UNIVERSITY OF THE FRASER VALLEY

OFFICIAL UNDERGRADUATE COURSE OUTLINE (page 1)

COURSE IMPLEMENTATION DATE: January 2012
COURSE REVISED IMPLEMENTATION DATE: January 2013
COURSE TO BE REVIEWED: November 2017
(six years after UEC approval) *(month, year)*

OFFICIAL UNDERGRADUATE COURSE OUTLINE INFORMATION

Students are advised to keep course outlines in personal files for future use.
Shaded headings are subject to change at the discretion of the department – see course syllabus available from instructor

STAT 431	Mathematics & Statistics	3
COURSE NAME/NUMBER	FACULTY/DEPARTMENT	UFV CREDITS

Data Mining
COURSE DESCRIPTIVE TITLE

CALENDAR DESCRIPTION: Advances in data collection and computer storage technology have generated a very large volume of data sets in business, internet, medicine, and a variety of scientific fields. Traditional methods of statistical data analysis have been challenged. New methodologies and algorithms in Computer Science, Statistics, and Business Intelligence are then developed. Data mining provides the techniques of extracting useful information and hidden patterns from this massive amount of data. The main topics in this course are data exploration, classification, decision trees, Bayesian classifiers, frequent item sets, association rules, clustering, K-means, EM algorithm, and anomaly detection. Statistical software such as SAS will be used to implement the algorithms. Students are expected to complete a group project based on a large data set.

Note: This course is offered as STAT 431 (formerly MATH 431) and COMP 431. Students may take only one of these for credit.

PREREQUISITES: STAT 271, STAT 331/COMP 331, and CIS 230.
Revised prerequisite effective September 2015:
STAT 271, STAT 331/COMP 331, and COMP 230 (formerly CIS 230).

SYNONYMOUS COURSE(S):
(a) Replaces: **MATH 431**
(b) Cross-listed with: **COMP 431**
(c) Cannot take: **COMP 431** for further credit.

SERVICE COURSE TO: *(department/program)*

TOTAL HOURS PER TERM: 45

STRUCTURE OF HOURS:

Lectures:	35	Hrs
Seminar:		Hrs
Laboratory:	10	Hrs
Field experience:		Hrs
Student directed learning:		Hrs
Other (specify):		Hrs

TRAINING DAY-BASED INSTRUCTION:
Length of course:
Hours per day:

OTHER:
Maximum enrolment: **36**
Expected frequency of course offerings: Every other year
(every semester, annually, every other year, etc.)

WILL TRANSFER CREDIT BE REQUESTED? ☐ Yes ☒ No
TRANSFER CREDIT EXISTS IN BCCAT TRANSFER GUIDE: ☐ Yes ☒ No

Course designer(s): **David Chu, Paul Franklin**
Department Head: **Greg Schlitt** Date approved: **March 5, 2012**
Supporting area consultation (Pre-UEC) Date of meeting: **March 20, 2012**
Curriculum Committee chair: **Norm Taylor** Date approved: **April 20, 2012**
Dean/Associate VP: **Ora Steyn** Date approved: **May 4, 2012**
Undergraduate Education Committee (UEC) approval Date of meeting: **May 23, 2012**

LEARNING OUTCOMES:

Upon successful completion of this course, students will be able to:
1. use summary statistics to describe univariate and multivariate data sets;
2. apply different definitions of distances to measure the similarity and dissimilarity between data objects;
3. classify discrete and continuous data using decision trees;
4. employ different classifiers (rule-based, nearest-neighbor, naïve Bayes) to solve various classification problems;
5. generate various frequent itemsets and set up association rules in market basket analysis;
6. evaluate association patterns by interest factor;
7. use the K-means clustering technique to classify data in different clusters;
8. determine the correct number of clusters;
9. apply the EM algorithm to cluster data by estimating the parameters of a mixture model;
10. detect outliers in univariate and multivariate normal distributions;
11. complete a group project based on a large data set and present the findings.

METHODS: *(Guest lecturers, presentations, online instruction, field trips, etc.)*

Lectures, and use of computer software SAS or Data Mining SQL Server 2008.

METHODS OF OBTAINING PRIOR LEARNING ASSESSMENT RECOGNITION (PLAR):

☐ Examination(s) ☐ Portfolio assessment ☐ Interview(s)

☒ Other (specify): Course Challenge; see PLAR policy (94) at http://ufv.ca/secretariat/policies/

TEXTBOOKS, REFERENCES, MATERIALS: *[Textbook selection varies by instructor. Examples for this course might be:]*

Introduction to Data Mining by P. Tan et al., Pearson, 2006.
Data Mining---Concepts, Models, Methods and Algorithms (2nd edition) by M. Kantardzic, Wiley-IEEE Press, 2011.

SUPPLIES / MATERIALS:

STUDENT EVALUATION: *[An example of student evaluation for this course might be:]*

Assignments	15%
Test	20%
Project	25%
Final exam	40%

The above percentages may vary among instructors and years. The final exam is comprehensive. Students must obtain at least 40% on the final exam to pass the course.

COURSE CONTENT: *[Course content varies by instructor. An example of course content might be:]*

Introduction: The origins of data mining, data mining tasks.

Exploring data: Types of data, data quality, data preprocessing, measures of similarity and dissimilarity, Minkowski distance, summary statistics, visualization, OLAP and multi-dimensional data analysis.

Classification: Decision tree induction, model over-fitting, evaluating the performance of a classifier, methods for comparing classifiers, rule-based classifiers, nearest-neighbor classifiers, Bayesian classifiers, artificial neural network, support vector machine, ensemble methods.

Association analysis: Frequent itemset generation, rule generation, maximal frequent itemsets, closed frequent itemsets, FP-Growth algorithm, evaluation of association patterns, interest factor, handling categorical and continuous attributes, sequential patterns.

Cluster analysis: Different types of clusters, K-means, bisecting K-means, agglomerative hierarchical clustering, density-based clustering algorithm, determining the correct number of clusters, supervised measures of cluster validity, fuzzy clustering, EM algorithm, self-organizing maps, minimum spanning tree clustering.

Anomaly detection: Causes of anomalies, determining outliers in univariate and multivariate normal distributions, proximity-based outlier detection, density-based outlier detection, clustering-based techniques.