University of Alberta
Department of Mathematical and Statistical Sciences

STAT 441/505 Q1 Applied Statistical Methods for Data Mining
WINTER 2016

Instructor:	Linglong Kong
Office:	CAB 431
Phone:	(780) 492-0205
E-mail:	lkong@ualberta.ca
Personal Web Page:	www.math.ualberta.ca/∼lkong

Office Hours:	MWF 10:00 - 11:20 or by appointment
Lecture Room and Time:	V 103, MWF 09:00 - 09:50
Laboratory Room and Time:	CAB 345, W 12:00 - 12:50
Course Web Page:	eclass/Moodle

Course Description:
Principles of statistical model building and analysis applied in linear and generalized linear models and illustrated through multivariate methods such as repeated measures, principal components, and supervised and unsupervised classification.

Course Prerequisites:
STAT 368 or 378. Knowledge on regression analysis is highly recommended.

Course Objectives and Expected Learning Outcomes:
Introduction to linear regression. Supervised classification by logistic regression, linear and quadratic discriminant analysis, and support vector machine. Model selection including classical methods and shrinkage methods, say, LASSO. Nonparametric methods including basis expansion and kernel smoothing method. Model assessment and inference. Tree-based methods including decision trees, bagging, boosting, and random forests. Neural networks and deep learning. Exploratory analysis or unsupervised learning including cluster analysis, principal component analysis, independent component analysis, factor analysis, canonical correlation analysis, and other dimensional reduction techniques, say, multidimensional scaling.

Required Textbook:
The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition), by T. Hastie, R. Tibshirani, J. H. Friedman. 2009, Springer.

Recommended or Optional Learning Resources:
An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Statistical Learning with Sparsity: The Lasso and Generalizations, by T. Hastie, R. Tibshirani, M. Wainwright. 2015, Chapman and Hall/CRC.
Applied Multivariate Statistical Analysis (6th edition), by R. A. Johnson and D. W. Wichern. 2007, Pearson.

Software:
The only software we are going to use is R. This software is free and you can download it from: http://www.r-project.org/.

Grade Evaluation:
The course mark will be calculated based on the following breakdown:

Course Component	**Weight of Total Mark**	**Date**
Assignments	30%	See below
Midterms	30% = 15% + 15%	M Feb 1 & Mar 7 09:00 - 09:50
Final Exam	40%	M Apr 18 09:00 - 11:00

Note: The date of the final examination is set by the Registrar and takes precedence over the final examination date reported in this document. Students must verify this date on BearTracks when the Final Exam Schedule is posted.

The final letter grade will be determined from the course mark as follows: An overall course mark of 50% or more guarantees a passing grade of at least D (STAT 441). An overall course mark of 65% or more guarantees a passing grade of at least C+ (STAT 505). An overall course mark of 90% or more guarantees a grade of at least A.

Grades are unofficial until approved by the Department and/or Faculty offering the course.

Assignments:
There are totally four equal weighted assignments, which should be handed to the instructor by the end of the due date.

Assignment No.	**Weight of Total Mark**	**Due Date**
Assign 1	10%	F Jan 29
Assign 2	10%	F Mar 4
Assign 3	10%	F Apr 8

General Comments:
Lecture notes, codes and data sets will be posted on the course website. Assignments will be posted on the course website and announced in the class.

Excused Absence Where the Cause is Religious Belief:
For an excused absence where the cause is religious belief, a student must contact the instructor(s) within two weeks of the start of Fall or Winter classes to request accommodation

for the term (including the final exam, where relevant). Instructors may request adequate documentation to substantiate the student request.

Missed Term Work
A student who cannot write a midterm or complete a term assignment due to incapacitating illness, severe domestic affliction or other compelling reasons can apply for an excused absence. To apply for an excused absence, a student must inform the instructor within two working days following the scheduled date of the term work or term exam missed, or as soon as the student is able, having regard to the circumstances underlying the absence. In all cases, instructors may request adequate documentation to substantiate the reason for the absence at their discretion.

