

CSCI 556: Data Analysis & Visualization

Instructor: Dr. Jinoh Kim



This course...

- ❖ Entry-level course for AI/data systems track
 - Topics are introductory without in-depth discussion of algorithms and methods
- ❖ Not recommend to those who have taken any "advanced" course in machine learning or data mining (e.g., CSCI527 Data Mining, CSCI574 Machine Learning)



First Week

- ❖ Syllabus discussion
- ❖ Course overview
- ❖ Action items for first week



The Instructor

- ❖ Jinoh Kim, PhD (call me “Dr. Kim”)
 - Asst/Assoc Professor @ TAMUC, 2012-present
 - Assistant Professor @ LHUP, 2011-2012
 - Researcher @ Berkeley Lab, 2010-2011
 - PhD from U of MN

- ❖ Research areas:
 - Networked systems and security
 - Systems/network telemetry and analytics



Office Hours

- ❖ Office: CS/JOUR 217
- ❖ Office hours (**Appointment-based**):
 - Will be posted on the course page
- ❖ Email communications
 - To: Jinoh.Kim@tamuc.edu
 - Indicate the course number in the email subject line
- ❖ Welcome your visit not only for the course but also for your research and career build-up

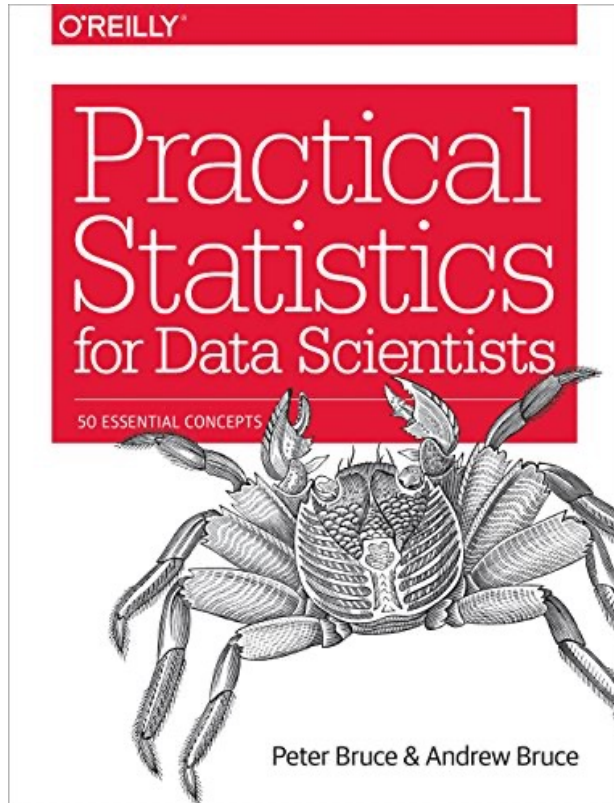


Communications

- ❖ Online course
- ❖ Communicating thru (1) course page and (2) email communications
 - REQUIRED: Read announcements from the course page frequently (at least three times a week) and email messages from the instructor without any significant delays (less than 48 hours).
 - Instructor may not reply over weekend (Fri evening ~ Mon morning)
- ❖ Instructor email: Jinoh.Kim@tamuc.edu



Textbooks



(Optional) Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition. ISBN-13: 978-0128042915, ISBN-10: 0128042915

- ❖ (Mandatory) Practical Statistics for Data Scientists (Essential Concepts), 1st Edition, ISBN 10: [1491952962](#) ISBN 13: [9781491952962](#)
- ❖ Course materials (e.g., slides, hand-outs) will be



Course Highlights

- ❖ Learn data analysis concepts
 - Data exploratory analysis and statistics
 - Classification, regression, clustering
- ❖ Utilize an open-source data analysis and visualization tool: Weka
 - <http://old-www.cms.waikato.ac.nz/~ml/weka/>



Using Weka (Example)

The image shows two windows from the Weka software interface. The left window is 'Weka Explorer' and the right window is 'Weka Clusterer Visualize: 14:18:48 - SimpleKMeans (iris)'.

