Note: **This is the syllabus from Spring 2017. Spring 2018 syllabus is under revision and will include content on "Classificiation". Final verion of the Spring 2018 syllabus will be made available prior to start of Spring 2018 semester.**

CS 412: Introduction to Data Mining Course Syllabus

Course Description

This course is an introductory course on data mining. It introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on two major data mining functions: (1) pattern discovery and (2) cluster analysis.

In the first part of the course, which focuses on pattern discovery, you will learn why pattern discovery is important, what the major tricks are for efficient pattern mining, and how to apply pattern discovery in some interesting applications. The course provides you the opportunity to learn concepts, principles, and skills to practice and engage in scalable pattern discovery methods on massive data; discuss pattern evaluation measures; study methods for mining diverse kinds of frequent patterns, sequential patterns, and sub-graph patterns; and study constraint-based pattern mining, pattern-based classification, and explore their applications.

In the second part of the course, which focuses on cluster analysis, you will learn concepts and methodologies for cluster analysis, which is also known as clustering, data segmentation, or unsupervised learning. We will introduce the basic concepts of cluster analysis and then study a set of typical clustering methodologies, algorithms, and applications. This includes partitioning methods, such as k-means, hierarchical methods, such as BIRCH, density-based methods, such as DBSCAN, and grid-based methods, such as CLIQUE. We will also discuss methods for clustering validation. The learning will be enhanced by clustering software and programming assignments.

The technical contents of the course are based on the textbook *Data Mining: Concepts and Techniques* (3rd ed), as well as the on-campus course CS 412 – Introduction to Data Mining, which is offered in the Department of Computer Science at the University of Illinois. Please note several themes covered in the textbook are not covered in this online course, including (1) data preprocessing and preparation, (2) data warehouse and data cube technology, and (3) classification. This is because these themes have been covered or will be covered, with possible in-depth treatment, in several other courses offered in the Data Science Online Master program. Therefore, this course will focus on the in-depth study of the two major data mining functions illustrated above.

Course Goals and Objectives

Upon successful completion of this course, for pattern discovery, you will be able to:

- Recall important pattern discovery concepts, methods, and applications, in particular, the basic concepts of pattern discovery, such as frequent pattern, closed pattern, max-pattern, and association rules.
- Identify efficient pattern mining methods, such as Apriori, ECLAT, and FPgrowth.
- Compare pattern evaluation issues, especially several popularly used measures, such as lift, chi-square, cosine, Jaccard, and Kulczynski, and their comparative strengths.
- Compare mining diverse patterns, including methods for mining multi-level, multi-dimensional patterns, qualitative patterns, negative correlations, compressed and redundancy-aware top-k patterns, and mining long (colossal) patterns.
- Learn well-known sequential pattern mining methods, including methods for mining sequential patterns, such as GSP, SPADE, PrefixSpan, and CloSpan.
- Learn graph pattern mining, including methods for subgraph pattern mining, such as gSpan, CloseGraph, graph indexing methods, mining top-k large structural patterns in a single large network, and graph mining applications, such as graph indexing and similarity search in graph databases.
- Learn constraint-based pattern mining, including methods for pushing different kinds of constraints, such as data and pattern-based constraints, anti-monotone, monotone, succinct, convertible, and multiple constraints.
- Learn pattern-based classifications, including CBA, CMAR, PatClass, and DPClass.
- Enjoy various pattern mining applications, such as mining spatiotemporal and trajectory patterns and mining quality phrases.
- Explore further topics on pattern analysis, such as pattern mining in data streams, software bug mining, pattern discovery for image analysis, and privacy-preserving data mining.

For cluster analysis, you will be able to:

- Recall basic concepts, methods, and applications of cluster analysis, including the concept of clustering, the requirements and challenges of cluster analysis, a multi-dimensional categorization of cluster analysis, and an overview of typical clustering methodologies.
- Learn multiple distance or similarity measures for cluster analysis, including Euclidean and Minkowski distances; proximity measures for symmetric and asymmetric binary variables; distance measures between categorical attributes, ordinal attributes, and mixed types; proximity measures between two vectors – cosine similarity; and correlation measures between two variables – covariance and correlation coefficient.
- Learn popular distance-based partitioning algorithms for cluster analysis, including K-Means, K-Medians, K-Medoids, and the Kernel K-Means algorithms.
- Learn hierarchical clustering algorithms, including basic agglomerative and divisive clustering algorithms, BIRCH, a micro-clustering-based approach, CURE, which explores well-scattered representative points, CHAMELEON, which explores graph partitioning on the KNN Graph of the data, and a probabilistic hierarchical clustering approach.
- Learn the density-based approach to cluster analysis, which can group dense regions of arbitrary shape, such as DBScan and OPTICS.

- Learn the grid-based approach, which organizes individual regions of the data space into a grid-like structure, such as STING and CLIQUE.
- Study concepts and methods for clustering evaluation and validation by introducing clustering validation using external measures and internal measures, and the measures for evaluating cluster stability and clustering tendency.

Textbook and Readings

Although the lectures are designed to be self-contained, it is recommended (but not required) to reference the textbook: Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Waltham: Morgan Kaufmann.

You can download a PDF version of the chapters 1, 6, 7 and 2, 10, 11, 13 from *Data mining: Concepts and techniques* (3rd ed.) for free. Note that these are all the chapters related to the topics covered in this course, so the free PDF version of the chapters is sufficient for this course.

