CS722/822: Machine Learning

Instructor: Jiangwen Sun

Computer Science Department

Course Information

- Instructor: Dr. Jiangwen Sun
  - Office: E&CS 3204
  - Phone: (757) 683-7712
  - Email: jsun@odu.edu
  - Web: https://www.cs.odu.edu/~jsun

- Blackboard
  - https://www.blackboard.odu.edu
  - Login with your MIDAS ID and password

- TA: TBD

Jiangwen Sun, Ph.D.

- Ph.D. in Computer Science & Engineering
- B.S. in Clinical Medicine
- Research interests: machine learning, computational systems medicine, data mining, medical informatics and health informatics
- Office hours: W 11:00am-12:00pm (other times by appointment)

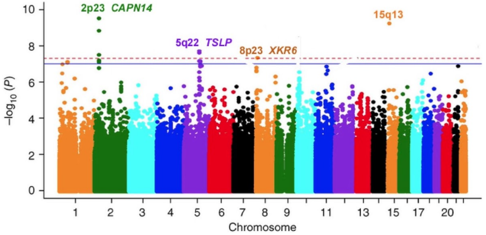
More on My Research

Disease Classification

- Reflect characteristics in both clinical presentation and etiology

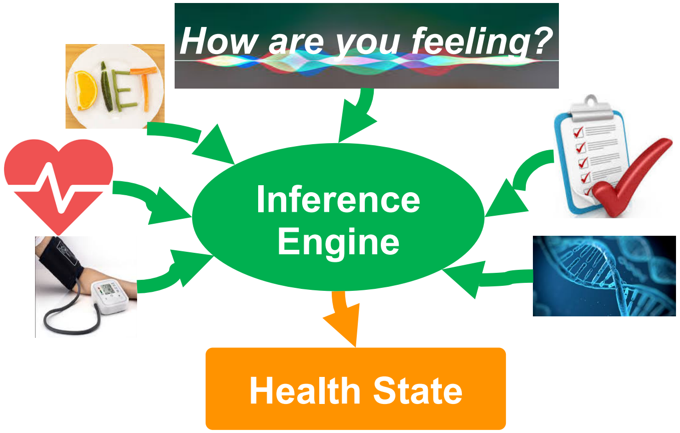
Phenome-Genome Association

- Identifying genetic factors
- Genome prediction



Machine Learning & Big Data Analytics

Artificial Physician

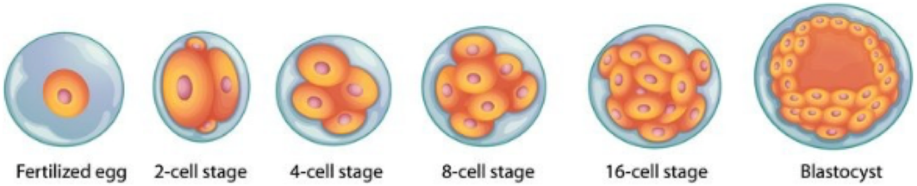