Weka Explorer - Clusterer

Buttons: Preprocess, Classify, Cluster, Associate, Select attributes, Visualize

Clusterer: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2

Cluster mode

- ☐ Use training set
- ☐ Supplied test set (Set...)
- ☐ Percentage split (% 66)
- ☒ Classes to clusters evaluation (Nom) class
- ☒ Store clusters for visualization

Buttons: Ignore attributes, Start, Stop

Result list (right-click for options)

- 14:17:23 - EM
- 14:17:42 - EM
- 14:18:03 - SimpleKMeans
- 14:18:48 - SimpleKMeans

Clusterer output

	(150.0)	(100.0)	(50.0)
sepalength	5.8433	6.262	5.006
sepalwidth	3.054	2.872	3.418
petallength	3.7587	4.906	1.464
petalwidth	1.1987	1.676	0.244

Time taken to build model (full training data) :
=== Model and evaluation on training set ===

Clustered Instances

Cluster	Count	Percentage
0	100	67%
1	50	33%

Class attribute: class
Classes to Clusters:

```
0 1 <-- assigned to cluster
0 50 | Iris-setosa
50 0 | Iris-versicolor
50 0 | Iris-virginica
```

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa

Incorrectly clustered instances : 50.0 33.3333 %

Weka Clusterer Visualize: 14:18:48 - SimpleKMeans (iris)

X: Instance_number (Num) Y: sepalength (Num)
Colour: Cluster (Nom) Select Instance

Buttons: Reset, Clear, Open, Save

Jitter: ☐

Plot: iris_clustered

A scatter plot showing data points colored by cluster. The x-axis is labeled 'Instance_number' and the y-axis is labeled 'sepalength'. The plot shows two distinct clusters of points, one colored red (cluster 0) and one colored blue (cluster 1).