If you would like to purchase the entire textbook, the publisher has an exclusive offer just for Coursera students. You can save 30% on either the print or eBook version of *Data Mining: Concepts and Techniques, 3rd Edition* and receive free shipping on all orders. Here is how it works:

- Add the book to your cart.
- Enter code **COMP317** and click **Apply**.
- The discount will be applied to the list price and cannot be combined with other promotions.

Course Outline

This 4-credit hour course is 16 weeks long. You should invest **6–8** hours every week in this course.

The course is composed of two parts. Part 1 of the course, Week 1 to Week 9, focuses on pattern discovery. Part 2 of the course, Week 10 to Week 16, focuses on cluster analysis. All of the course content will be released on the first day of class, with the exception of the 2 proctored exams, which will not be released until the day of each exam (for more information on the proctored exams, read the section Elements of This Course below). Although all content (except for exams) is made available to the entire class on the first day, the course follows a schedule (see the table below).

Week	Duration	Topics

1	1/17–1/22	Course Orientation; Course Part 1 Pattern Discovery Overview; Pattern Discovery Basic Concepts; Efficient Pattern Mining Methods; Pattern Discovery Programming Assignment 1
2	1/23–1/29	Pattern Evaluation; Mining Diverse Frequent Patterns
3	1/30–2/5	Sequential Pattern Mining; Pattern Mining Applications: Mining Spatiotemporal and Trajectory Patterns
4	2/6–2/12	Constraint-Based Mining
5	2/13–2/19	Graph Pattern Mining
6	2/20–2/26	Pattern-Based Classification
7	2/27–3/5	Pattern Mining Applications: Mining Quality Phrases from Text Data; Advanced Topics on Pattern Discovery
8	3/6–3/12	Pattern Discovery Programming Assignment 2; Preparation for Part 1 Exam
9	3/13–3/19	Course Part 1 Exam on Pattern Discovery
10	3/20–3/26	Spring break
11	3/27–4/2	Course Part 2 Cluster Analysis Overview; Cluster Analysis Introduction; Similarity Measures for Cluster Analysis

12	4/3–4/9	Partitioning-Based Clustering Methods; Hierarchical Clustering Methods
13	4/10–4/16	Hierarchical Clustering Methods (continued); Density-Based and Grid-Based Clustering Methods; Cluster Analysis Programming Assignment 1
14	4/17–4/23	Methods for Clustering Validation; Cluster Analysis Programming Assignment 2
15	4/24–4/30	Preparation for Part 2 Exam
16	5/1–5/7	Course Part 2 Exam on Cluster Analysis

MOOC Version and CS 412 Content Mapping

If you have taken the MOOC version of the course, namely Pattern Discovery and Cluster Analysis, below is how the content in those two MOOCs maps to this course.

MOOC Equivalent	**CS 412**
Pattern Discovery MOOC	Week 1–3, 7, and 8
No MOOC equivalent	Week 4, 5, 6, and 9
Cluster Analysis MOOC	Week 11–14
No MOOC equivalent	Week 15 and 16

Assignment Deadlines

For all assignment deadlines, please refer to the **Course Assignment Deadlines, Late Policy, and Academic Calendar** page.

Elements of This Course

The course is comprised of the following elements:

- **Lecture Videos.** In each week, the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within the browser by clicking on their titles or download the videos. You may also download the slides that go along with the videos.
- **In-Video Questions**. Some lecture videos have questions associated with them to help verify your understanding of the topics. These questions will automatically appear while watching the video if you stream the video through your browser. These questions do not contribute toward your final score in the class.
- **Lesson Quizzes**. Each week may contain one or multiple lessons. A lesson is a series of videos on a certain topic, which concludes with a lesson quiz. You will be allowed 2 attempts for each quiz. There is no time limit on how long you take to complete each attempt at the quiz. Each attempt may present a different selection of questions to you. Your highest score will be used when calculating your final score in the class.
- **Programming Assignments.** There are 4 total programming assignments in this course – 2 are designed around the topic of pattern discovery and the other 2 on cluster analysis. For more information about the programming assignments, please read the instructions on programming assignment in respective weeks.
- **Proctored Exams**. There are 2 proctored exams in this class. The Part 1 Exam will be released during Week 9. The Part 2 Exam will be released during Week 16. Both exams will be proctored via a proctoring service called ProctorU. For more information about ProctorU and the proctor exams, read the **Proctored Exam** page.

Grading Distribution and Scale

Grading Distribution

Assignment	Frequency	Percentage Weight of Final Grade
Lesson Quizzes	17	17 x 2% per quiz = 34%

Programming Assignments (or MP)	4	4 x 4% per MP = 16%
Course Part 1 Exam	1	30%
Course Part 2 Exam	1	20%

Grading Scale

Letter Grade	**Percent Needed**	**Letter Grade**	**Percent Needed**	**Letter Grade**	**Percent Needed**
A+	**95%**	**B+**	**80%**	**C**	**65%**
A	**90%**	**B**	**75%**	**D**	**60%**
A-	**85%**	**B-**	**70%**	**F**	**Below 60%**

View Grades

You can view your grade on each assignment by clicking the **Assignments** tab on the left menu bar.