Rogers + Girolami, 2nd ed., A First Course in Machine Learning

DS 303X: Concepts and Applications of Machine Learning. Cr. 3-0. F.S.
(To be submitted, Fall, 2017)
Prerequisites: *STAT 347*

Description:
Machine learning concepts such as training and test sets; feature extraction; principles of machine learning techniques; regression; classical pattern recognition methods; advanced topics in pattern recognition; unsupervised learning techniques; assessment and diagnostics: overfitting, error rates, residual analysis, model assumptions checking, feature selection; communicating findings to stakeholders in written, oral, verbal and electronic form, and ethical issues in data science.

Course outcomes/objective:
Upon successful completion of the course, students will have an understanding of the key concepts of machine learning and learn a variety of machine learning-related concepts such as linear models, nonparametric regression, logistic regression, and support vector machines. Students should be able to:

1. Define key concepts in machine learning such as training and test sets, and feature extraction
2. Apply machine learning techniques to solve data science problems
3. Evaluate and diagnose learned data models
4. Understand ethical concerns in data science
5. Communicate the output of data analysis pipelines to stakeholders
6. Function on multi-disciplinary teams

Course content/major topics to be addressed:
Major topics include:

- Introduction – What is machine learning? Motivating case studies.
  - Basic machine learning concepts: training and test sets, feature extraction
- Principles of machine learning techniques
  - Regression
    - Ordinary least squares
    - Inference for linear regression
    - Robustness
    - Beyond linearity: nonparametric regression, splines, local polynomial regression
  - Classical pattern recognition methods and their assumptions
    - Logistic regression
    - Bayes theorem (LDA, QDA, Naive Bayes)
    - K-nearest neighbors
  - Advanced topics in pattern recognition
    - Support vector machines and its variants
    - Tree based methods (Bagging, Random forest, Boosting)
  - Unsupervised learning
    - Principal component analysis
    - Clustering (k-means, hierarchical)
- Model assessment and diagnostics
  - overfitting, error rates, residual analysis, model assumptions checking
  - introduction to feature selection
- Selected applications in various domains
- Ethical issues in data science

New Experimental Course Proposal

DS 303 (3-0)

Concepts and Applications of Machine Learning

Instruction Type (Hours per Week): Lecture 3; Lab 0

Prerequisites:
Math 207, Math 265, STAT 347

Description:
Machine learning concepts such as training and test sets; feature extraction; principles of machine learning techniques; regression and related topics such as ordinary least squares, inference for linear regression, robustness, beyond linearity: nonparametric regression, splines, local polynomial regression; classical pattern recognition methods such as logistic regression, Bayes theorem (LDA, QDA, naive Bayes); K-nearest neighbors; advanced topics in pattern recognition such as support vector machines and its variants, tree based methods (Bagging, Random forest, Boosting); unsupervised learning techniques such as principal component analysis, clustering (k-means, hierarchical); assessment and diagnostics: overfitting, error rates, residual analysis, model assumptions checking, feature selection; communicating findings to stakeholders in written, oral, verbal and electronic form, and ethical issues in data science.

Reason for proposal:

With the explosion of big data problems and data science, machine learning has become a critical subject for many scientific areas (computer science, engineering, etc.) as well as marketing, finance, and other business disciplines. People with machine learning skills are in high demand. Hence, it is extremely important to provide data science undergraduate students with a basic knowledge and tools to analyze data sets and perform model checking.

This upper-level course in data science is built as one of the core courses for the Data Science Undergraduate program. There is a large demand for data science professionals today both in Iowa and nationally. Machine learning and statistics provide the technical basis of data science. The course aims to provide an introduction to the concepts and techniques of machine learning and statistics used in practical data analysis.

Course outcomes/objective:

Upon successful completion of the course, students will have an understanding of the key concepts of machine learning and learn a variety of machine learning-related concepts such as linear models, nonparametric regression, logistic regression, and support vector machines. Students should be able to:

1. Define key concepts in machine learning such as training and test sets, and feature extraction
2. Apply machine learning techniques to solve data science problems
3. Evaluate and diagnose learned data models
4. Understand ethical concerns in data science
5. Communicate the output of data analysis pipelines to stakeholders
6. Function on multi-disciplinary teams

Course content/major topics to be addressed:

Students will learn the concepts and applications of machine learning techniques. They will write programs to solve problems in hands-on Data science projects using machine learning and data modeling techniques. Major topics include:

- Introduction – What is machine learning? Motivating case studies. – 1 lecture
- Basic machine learning concepts - 1 lecture
    - training and test sets
    - feature extraction

- Principles of machine learning techniques
    - Regression - 6 lectures
        - Ordinary least squares
        - Inference for linear regression
        - Robustness
        - Beyond linearity: nonparametric regression
            - Splines
            - Local polynomial regression
    - Classical pattern recognition methods and their assumptions – 5 lectures
        - Logistic regression
        - Bayes theorem (LDA, QDA, Naive Bayes)
        - K-nearest neighbors
    - Advanced topics in pattern recognition – 5 lectures
        - Support vector machines and its variants
        - Tree based methods (Bagging, Random forest, Boosting)
    - Unsupervised learning – 6 lectures
        - Principal component analysis
        - Clustering (k-means, hierarchical)
- Model assessment and diagnostics – This is a crosscutting topic, most of which is first introduced during the regression module, and then revisited throughout this course.
    - overfitting
    - error rates
    - residual analysis
    - model assumptions checking
    - introduction to feature selection
- Selected applications in various domains – Examples and applications would be used throughout this course to concretely explain concepts.
- Ethical issues in data science - This is a crosscutting topic, most of which is first introduced during the introductory module, and then revisited throughout this course.

Assessment Plans:

Student performance will be measured via programming assignments, laboratory exercises, written and programming exams, and a term project. The content of all of the assignments above are continuous and support each other. The weights of these assessment mechanisms are as follows:

- Problem sets (25%)
- Laboratory assignments involving using machine learning software (25%)
- A term project requiring written report and oral presentation (25%),
- One exam (25%)

Relationship of this course to existing courses in other departments and programs (supporting, overlap, etc.):

This course may have small overlap on survey of the basic machine learning techniques with COM S 474 Introduction to Machine Learning and IE 483 Knowledge Discovery and Data Mining. COM S 474 focuses on machine learning techniques and mainly targets students in computer science and related disciplines. IE 483 focuses on algorithm techniques that can be used for data mining tasks in manufacturing and service industries. This course targets students from any disciplines interested in Data Science. Students could further their knowledge in machine learning and data mining by taking COMS 474 or IE 483 after taking this course.

This course will also serve as a core course for the Undergraduate Data Science Major.