January 28, 2002, COMS 6998-01 Advanced Machine Learning

Prof. Tony Jebara, CEPSR 605 jebara@cs.columbia.edu
http://www.cs.columbia.edu/~jebara/6998-01
Office Hours: Tuesday 2-4, Thursday 11am-12am, or appt.

Text book Online, Duda Hart & Stork, Bishop

7. Email me for permission

Handout Outline (webpage), Grading, Homework & Project

Pre-reqs: Linear Algebra, Calculus
Intro. Machine Learning or Stats

Class List: names, emails, listener or for credit

Course Structure: Seminar-like, hands-on,

interactive, MATLAB, get your hands dirty,
research, topics, new and controversial ideas
as well as formal. Email-list: discussion group

Learn a set of tools for many applications

& research.

ML: fields are expanding, hard to keep up, Jack of all trades

Physics, Math, Statistics, Economics OR Neuroscience, Psych, Biology
ML: model complex and non-deterministic systems

Physics \( E=mc^2 \) many other domains: unknown cnps on vars
- partial observations
- unknown | uncertain | incomplete models
- high dimensions 100,000 vars, not 3
- noise
- complex (non-linear)
- stochastic
- refine models with data observations

Applications:
- Speech Rec \( \rightarrow \) HMMs
- Computer Vision (face rec, digits, supervised)
  all share
- Time Series Prediction (weather, financial...)
- Common tools and theory
  - Genomics (DNA, micro-arrays, SVMs, splice sites)
  - NLP & Parsing HMMs, CRFs \( \rightarrow \) Whizbang
  - Text clas, IR (documents, spam, TSUMs)
  - Medical: AHA-AT \( \rightarrow \) 200 diseases

History:
1942: Norbert Wiener, Cybernetics, Multi-fields
1943: Shannon
1956/5: Minsky, McCulloch, Herbert Simon, John McCarthy, \( \rightarrow \) \( \rightarrow \) HMM Bowers
  Symbolic AI, Log., Inconsistency, Rule-based
1957: Rosenblatt Perceptron
  \( \rightarrow \) AI, 2005:
1960: Minsky-Papert: Perceptron Linear \( \rightarrow \) Graphical Models 1990
  Lauritzen, Pearl
1994: Werbos, PhD Backprop, Neuro
1986: Rumelhart & Mclachlan, HIF, conjugate
1980's: NN, ANNs, Genetic Alg, Fuzzy Logic, Black Boxes
1990's: Bayesian & Statistical & Structure & Priors
Graphical Models: MRF, HMM, Sig, Belief Nets, MRFs
SVMs, Computation
Course Outline: Slow & Faster, email me on pace
old topics + new research
EM - 3 views

General ML Recipe:
Problem
Data, Features (Quantitative/Qualitative)
\downarrow \text{Vectors, Distributions, Prior}
\downarrow \text{Knowledge, Parametrization, Structure, Model}
\downarrow \text{Criterion (task)}  \quad \text{(min squared error)}
\downarrow \text{Algorithm (lead if intractable, instead)}

Tasks:
- Modeling & Density Estimation (Compression, Coding, Analysis)
- Clustering
- Classification (binary, multiclass X, K, K, y, yi, yi) 859
- Regression
- Structure Learning
- Feature Selection, Subspace
- Transduction
- Anomaly Detection

Discriminative vs. Generative: p(x|y) vs.

\text{General yet suboptimal} \quad \text{Gt whatever by manipulating this,}

\text{Analytic pdf over all variables in system}
Probability Theory Review:

Joint
\[ \sum_{y} p(x,y) > 0 \]

Marginalization:
\[ \sum_{y} p(x,y) = p(x) \]

Condition:
\[ p(x|y) = \frac{p(x,y)}{p(y)} \]

Bayes Rule:
\[ p(x|y) = \frac{p(y|x)p(x)}{p(y)} \]

\( X \perp Y \): independent \( \implies p(x,y) = p(x)p(y) \) or \( p(x|y) = p(x) \)

\( X \perp Z \mid Y \): cond. indep. \( \implies p(x,z|y) = p(x|y)p(z|y) \)

But
\[ p(x,z|y) = p(x|y) + p(z|y) \]

Classifier:
\[ p(y|x) \]
\[ \hat{y} = \arg\max_y p(y|x) \]
Regression:
\[ p(y|x) \]
\[ \hat{y} = \sum_y y p(y|x) \]

Anomaly detection:
\[ p(y) > \text{threshold} \implies \text{true, else anomaly} \]

Graphical Model Rep:

Nodes \( \rightarrow \) Variables

Independence = no link

Dependence = link

Cyclical

Arrow = parent-child (\( \sim \) causal)

\( p(x_1, \ldots, x_n) = \prod_i p(x_i | p(x_i)) = \prod_i p(x_i | \pi(x_i)) \)

Efficiency tables, later on...

Chain of events

\[ x = \text{trip} \]
\[ y = \text{fall down stairs} \]
\[ z = \text{bruise} \]

\( x \perp z \mid y \)

\[ p(x|z,y) = \frac{p(x,y,z)}{p(z,y)} = p(x|y) \]
One cause with 2 effects

\[ \begin{align*}
  y &= \text{flu} \\
  x &= \text{sore throat} \\
  z &= \text{temperature}
\end{align*} \]

\[ x \perp \!\!\!\!\!\!\perp y \]

2 causes, 1 effect
Explaining away

\[ \begin{align*}
  x &= \text{rain} \\
  y &= \text{wet driveway} \\
  z &= \text{car oil leak}
\end{align*} \]

- How to get \( p(x,y,\ldots) \)? Parametric model \( p(y|x,\ldots|\theta) \)

Estimation: Two Schools: Bayesian & Frequentist
Both have own advantages, we focus on Bayesian

Frequentist: Classical, objective, no priors

Some things are not distributions

Can talk about \( p(\text{event}) \) if never get data.

\[ p(\theta) \] only true model, not R.V.

\[ p_0(x,y) = p(x|y,\theta) \]

Can't say prob. a coin will be 50% heads we observe.

Plug in \( \hat{\theta} \) is generated with estimator, MLE

Bayesian: Subjective, pdf on anything (i.e. uncertainty)

Even on deterministic values, like speed of light

Bayes rule:

\[ p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \]

You pick it

Simplicity

Occam

Bayes 1702-1761