Basic Information

Class meets:
  Time: MW 2:30 – 3:20pm
  Place: BU-327

Lab section:
  Time: F 11:15 – 12 :05pm
  Place: I-109

Instructor:
  Predrag Radivojac
  Office: Eigenmann 1005
  Email:.predrag@indiana.edu.
  Web: http://www.informatics.indiana.edu/predrag

Office Hours:
  Time: MW 4-5pm or by appointment
  Place: Informatics Building, Eigenmann 1005

Course Web Site:
  Can be found from the School of Informatics course page
Some “Examples”

- “rules” for American presidents (before 2004 elections)
  - if the Washington Redskins win their last home game before the election, the incumbent’s party will be re-elected
  - no Republican has ever won a presidential election without carrying Ohio
  - no incumbent with a four-letter last name has ever been re-elected (Polk, Taft, Ford, Bush Sr.)
  - Americans won’t unseat a wartime President
What are the Tasks?

- Discovering patterns and rules
  - election “rules”, fraudulent telephone calls example
- Exploratory data analysis
  - interactive and visual techniques
- Descriptive modeling
  - cluster analysis, density estimation
- Predictive modeling
  - goal is to build a model that will enable one variable to be predicted from some other variables
- Information retrieval
  - web mining (Google), image mining etc.
Predictive Models
Classification Problems

- face recognition
- digit recognition
Clustering and Density Estimation

three clusters

1D and 2D prob. density
More Complex Clustering
What is Data Mining?

- practical discipline (but does not lack theory)
- aims to identify interesting new relationships and patterns hidden in numerous databases
- concerned with knowledge discovery
- data mining combines elements
  - statistics
  - machine learning
  - artificial intelligence
  - database management
  - other fields
Overview of the Course

Textbook information:

Required:
- Principles of Data Mining - by David Hand, Heikki Mannila, and Padhraic Smyth

Recommended:
- Data Mining - by Ian H. Witten and Eibe Frank
- Elements of Statistical Learning - by Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman

Supplementary material will be provided in class.
Overview of the Course

- introduction to probability
  - probability distributions
  - entropy
- data preparation techniques
- mining association rules
  - frequent itemsets
  - apriori algorithm
- classification and regression
  - linear discriminant analysis
  - decision trees
  - neural networks
  - support vector machines
Overview of the Course

- model selection and evaluation of experimental results
  - what algorithm to use
  - how to design evaluation process
- clustering
  - K-means algorithm
  - hierarchical clustering
- data analysis and visualization
  - dimensionality reduction
Grading Policy

- Midterm exam: 30%
- Final exam: 30%
- Homework assignments: 30%
- Class participation: 10%

- Midterm exam will be held in week 8 or 9
- Final exam will be as scheduled by the School of Informatics
Late Assignment Policy and Academic Honesty

- The homework assignments are due in class, on the specified due date.
- No late assignments will be accepted unless there are legitimate circumstances.
- All assignments are individual.
- All the sources used for problem solution must be acknowledged (people, web sites, books, etc.).