Class colour

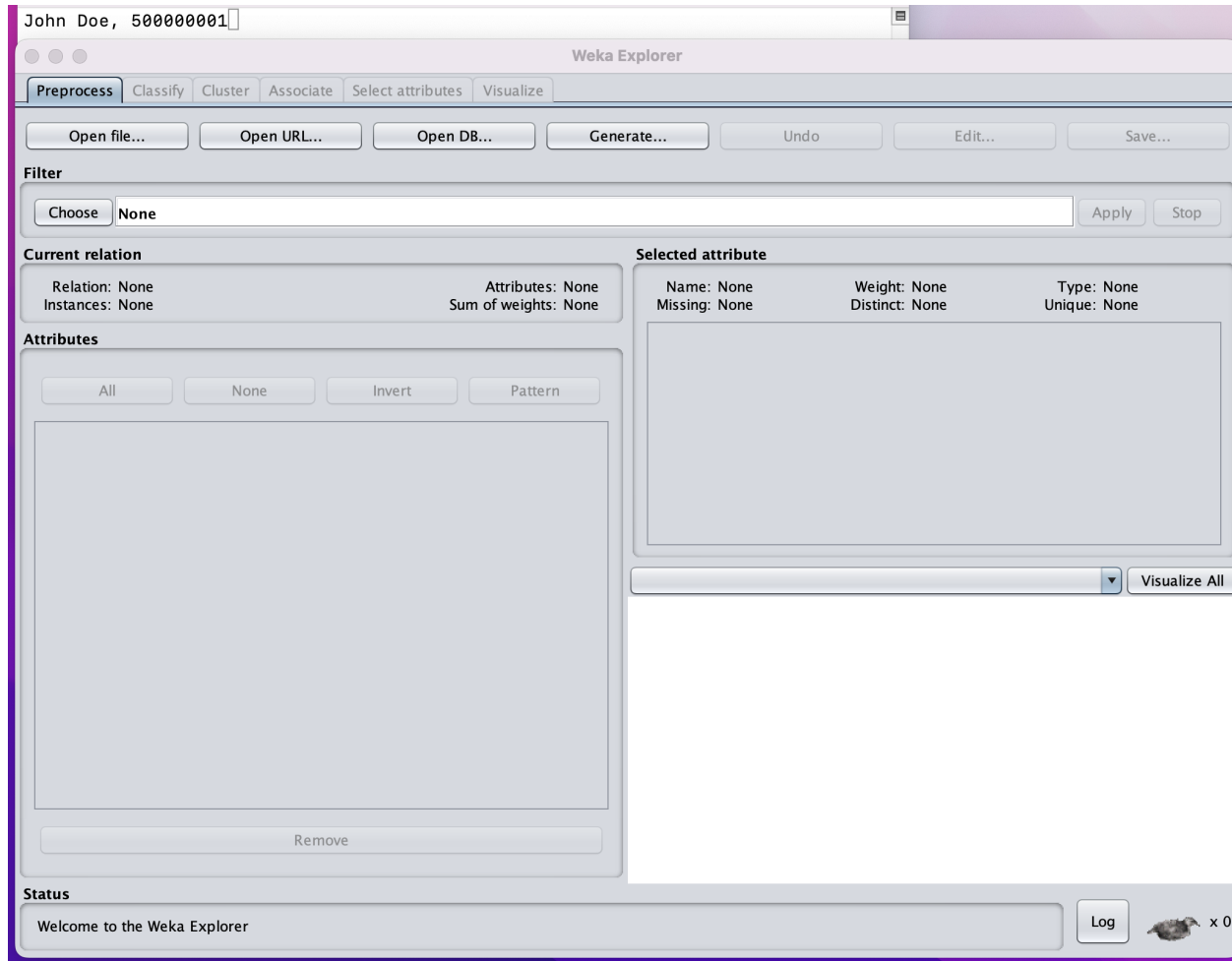
cluster0 cluster1

Extra credit (1%)

- ❖ Install WEKA and submit a captured Explorer screenshot to “WEKA install” under Assignments
 - Download: https://waikato.github.io/weka-wiki/downloading_weka/
 - Watch and try: <https://youtu.be/TFIyh5PKaql>
 - Weka manual is available in the course page
- ❖ Deadline: Posted on the course page – No grace period and partial points after this due date
 - Any extension/make-up/resubmission request will NOT be accepted



Extra credit – example capture



Submit a PDF document containing the capture!

Class Schedule

- ❖ Part 1: Introduction, exploratory analysis & statistics, input/output (Week 1-5)
- ❖ Midterm exam 1 (Week 6)
- ❖ Part 2: Regression, prediction, classification (Week 7-10)
- ❖ Midterm exam 2 (Week 11)
- ❖ Part 3: Statistical classification, unsupervised learning (Week 12-14)
- ❖ Part 4: Advanced topics – feature selection, projection (Week 15)
- ❖ Final exam (Week 16)



Grading

❖ Follow absolute scale:

- A = 90%-100%
- B = 80%-89%
- C = 70%-79%
- D = 60%-69%
- F = 59% or Below

❖ Components:

Components	Weight	Remarks
Assignments	30%	One lowest score will be dropped
Midterm exam	35%	Two exams: the lower score will be used
Final exam	35%	



Assignments

- ❖ Homework should be typed and submitted in a single PDF file (unless otherwise instructed)
 - Handwritten answers will not be accepted and graded
- ❖ On-time submission encouraged
 - The deadline for the assignment can be extended with a 15% penalty per day, up to two days (48 hours)
 - Any submission later than 48 hours after the deadline will not be accepted and graded.
- ❖ N assignments and the lowest one will be discarded from final grade calculation
- ❖ No make-up/extension/resubmission will be accepted and responded



Exams

- ❖ Three exams (online): 2 midterm and 1 final exam
- ❖ Tentative schedule is available on the course page
- ❖ Rescheduling of exams:
 - In case of time conflicts (e.g., with your work schedule), you can make a rescheduling request and take the exam one day earlier
 - **The rescheduled date cannot be anytime later than the regular exam date**
 - The rescheduling request should be received by the instructor by Friday (before the exam week) with a valid document
 - Any personal reason (e.g., family travel) will not be considered for rescheduling



Exam policy and format

- ❖ **Make-up policy:** Makeup exams will not be given for any reason. However, students will have two midterm exams, and the higher score will only be considered for the final grade calculation. If a student is unable to take the final exam for any emergency reason, the student may receive an 'X' (incomplete)
- ❖ **Exam format:** Due to increasing cheating incident reporting, this course employs the following format for online testing: (1) You can see one question per page, (2) you cannot proceed to the next question unless you answer the current question, and (3) it is not allowed to go back to the previous question. This should be strictly applied to minimize any possibility of academic dishonesty/misconduct



Rebuttal Window

- ❖ **REBUTTAL WINDOW:** For any piece of grading, the length of rebuttal window is limited to two weeks since the day when the grade is being posted. You can contact the instructor through email if you want to make a rebuttal request (but within the rebuttal window).



