

What You Should Know (by now)

Introduction to Statistical Machine Learning
ANU COMP 6467/4670, Sem 1, 2008

Scott Sanner
NICTA / RSISE
First.Last@nicta.com.au

Note

- This lecture covers remaining material not covered in mid-semester review lecture
- View those slides before this set

Your Linear Classifier Toolbox

- Linear or Log-linear Discriminators
 - Naïve Bayes
 - Logistic Regression
 - Linear Discriminant Analysis
 - Fit Gaussians with same covariance
 - Also Fisher linear discriminant
 - Support Vector Machines (SVMs)
 - And most training methods with convex loss / regularization
- Remember
 - Very powerful, can use non-linear features!
 - Optimal solution guaranteed w.r.t. assumptions

Your Non-linear Classifier Toolbox

- Caveat for Non-linear methods
 - Non-linear methods usually can't guarantee optimum
 - But good optimization methods can mitigate this
- Non-linear Discriminant Analysis
 - Fit any probabilistic model by *density estimation*
 - Exponential family (most functions)
 - Just estimate sufficient statistics
 - e.g., Gaussian requires mean, covariance
 - If model correct:
 - gives Bayes optimal decision
 - If model incorrect:
 - may still work reasonably, may not

Your Non-linear Classifier Toolbox

- Neural Networks
 - Non-linear for 1+ hidden layer
 - Capable of learning all boolean functions
 - Efficient backpropagation training
 - Methods apply generally to any continuous functions
- Decision Trees
 - Use entropy (or Gini) for boolean splits
 - Use (non-linear) discriminator for real-valued splits
 - One of most used methods
 - Easy, interpretable, expressive
 - Good feature selection methods

Your Regression Toolbox

- Linear Regression
 - Standard linear least squares
 - Make sure can derive (see tutorial notes)
 - Lasso and related methods
 - L_1 (rather than L_2) normalization promotes sparsity
 - Support vector regression
- Remember
 - Very powerful, can use non-linear features!
 - Optimal solution guaranteed given assumptions

Your Nonlinear Regression Toolbox

- Caveat for Non-linear methods
 - Non-linear methods usually can't guarantee optimum
 - But good optimization methods can mitigate this
- Non-linear Regression
 - Neural nets
 - If need output other than $[-1,1]$, don't use logistic function
 - Again: applies generally to continuous functions
 - Regression trees
 - Decision trees with constants, functions at leaves

Model (and Feature) Selection

- Can use different models / features
 - Features:
 - Linear, polynomial, general function (log,exp)
 - Hypothesis space:
 - Linear, polynomial, general function for discriminant / regression
- How to select “correct features / model”?
- Model-selection criteria
 - Theoretical:
 - Minimum Description Length, BIC, AIC
 - Empirical:
 - Cross-validation (K-fold, leave-one-out, etc...)
 - But caveats to cross-validation with low data

Graphical Models

- In one sentence
 - Structured probabilistic models, inference
- Motivating example with n binary variables
- Need to use GMs when
 - Cannot represent 2^n probability table entries
 - Cannot learn low variance estimate of 2^n parameters given data
 - In this case, better to choose bias over variance in tradeoff
 - Cannot efficiently do inference in $O(2^n)$ computations

Unsupervised & Semisupervised

- Unsupervised
 - Unlabelled data
 - Want to learn latent variables predictive of data structure
 - Methods
 - PCA, ICA
 - EM (special case of K-means)
- Semisupervised
 - Use labelled + unlabelled data to find latent variables
 - Perform classification / regression in latent space!
 - Great when lots of data, but little is labelled
 - e.g. web information extraction and EM / co-training

Temporal (Graphical) Models

- When data non i.i.d., but temporally related
 - Time series, sequential data
 - Examples
 - Text, speech, video
 - Tracking objects
- Most common models
 - Markov Model (fully observed)
 - Hidden Markov Model (partially observed)
- Many inference approaches
 - What was most likely or expected sequence
 - What will be next datum?
 - Note, this special inference case is *sequential prediction*

Offline vs. Online

- Offline
 - Given data all at once
 - Build classifier / regressor
 - Case most studied in this course
- Online (sequential prediction)
 - Data comes one by one
 - Arguably fits many applications better... continuous learning!
 - Make best prediction given data so far
 - Can look at
 - Incremental updates to classifiers / regressors
 - Bayesian framework where constantly update distribution