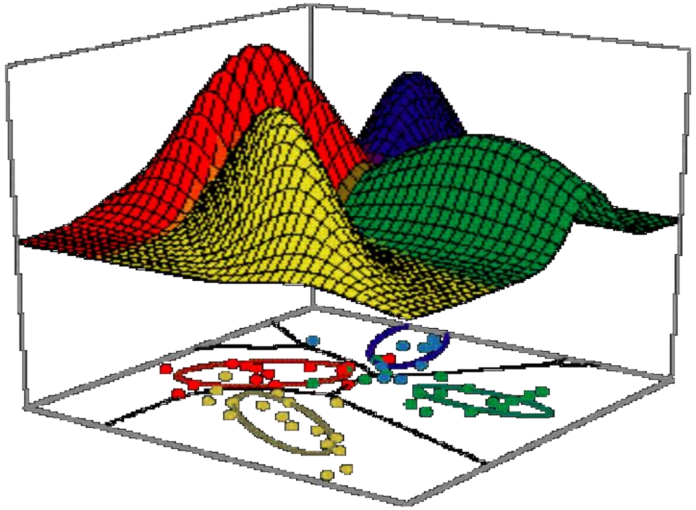
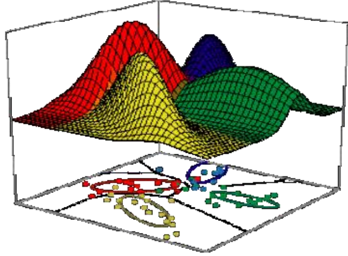
SYSC5405/BIOM5405



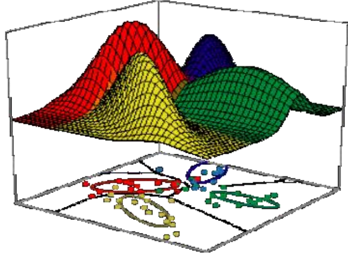
Pattern Classification and Experiment Design

Instructor



- **Prof. James Green, P.Eng.**
- **Web:** http://www.sce.carleton.ca/faculty/green
- **Office:** 6203 Canal Building
- **Email:** jrgreen@sce.carleton.ca
  - Please include "SYSC5405/BIOM5405:" in the subject to ensure that your email is not missed.
  - Emails should only be sent from your Cmail account - all other emails may be ignored at the instructor's discretion.
- **Office hours:** By appointment.

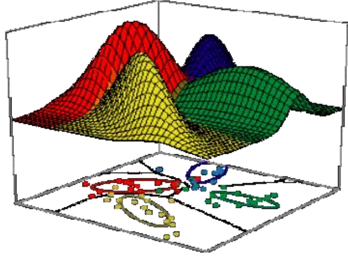
Textbook



- The course textbook is Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, Wiley, Second Edition, ISBN 0-471-05669-3, 2001.
- Additional textbooks covering experiment design:
  - Paul Cohen, Empirical Methods for Artificial Intelligence, MIT Press, ISBN 0-262-03225-2, 1995.
  - Sholom Weiss and Casmir Kulikowski, Computer Systems That Learn, Morgan Kaufmann, 1991.

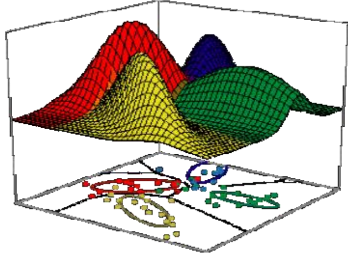
Note that one copy of each text should now be on reserve in the library.

Prerequisites



- Undergraduate introductory probability and statistics.
- Students are expected to have a working knowledge of calculus, linear algebra and basic probability theory.
  - A qualified student should understand Appendix A of the textbook without any problem.
    - Except: A.3, A.8
- In addition, proficiency in programming (e.g., MATLAB) is needed for the project and some of the assignments.

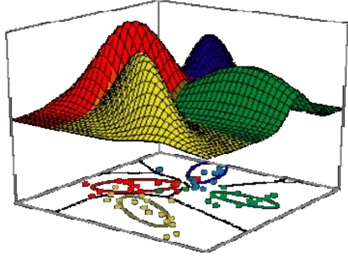
Course Materials



- Course website is managed through BrightSpace
  - Will be updated regularly
  - UofO students should automatically receive access:

  https://gradstudents.carleton.ca/faculty-of-graduate-and-postdoctoral-affairs-access-to-brightspace/
  - Slides will be posted before lectures, if possible
    - Any differences between posted draft slides and presented slides will be highlighted during class
- Course outline is on BrightSpace page.