Inference Engine

Health State

Understand Human Embryo Development

- Difference across developmental stages
- Developmental stage regulation

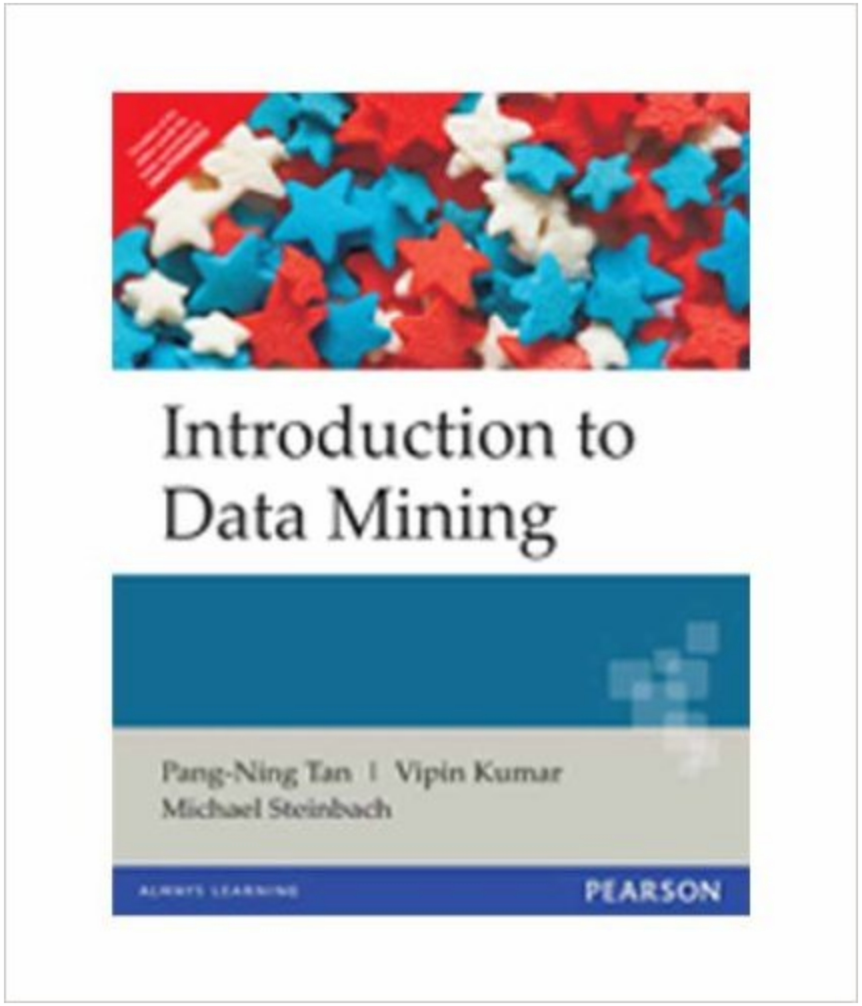


Course Information

- Prerequisite: Basic linear algebra, basic probability, calculus, optimization and basics of programming (Python, or MATLB, or R)
- Objectives:
  - Master basic concepts of machine learning, such as overfitting, different kinds of learning problems
  - Understand the foundations of some commonly-used machine learning methods and algorithms
  - Get informed of the state of the art in the field
  - Get familiar enough with some machine learning methods in solving practical problems
- Format:
  - Lectures, Paper reviews, In-class quizzes, Homework assignments, A final quiz, and A term project

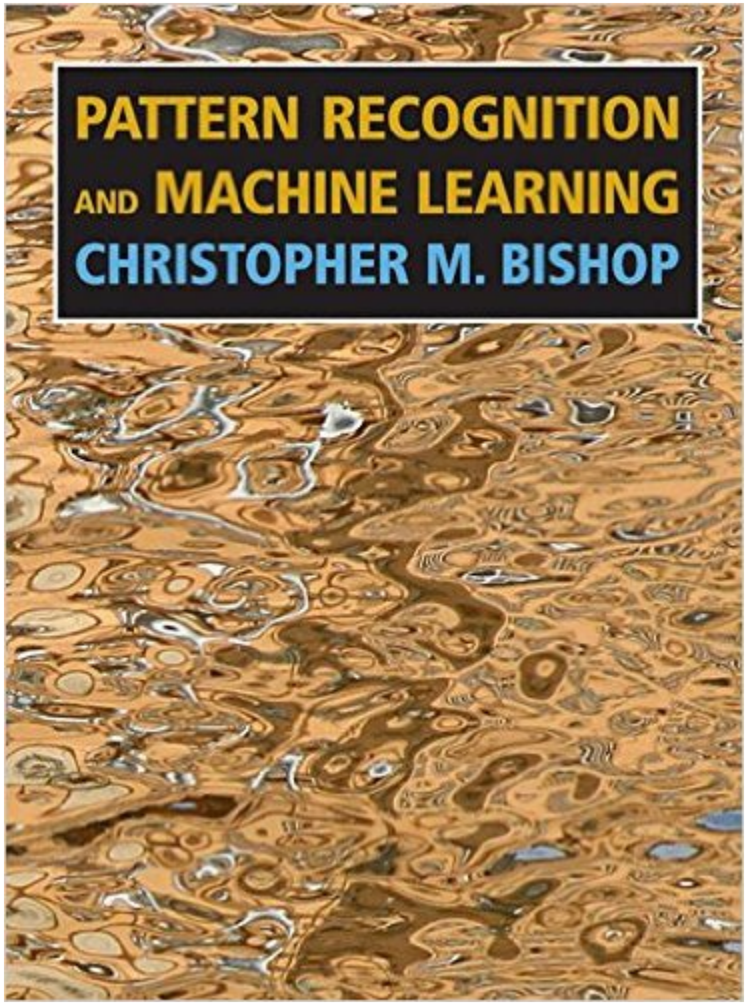
Optional textbooks

- Course textbook (not required):
  - *Introduction to Data Mining* (2005) by Pang-Ning Tan, Michael Steinbach, Vipin Kumar