An excused absence is a privilege and not a right; there is no guarantee that an absence will be excused. Misrepresentation of Facts to gain an excused absence is a serious breach of the *Code of Student Behaviour*.

Missed Final Examination:
A student who cannot write the final examination due to incapacitating illness, severe domestic affliction or other compelling reasons can apply for a deferred final examination. Students who failed at the start of term to request exam accommodations for religious beliefs are expected to follow the normal deferred final examination process. Such an application must be made to the student's Faculty office within two working days of the missed examination and must be supported by a Statutory Declaration (in lieu of a medical statement form) or other appropriate documentation (Calendar section 23.5.6). Deferred examinations are a privilege and not a right; there is no guarantee that a deferred examination will be granted. Misrepresentation of Facts to gain a deferred examination is a serious breach of the *Code of Student Behaviour*.

Any deferred final examinations are scheduled as follows:

Date: Saturday May 7, 2016
Time: 9:00 am
Location: CAB 357

Students should meet outside CAB 357 to register for their exam at 8:30 am; the exams will begin at 9:00 am.

Re-examination: (STAT 441 only)
A student who writes the final examination and fails the course may apply for a re-examination. Re-examinations are rarely granted in the Faculty of Science. These exams are governed by University (Calendar section 23.5.5) and Faculty of Science Regulations (Calendar section 192.5.3). Misrepresentation of Facts to gain a re-examination is a serious breach of the *Code of Student Behaviour*.

STUDENT RESPONSIBILITIES

Academic Integrity:
The University of Alberta is committed to the highest standards of academic integrity and honesty. Students are expected to be familiar with these standards regarding academic honesty and to uphold the policies of the University in this respect. Students are particularly urged to familiarize themselves with the provisions of the *Code of Student Behaviour* (online at www.governance.ualberta.ca) and avoid any behaviour which could potentially result in suspicions of cheating, plagiarism, misrepresentation of facts and/or participation in an offence. Academic dishonesty is a serious offence and can result in suspension or expulsion from the University.

All forms of dishonesty are unacceptable at the University. Any offense will be reported to the Senior Associate Dean of Science who will determine the disciplinary action to be taken. Cheating, plagiarism and misrepresentation of facts are serious offenses. Anyone who engages in these practices will receive at minimum a grade of zero for the exam or paper in question and no opportunity will be given to replace the grade or redistribute the weights. As well, in the Faculty of Science the sanction for **cheating** on any examination will include **a disciplinary failing grade** (NO EXCEPTIONS) and senior students should expect a period of suspension or expulsion from the University of Alberta.

Collaboration on Assignments:
Every term there are several students who receive academic penalties for copying assignments. Here are some tips to avoid copying on assignments:

1. Do not write down something that you cannot explain to your TA or instructor.
2. When you are helping other students, avoid showing them your work directly. Instead, explain your solution verbally. Students whose work is copied also receive academic sanctions.
3. If you find yourself reading another student's solution, do not write anything down. Once you understand how to solve the problem, remove the other person's work from your sight and then write up the solution to the question yourself. Looking back and forth between someone else's paper and your own paper is almost certainly copying and will result in academic sanctions for both you and your fellow student.
4. If the instructor or TA writes down part of a solution in order to help explain it to you or the class, you cannot copy it and hand it in for credit. Treat it the same way you would treat another student's work with respect to copying, that is, remove the explanation from your sight and then write up the solution yourself.
5. There is often more than one way to solve a problem. Choose the method that makes the most sense to you rather than the method that other students happen to use. If none of the ideas in your solution are your own, there is a good chance it will be flagged as copying.

You are recommended to read the Appropriate Collaboration link on the Office of Student Judicial Affairs website (http://www.osja.ualberta.ca/Students/AppropriateCollaboration.aspx).

Exams:
Your student photo I.D. is required at exams to verify your identity. Students will not be allowed to begin an examination after it has been in progress for 30 minutes. Students must remain in the exam room until at least 30 minutes has elapsed. Electronic equipment cannot be brought into examination rooms.