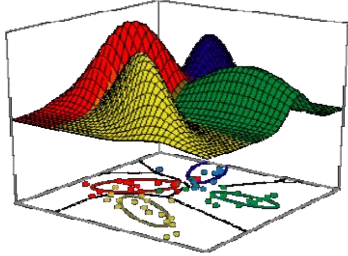
Grading



The final grade will consist of:

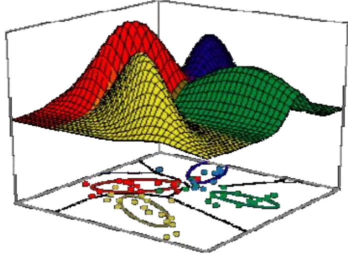
30%	**Assignments**	One or more assignments will be a mix of mathematical exercises and implementations/evaluations of pattern classification techniques. • Will require programming (choose between MATLAB, Python, R, etc.) • ***Portions may be used for evaluation***. • ***Peer evaluation may be used***
30%	**Term Project**	A course project will involve all students analyzing the same dataset using a variety of pattern classification approaches. Results will be presented in class with a mock-competition over a blind test dataset.
40%	**Final**	The final exam will cover all topics in the course. It will likely be a mixture of multiple choice and short answer questions.

Deadlines (subject to change)



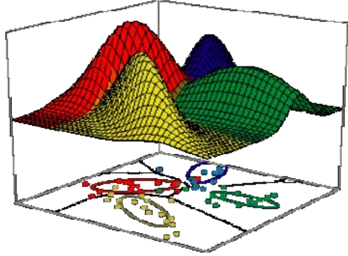
- Assignments
  - 1st assignment: due in < 2 weeks!
  - Assignments to follow ~bi-weekly
  - May include summarizing papers from the literature
  - May include oral presentations
  - Evaluation may be based on a subset of each assignment.
  - Peer evaluation may be used…

Deadlines (subject to change)



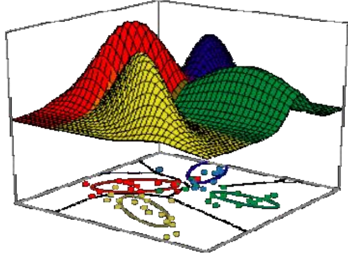
- Project:
  - The project commences on **Tuesday 22 Nov**.
  - Each group should choose a method and prepare a short 5-minute presentation describing their method, implementation, and planned approach for **Tuesday 29 November**.
  - In total, you'll have ~two weeks to develop and fine-tune your method.
  - On **Tuesday 6 December**, you'll make another 5-minute presentation where you must declare your expected performance over a blind test set.
  - The blind test data will then be released, and you'll have approximately **24 hours** to run your method on this blind test set.
  - Final competition results will be announced on **Thursday 8 December** (last class).
  - A final project report will be due shortly after the last class.
  - Detailed instructions will be provided in November.
  - One person from each group can choose a method in the Project section of the BrigthSpace page, starting on **22 Nov**. First come, first served! If you want to use a method that is not listed, I will happily add it.

Deadlines (subject to change)



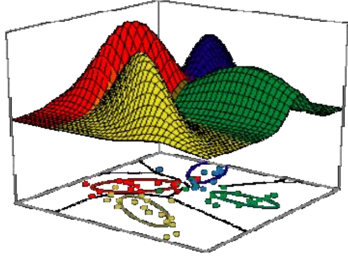
- Final Exam:
  - Formally scheduled.
  - In person.
  - Multiple choice and short answer

Plagiarism



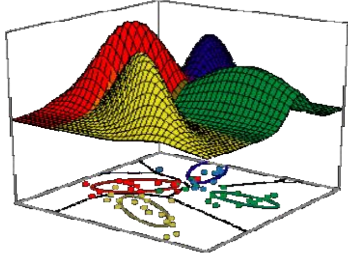
- Cite everything!
- Citing is not enough – cannot copy sentences from other's work, must use your own words.
  - Easy to spot, easy to Google…
- Suggested approach:
  - Take point-form notes while reading, keeping track of where you read it
  - Later, write your own paper based on your notes
  - This way, not tempted to use original phrasing.
- Suggest using a reference manager (e.g., Mendeley)
  - Keep track of papers you read and notes you take

