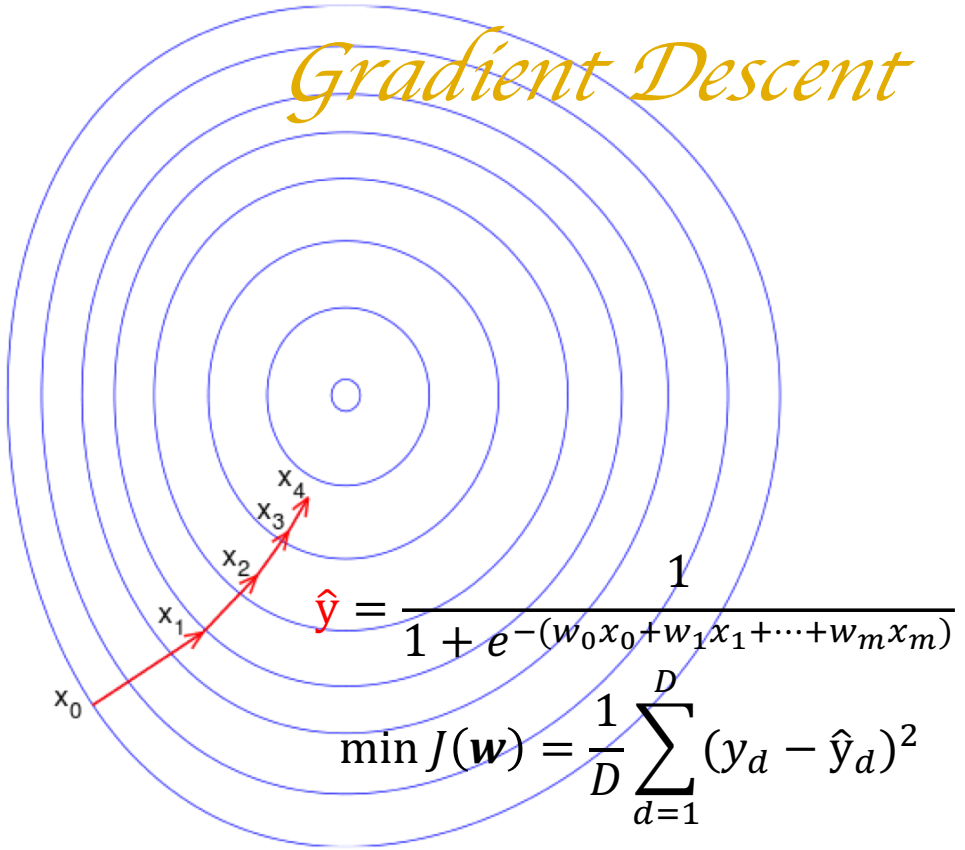
Final exam (May 8)

Artificial Neural Networks
(Feb. 22)

Q33: Describe the optimization process.

$$\mathbf{x} \leftarrow \mathbf{x} - R\nabla f(\mathbf{x})$$

$$x_i \leftarrow x_i - R\frac{\delta f}{\delta x_i}$$



Chapter 1:
Introduction (Jan. 16)

Chapter 2 - 3:
Data preprocessing
(Jan. 18 – Jan. 30)

Chapter 8 - 9:
Classification
(Feb. 1 – Feb. 22)

Mid-term exam (March 1)

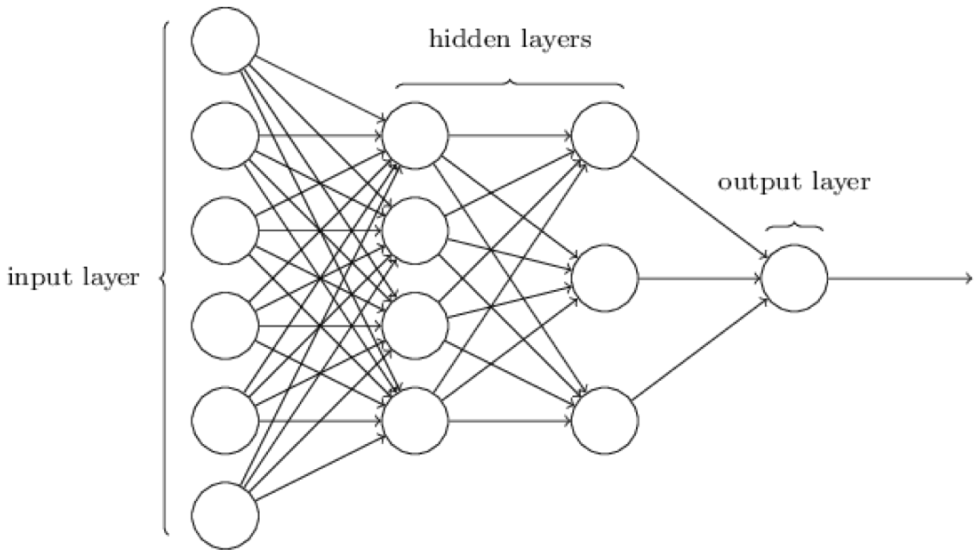
Chapter 10:
Clustering
(March 20 – April 3)

Chapter 6 - 7:
Frequent pattern mining
(April 5 – April 19)

Final exam (May 8)

Artificial Neural Networks
(Feb. 22)

Q34: Describe input/hidden/output layers and count parameters.



# Research We Introduced

- Data Preprocessing
  - Causal analysis using Propensity Score Matching with Q-Q plot
  - Twitter botnet account detection using SVD
- Classification
  - Scientific concept typing using kNN
  - Corpus-based set expansion using Rank Ensembles
  - Others
    - ND Football game prediction using Decision Trees and Naïve Bayes (HW2)
    - Soccer goal detection using SVMs and LibSVM
    - ImageNet using Deep Convolutional Neural Networks (152 layers)
    - …


KEEP
CALM
AND
GOOD
LUCK!