Cell Phones:
Cell phones are to be turned off during lectures, labs and seminars. Cell phones are not to be brought to exams.

Audio or Video Recording:
Audio or video recording, digital or otherwise, of lectures, labs, seminars or any other teaching environment by students is allowed only with the prior written consent of the instructor or as a part of an approved accommodation plan. Student or instructor content, digital or otherwise, created and/or used within the context of the course is to be used solely for personal study, and is not to be used or distributed for any other purpose without prior written consent from the content author(s).

Students Eligible for Accessibility-Related Accommodations (students registered with Student Accessibility Services – SAS):
Eligible students have both rights and responsibilities with regard to accessibility-related accommodations. Consequently, scheduling exam accommodations in accordance with SAS deadlines and procedures is essential. Please note adherence to procedures and deadlines is required for U of A to provide accommodations. Contact SAS (www.ssds.ualberta.ca) for further information.

Student Success Centre:
Students who require additional help in developing strategies for better time management, study skills, or examination skills should contact the Student Success Centre (2-300 Students Union Building).

Policy about course outlines can be found in section 23.4(2) of the University Calendar.

Disclaimer:
Any typographical errors in this Course Outline are subject to change and will be announced in class.

Copyright:

Tentative course outline and important dates				
Lec	**Date**	**Reading**	**Course materials**	**Comments**
1	M Jan 4	chapter 1	Introduction	
2	W Jan 6	chapter 2	Statistical machine learning overview	
3	F Jan 8	chapter 3	Linear regression I	
4	M Jan 11	chapter 3	Linear regression II	
5	W Jan 13	chapter 3	Model selection I	
6	F Jan 15	chapter 3	Model Selection II	
7	M Jan 18	chapter 4	Logistic regression I	
8	W Jan 20	chapter 4	Logistic regression II	
9	F Jan 22	chapter 4	Linear discriminant analysis I	
10	M Jan 25	chapter 4	Linear discriminant analysis II	
11	W Jan 27	chapter 12	Support vector machine I	
12	F Jan 29	chapter 12	Support vector machine II	**Assign 1 Due**
	M Feb 1	**Midterm Exam**		
13	W Feb 3	chapter 5	Nonparametric methods I	
14	F Feb 5	chapter 6	Nonparametric methods II	
15	M Feb 8	chapter 7	Model assessment	
16	W Feb 10	chapter 8	Model inference	
17	F Feb 12	chapter 9	Tree-based methods I	
18	M Feb 22	chapter 9	Tree-based methods II	
19	W Feb 24	chapter 10	Tree-based methosd III	
20	F Feb 26	chapter 15	Tree-based methods IV	
21	M Feb 29	chapter 11	Neural networks I	
22	W Mar 2	chapter 11	Neural networks II	
23	F Mar 4	chapter 11	Deep Learning	**Assign 2 Due**
	M Mar 7	**Midterm Exam**		
24	W Mar 9	chapter 13	Cluster analysis I	
25	F Mar 11	chapter 14	Cluster analysis II	
26	M Mar 14	chapter 14	Principal component analysis I	
27	W Mar 16	chapter 14	Principal component analysis II	
28	F Mar 18	chapter 14	Independent component analysis I	
29	M Mar 21	chapter 14	Independent component analysis II	
30	W Mar 23	chapter 14	Factor analysis I	
31	W Mar 30	chapter 14	Factor analysis II	
32	F Apr 1	chapter 14	Canonical correlation analysis I	
33	M Apr 4	chapter 14	Canonical correlation analysis I	
34	W Apr 6	chapter 14	Multidimensional scaling I	
35	F Apr 8	chapter 14	Multidimensional scaling II	**Assign 3 Due**
	M Apr 18	**09:00 - 11:00 Final Exam**		

Please fill out this page and return it to me

Name:

Degree Program:

Please list the STAT and MATH courses you have previously taken. Include the names or topics of the courses, if they were not taken here.

Please list the STAT and MATH courses you are taking this year.

Why are you taking this course?