Plagiarism



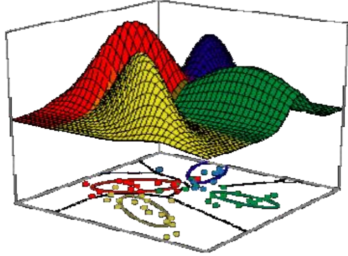
- Assignments:
  - Please include citations to any resources you use to solve the questions
    - E.g., open-source implementations of algorithms, tools used, etc.
  - It is OK to discuss a question and a possible <u>approach</u> with your colleagues, but the solution must be <u>your own work</u>.
    - Using a colleague’s code or words is not acceptable
- Literature review:
  - Make sure you cite all resources. Make sure you paraphrase or use quotes when summarizing the contents of a paper (see next slide)
- Final exam:
  - You <u>cannot discuss the exam or collaborate with colleagues.</u>

Course Rationale



Many graduate students are engaged in research involving classifying data using pattern recognition, machine learning, or artificial intelligence techniques. These students require a fundamental course early in their program that covers the topics of experiment design and conservative and correct reporting of research results as well providing a survey of the major approaches available. Such a course will be helpful to the earliest stages of their research in both critically evaluating the state of the art and also in forming their own research plan. Where possible, each method of pattern classification will be accompanied by a discussion of a recently published application, mainly from the field of biomedical informatics. The fundamental techniques introduced in this course will have wide appeal to many students, while the discussions of biomedical applications will have particular appeal to students in the growing biomedical engineering program.

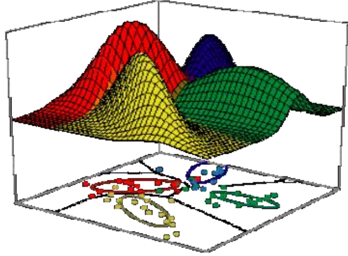
Course Outline (subject to change)



- **Weeks 1-2:** Introduction to Pattern Classification. Data pre-processing, analysis, outlier detection, and transformations. Experiment design (feature selection & dimensionality, selecting classifier structure, test protocols, cross-validation, data partitioning, etc). Hypothesis testing. Avoiding fundamental errors of testing on the training set and training on the test set.
- **Week 3:** Reporting results. How to accurately and honestly report classification system performance. True error vs. apparent error. Confidence intervals, statistical tests to compare methods, receiver operator characteristic curves, sensitivity, specificity, confusion matrices, P-values. Critical assessment of reported results in the literature.

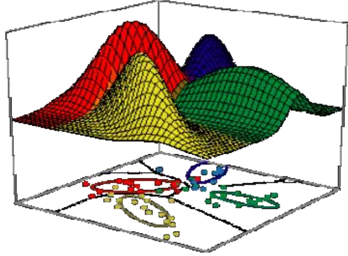
Note that by introducing experiment design and reporting of results early in the course, applications discussed during all subsequent topics will be evaluated using these fundamental principles. The remainder of the course will survey a number of approaches to pattern classification. The depth of coverage will vary and will depend on time available. Relevant applications of pattern classification techniques from the currently literature will be discussed.

Course Outline (subject to change)



- **Week 4:** Classification fundamentals. Brief review of Bayesian classification. Error rates. Classifiers, discriminant functions, and decision surfaces. The normal density. Missing features, Bayesian belief networks, and naïve Bayes rule.
- **Week 5:** Maximum likelihood and Bayesian parameter estimation. Non-parametric techniques such as Parzen windows, probabilistic neural networks, and K-nearest neighbour estimators and classifiers.
- **Week 6:** Decision trees and decision forests. Training, pruning, splitting and stopping criteria.
- **Week 7:** Linear and nonlinear discriminant analysis. Linear discriminant functions and decision surfaces. Perceptron criterion, relaxation procedures, and MSE procedures. Generalized linear discriminant functions and support vector machines (briefly).

Course Outline (subject to change)



- **Week 8:** Nonlinear system identification in the context of pattern classification. Neural networks: network structure, feedforward operation and classification, backpropagation training.
- **Week 9:** Markov chains, hidden Markov models, and expectation maximization.
- **Week 10:** Meta-learner and re-sampling approaches including bagging and boosting. Combination of multiple experts: voting strategies and cascaded classifiers. Learning with queries.
- **Week 11:** Unsupervised clustering (hierarchical, K-means, SOMs). Mixture densities, criterion function for clustering, and the number of cluster problem and cluster validation.
- **Week 12/13:** Student project presentations and *competition*. Review.