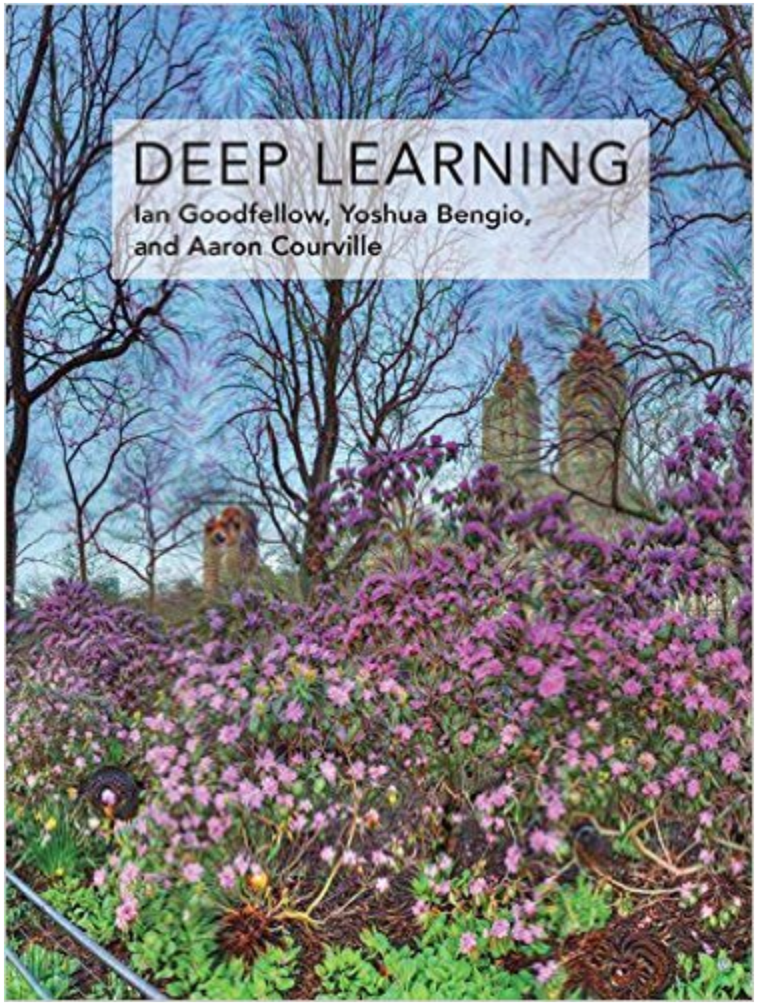
Optional textbooks

- Course textbook (not required):
  - *Pattern Recognition and Machine Learning* (2006) Christopher M. Bishop



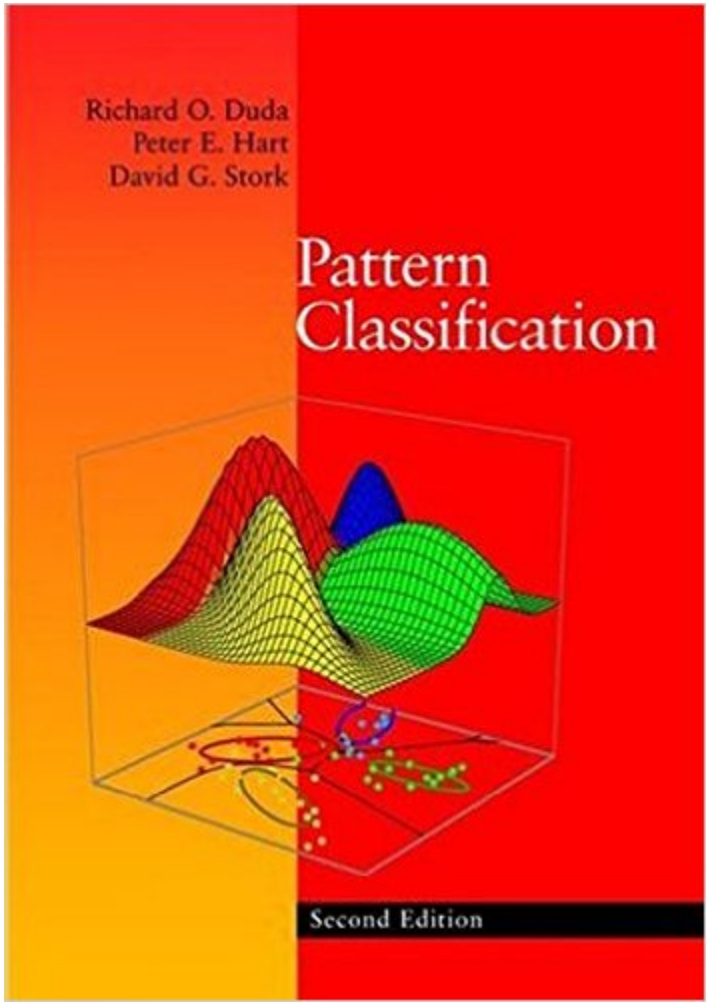
Optional textbooks

- Course textbook (not required):
  - Deep Learning (2016) Ian Goodfellow, Yoshua Bengio, and Aaron Courville



