

## Final project pointers

CISC 5800  
Professor Daniel Leeds

## Homework 3 – part C

- Page 6 has been deleted
- Termination criterion for feature removal:  
"Testing error" equivalent to (100% - Percent\_Accuracy)  
Remove features until test error starts to **increase**

## Data set

- There is numeric and non-numeric data
  - Some non-numeric data can be mapped onto a number line
  - For non-mappable data, you can learn individual probabilities  
e.g.:  $P(\text{mood}=\text{"happy"} | \text{animal}=\text{"dog"}) = 0.6$   
 $P(\text{mood}=\text{"sad"} | \text{animal}=\text{"dog"}) = 0.2$
- There are missing features for certain data points
  - Could ignore those features for that data
  - Could guess "typical" features for that data

## Learning methods

- Bayes
- Naïve Bayes
- Support Vector Machines
- Logistic regression

## Software

- Statistics toolbox for Matlab: has svmclassify, pca
- Free software: SVM light, svd (Matlab version of pca), fastica

## Experimenting with "learning parameters"

By "learning parameters" I mean:

- Regularization
- Update steps
- Probability distributions to learn for Bayes/Naïve Bayes
- Number of training data points
- Number of repeated iterations on training data
- Slack variable constant size
- Kernel type