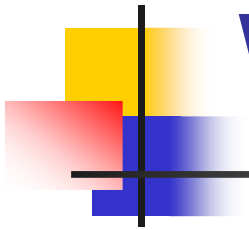
Cmput 466 / 551

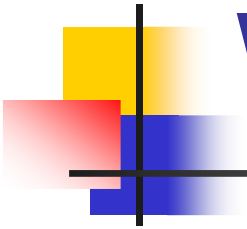
# Summary

R Greiner
University of Alberta



# What we covered (I)

- Machine Learning intro [ ½ ]
  - What it is… why relevant…
  - Supervised learning
    - … steps …
- Foundations [ 1 ]
  - Simple estimation (MLE)
- Linear Regression [1]
  - Minimize L2 loss – direct matrix computation
  - Gradient descent
  - Connection to Gaussian Error model
  - Different basis functions



# What we covered (II)

- Evaluation (supervised learning) [1]
  - Training set error ≠ "real" error
  - Hold out set… cross-validation …
  - In-fold feature selection / parameter setting
- Overfitting [1]
  - Bias-variance tradeoff
  - Regularization … L2
    - … just Bayesian MAP …
  - L1-regularization (NOT: feature selection,…)

# What we covered (III)

- Linear Classifiers [2 ½ ]
  - Perceptron – noiseless, margin, dual, …
  - LMS – gradient descent (Newton-Raphson)
  - Logistic Regression – P(y|x)
  - LDA, FDA, … – P(y,x)
- Support Vector Machines [2 ½ ]
  - Primal/Dual, Lagrange Multipliers, KKT
  - Max Margin
  - Quadratic programming
  - Kernels!
    - Different expressiveness, ok-complexity (using Dual)
  - Classification … or Regression

Fun Programming Exercise!

# What we covered (IV)

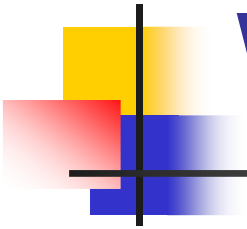
- Neural Nets [1 ½ ]
  Gen'l Multi-layered Feed-forward (logistic)
  - Motivation, Definition
  - Expressibility
  - Training: Gradient descent, Backpropagation
- Useful tricks [1]
  - Conjugate Gradient
  - Line Search
- Other aspects [ ½ ]
  - Auto-Encoders
  - Example
  - Not: Deep learning

# What we covered (V)

- Decision Trees [2]
  - Motivation, Definition
  - Expressibility
  - Training:
    - Greedy
    - Info Gain, …
  - Avoiding overfitting
    - Minimum Description Length
    - Chi$^2$
    - PostPruning
      - Validation set
    - Re-represent (rule sets)
- Issues:
  - real values
  - other splitting,
  - attribute costs
  - missing values
    - Why missing?

# What we covered (VI)

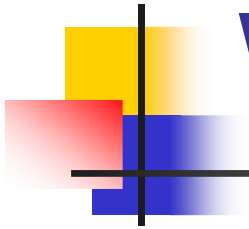
- Formal Model of Learnability (PAC) [1]
  - Motivation, Framework (Protocol)
    - Sample complexity
    - Computation complexity
  - Probably Approximately Correct
    - Union bounds;  Chernoff/Hoeffding inequality
    - Just O(.) … but insightful !
  - Consistency filtering algorithm
    - Learning Conjunctions, k-CNFs, …
  - Covering algorithm
    - Learning Decision-Lists
  - How sample-size depends on expressibility
    - Trade-off
- Not: Errors in data (unrealizable), sequential algorithms, …
- Not: ∞ space … VC Dimension



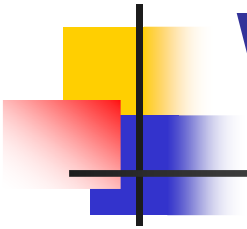
# What we covered (VII)

- Imbalanced data (in general)
- Cost Curves [1]
  - Guest lecture: R Holte

# What we covered (VIII)

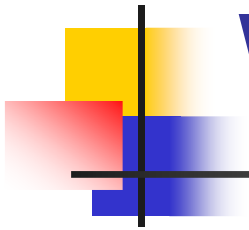


- Probability Foundation [1]
  - Foundations
  - Dutch-book arguments
  - MLE … and issues ⇒ Bayesian Framework
  - Beta (for Binomial) … conjugate prior … "smoothing"
  - Estimation for Gaussians



# What we covered (IX)

- Bayesian Belief Nets: Intro [2]
  - Motivation, Framework
    - Conditional + Just relevant ⇒ Belief Net
    - Conditional independence
  - Semantics of directed graph structure
    - d-separation
  - Useful generalization of CP-tables
  - Inference
    - Naïve Bayes, …
    - Complexity… NP-hard but with special cases
    - Max Expected Utility



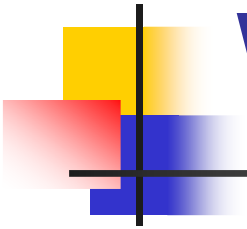
# What we covered (X)

- Belief Nets: Learn Parameters [1]
  - Here: "learning" = density estimation
    - Generative model
  - Complete data
    - Frequentist: Trivial (MLE = observed frequency)
    - Bayesian: Trivial (MAP ≈ observed frequency)
      - BDe prior
  - Incomplete data
    - Gradient descent
    - Expectation Maximization
    - Gibbs

# What we covered (XI)

- Hidden Markov Model [1]
  - Inference… Filtering, Prediction, Likelihood, …
  - Extensions
  - Learn parameters

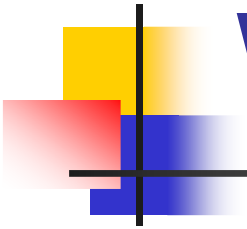
- Belief Nets: Learn Structure [0]
  - General situation: Score-based
    - Find structure that is best, for its best Cptable
    - Find parents that are VERY informative about child
  - Tree-structured
    - Simple Chow-Liu Algorithm … efficient
    - Tree-augmented Naïve Bayes
  - Bayesian framework, to "avoid" overfitting
    - Margin probability of structure (over all parameters)



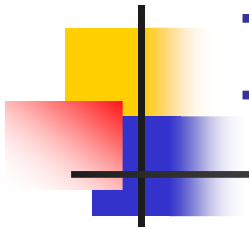
# What we covered (XII)

- Principal Component Analysis (PCA) [1/2 ]
  - Goal: reduce dimensions…
    while preserving variance
  - Relation to Eigen-vector/value
  - Eg: Eigenfaces
  - Where PCA is useful… or not

# What we DID NOT cover

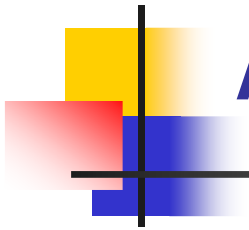


- Imbalanced data (in general)
- Ensemble methods
- Deep learning
- PCA
- Clustering
- Reinforcement learning



# Issues

- Text book ?
  - Is there a better one?
  - Is it needed?
  - Note that HTF is ON-LINE (free)
- Coursera assignments (?)
  - Too easy?  Useful? …
- No lab … did not seem needed?
- Project ?
  - Useful?
  - Is there a better way to do it?
- "Losing" 2 lectures for presentations
  - Use Lab time for additional lectures?



# Any feedback?

- If you have ideas / feedback / comments
  - What material to remove? … add?
  - Assignments?  Coursera or otherwise?
  - …
- Contact me:
  - rgreiner@ualberta.ca