Optional textbooks

- Course textbook (not required):
  - *Pattern Classification* (2nd edition, 2000) Richard O. Duda, Peter E. Hart and David G. Stork



Reading materials

- Additional class notes and copied materials will be given
- Reading material and paper links will be provided

Grading Policy

Close-book quizzes (3 in-class, and one final):	40%
Paper review and presentation (1):	10%
Non-programming-based HW (2):	5%
Programming-based HW (2-3):	15%
Term project (1):	30%

- Study group is encouraged. A group can consist of two to three students only. Each group works on the review of one paper and on a term project, and team members will receive the same grade for the paper review and term project. However, each student in the team needs to specify his/her roles in the project.
- All term projects will involve programming of certain recent algorithms published in high-quality machine learning venues.
- Choices: (1) from recent machine learning venules; or
  (2) from a list prepared by the instructor

Grading Policy

- Participation in paper reviews is very important, and presenting the paper itself accounts for 50% of the credits for paper review, the other half is judged by the instructor and/or other students
- Quizzes and HWs are graded by TA
- For programming-based HWs, students will be required to turn in the final ML model besides their codes. TA will rank the performance of the final models and grades will be given accordingly
- Final term projects will be graded by both TA and the instructor
- If you miss a quiz due to a qualified reason, there may be a take-home quiz to make up the credits; otherwise there is no make-up plan

Grading Policy - paper review

- Each study group will be responsible of teaching one specific paper published in the following venues in the past **three** years:
  - International Conference on Machine Learning (**ICML**)

    e.g., 2018 (https://icml.cc/)
  - Advances in Neural Information Processing System (**NIPS**)

    e.g., 2017 (https://nips.cc/)
  - ACM **SIGKDD** Conference on Knowledge Discovery and Data Mining

    e.g., 2018 (http://www.kdd.org/kdd2018/accepted-papers)

Term Project

- Each team implements the algorithms proposed in a paper that the team members presented during paper review
- Each team needs to present their results
- Each team needs to submit codes
- Each team needs to submit a project report
  - Recite the algorithms using your own language
  - Any details in your implementation and computational results
  - Conclusion (success or failure to replicate the results in the paper, if failed, why?)
- Extra credit
  - Apply the algorithm to additional datasets
  - Comparing with more other relevant algorithms
  - Modify the algorithms to address new problems
  - … …

Term Project

- The term project will require significant efforts, so start the earlier the better
- A lot of machine learning papers accompany with codes that implement their algorithms. If you find the associated codes, you can use them as reference but the implementation should be your own programming.

What is Machine Learning

- The ultimate goal of machine learning is the creation and understanding of machine intelligence
- Teach a computer to learn concepts using data or interactions with an environment– without being explicitly programmed
- Machine learning is a type of artificial intelligence that provides computers with the ability to learn. It focuses on the development of computer programs that can change when exposed to new data

Please give me daily-life examples of ML

Traditional Topics in AI

- Fuzzy set and fuzzy logic
  - Fuzzy if-then rules
- Evolutionary computation
  - Genetic algorithms
  - Evolutionary strategies
- Artificial neural networks
  - Back propagation network (supervised learning)
  - Self-organization network (unsupervised learning, will not be covered)

Modern topics in ML

- Data representation: find the best representation of data, especially complex data
- Feature selection: select the most relevant features for a target problem from a massive amount of variables
- Incomplete data: practical data almost always have missing entries, low-confidence entries
- Integrative modeling: deal with multiple kinds of data – videos, texts, audios
- Large scale: design scalable machine learning algorithms for big data

Topics to be covered

- Introduction of machine learning problems, basic concepts, review of basics of probability, basics of linear algebra
- Supervised learning
  - Regression (least squares regression, linear regression, polynomial regression, overfitting, ridge regression, LASSO, gradient descent, stochastic gradient descent)
  - Classification (k-nearest neighbor, logistic regression, ROC curve, evaluation metrics)