Academic Integrity

- ❖ Anyone attempting cheating will receive a zero on the work
- ❖ Subsequent cheating will result in a failing grade
 - May be reported to the university and the case filed officially
- ❖ Be aware that we have a high standard for academic integrity with zero-tolerance rule
- ❖ Use software tools such as turnitin and iThenticate



Plagiarism/Originality Check

Turnitin Document Viewer

https://submit.ac.uk/dv?s=1&o=15206419&u=632011&lang=en_us&session-id=798f685ecd4a25383b1354f4da7731c8

UN825: Educational Research Met... Submit UN825 essay here - DUE 12-Mar...

What's New Paper 4 of 4

Originality GradeMark PeerMark

Deconstructing research: A methodological critique of the

BY D. CLARK

turnitin 60% --

Match Overview

1	Submitted to Universit... Student paper	52%
2	Submitted to Universit... Student paper	1%
3	Submitted to Universit... Student paper	<1%
4	Bell, Doing Your Resea... Publication	<1%
5	Submitted to Universit... Student paper	<1%
6	www.isast.org Internet source	<1%
7	eprints.worc.ac.uk Internet source	<1%
8	Submitted to Universit... Student paper	<1%
9	Submitted to Universit... Student paper	<1%
10	Submitted to Universit... Student paper	<1%
11	Submitted to Universit... Student paper	<1%

is relatively new and the combination of two distinctly gender-aligned disciplines, habitually the feminine musician and the masculine technologist, makes for a new (and potentially interesting) area of investigation - a niche area that is largely untouched by researchers. To that end, as a first point of critique, the chosen area of analysis appears to hold relevance.

1

"...this project proposes to make suggestions towards a new inclusive recruitment strategy based upon the given findings, in an attempt to improve student recruitment and, where possible, aim to understand how to create a more inclusive environment for female students." (Clark 2009:1)

In reading the opening statement of the original project, there are some clear medium- to long-term objectives outlined. Such objectives were formed on the principle that "the aim [of research] is not only 'to know facts and to understand relations for the sake of knowledge. We want to know and understand in order to be able to act "better" than we did before' (Langeveld 1965: 4, as cited in Bell 2010: 27).

The aims of the project are addressed sufficiently, both in terms of perceived achievability and scope. However, the project states only the desired outcomes and not *how* those desired outcomes might be accomplished. However, it's important to recognise that the findings of this project would have potentially informed the 'how' as well as the 'why', as Bell reminds us that "whatever the size and scope of the study, you [the researcher] will in all cases be required to analyse and evaluate the information you collect and, in some cases, you might then be in a position to suggest desirable changes in practice." (2010: 28)

In general, it is clear that the project is attempting to, a) understand the reasons behind the apparent gender disparity, b) investigate the incidence of this phenomenon in other institutions, and c) attempt to nurture a more inclusive environment for female students.

It is now on this basis that this paper can begin to look in more depth at the epistemological issues raised by the original project.

Defining Concepts

For further details...

- ❖ Read the syllabus thoroughly!
- ❖ Course page for announcements, course materials, and grade posting
- ❖ **IMPORTANT: Syllabus can be changed over time!**
 - Will be announced for every change
 - Course page will keep the most recent copy of the syllabus



First week action items

- ❖ Read syllabus thoroughly
- ❖ Weka: install and run to earn extra credit!
- ❖ Watch “Getting Started with Weka - Machine Learning Recipes #10”
 - <https://www.youtube.com/watch?v=TFIyh5PKaql>