Topics to be covered

- Unsupervised learning
  - Cluster analysis (k-means, hierarchical clustering, DBSCAN, etc.)
  - Dimension Reduction (PCA, CCA, etc.)
- Important specific methods
  - Spectral clustering
  - Support vector machine
  - Neural networks (shallow multilayer perceptron and back propagation algorithm)
  - Deep learning (deep layers of neural networks)

Basic concepts in ML

- Different learning problems/tasks: unsupervised learning versus supervised learning, more subtle categories, such as reinforcement learning, semi-supervised learning, active learning, online learning
- Different supervised learning problems: Classification versus regression
- Mathematical representation of data: features or variables, samples or examples
  - Training/testing examples
- Representation of a model: hypothesis space

Supervised versus unsupervised

- Given a collection of examples (*training set*)
  - Each example contains a set of *features (independent variables)*, and a supervision label that serves the target or dependent variable.

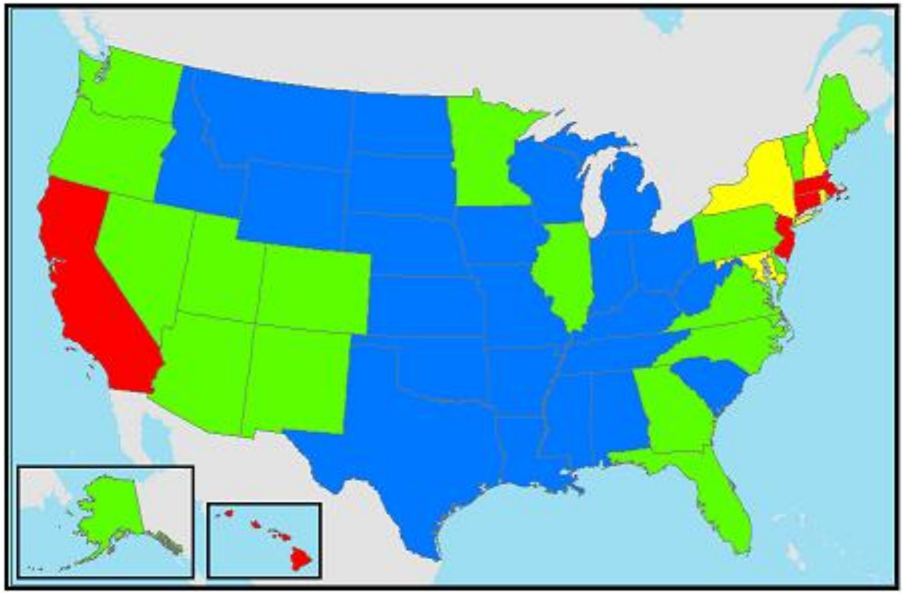
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Loss**
1	Yes	Single	125K	**100**
2	No	Married	100K	**120**
3	No	Single	70K	**-200**
4	Yes	Married	120K	**-300**
5	No	Divorced	95K	**-400**
6	No	Married	60K	**-500**
7	Yes	Divorced	220K	**-190**
8	No	Single	85K	**300**
9	No	Married	75K	**-240**
10	No	Single	90K	**90**

For instance, loss is the target, and each transaction record contains three features, what is the relationship between the features and the loss, can we use the features to compute loss

- **Supervised learning goal**: find a *model or a mapping* that best maps from the features to the target.

Supervised versus unsupervised

- Given a collection of examples, each consisting a set of features, but without a clear target



For instance, cluster analysis using lodging data of 50 states

Each state record contains median rent, and median home value

- **Unsupervised learning goal**: draw inferences from the data for exploratory analysis, such as finding hidden patterns or grouping of examples

In-class practice

Please judge if the following problems are supervised or unsupervised learning problems

- Given data about the size of houses on the real estate market, try to predict their price
- For a collection of 10,000 patients, find a way to automatically group them into subgroups that are similar by different variables, such as lifespan, location, race, etc
- Given a picture of a face, predict which person’s face it is
- Given a patient with a tumor, predict whether the tumor is malignant or benign
- Given a recoding of sound recorded from a party, try to identify if the sound comes from one source, or different sources, and try to segment the different sources

More recent learning problems

- Semi-supervised learning
  - A large dataset containing a small portion of labeled data but a majority of data without labels
- Active learning
  - When an oracle (an expert or a process) in the loop who can provide labels for a limited number of example at a time
- Online machine learning
  - When data becomes available in a sequential order, so at each step update the current best model
- Reinforcement learning
  - Learn how to take actions in an environment so as to maximize some notion of cumulative reward (label)