Scanner : 470



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

Prof. Tony Jebara, CEPSR 605 jebara@cs.columbia.edu

CS Dept. Area: ML for modeling people, behavior, medical, visual

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
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, researchy 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 separating, 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 eqns. on vars

- partial observations
- unknown / uncertain / incomplete models
- high dimensions 100 000 vars, not 3
- noise
- complex (non linear)
- stochastic
- refine models with data / real measurements and observations

Applications:

all share common tools and theory

- Speech Rec HMMs
- Computer Vision (face rec, digits, MRFs Super-res.,
- Time Series Prediction (weather, financial ...)
- Genomics (DNA, micro-arrays, SVMs, splice sites)
- NLP & Parsing HMMs, CRFs, Whizbang
- Text classif. & IR (documents, spam, TSVMs)
- Medical: QMR-DT 600 diseases 4000 symptoms

History:

~~1917: Karel Capek - Robot~~

1943: McCulloch & Pitts (Bio, Neuron)

1947: Norbert Wiener, Cybernetics, multi-fields

1949: Shannon

1950's: Minsky, Allen Newell, Herbert Simon, John McCarthy
Symbolic AI, Logic, INCONSISTENCY, Rule-Based

1957: Rosenblatt Perceptron

1969: Minsky & Papert: Perceptron linear ~~XOR~~ kill AI, 12-15yrs

1974: Werbos, PhD BackProp, Nonlinear

1986: Rumelhart & McLennen, MLP, Conjugate

Statistics

1970 HMM Baum

1977 EM (Demp.)

Graphical Models 1980
Lauritzen, Pearl

1980's: NN, RNN, Genetic Alg., Fuzzy Logic, Black Boxes.

1990's: Bayesian & Statistical & Structure & Priors

Graphical Models: EM, KF, HMM, Sig. Belief Nets, MRFs

Course → SVMs, Comput. Learning Theory, Boosting

Course Outline: slow & faster, email me on pace

old topics + new research: EM - 3 views

General ML Recipe: Problem

↓

Features (Quantitative) vectors, discrete...

↓

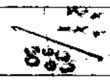
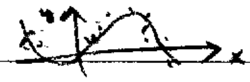
Data → model ← Prior knowledge, parametrization, structure

↓

criterion / task (min squared error)

↓

algorithm (bad if intractable, inexact)

Tasks:
- modeling & density estimation (compression, coding, analysis)
- clustering
- classification < binary, multiclass  $x_1, \ldots x_T$ $y_1, \ldots, y_T$ 
- regression 
- structure learning 
- feature selection, subspace
- transduction
- anomaly detection

overlap

Discriminative vs. Generative: $p(x, y, \ldots)$ ← analytic pdf over all variables in system

general yet suboptimal ← $p(x,y,\ldots)$ → get whatever by manipulating this.

assume we have $p(x,y..)$

joint

random vars

**Probability Theory Review:** $p(x,y) \geq 0$

(useful properties) $\sum_{x,y} p(x,y) = 1 \longrightarrow \int_{x,y} p(x,y) = 1$

marginalize: $\sum_y p(x,y) = p(x)$

condition: $p(x|y) = \frac{p(x,y)}{p(y)}$

bayes rule: $p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$

$x \perp\!\!\!\perp y$ : independent $\Longrightarrow p(x,y) = p(x)p(y)$ OR $p(x|y) = p(x)$

$x \perp\!\!\!\perp z \mid y$ : cond indep $\Longrightarrow p(x|z,y) = p(x|y)$

BUT $\longrightarrow p(x|z) \neq p(x)$

Classifier: $p(y|x)$ $\quad \hat{y} = \underset{}{\text{arg max}} \{ p(y=0|\vec{x}), p(y=1|\vec{x}) \dots \}$

regression: $p(y|x)$ $\quad \hat{y} = \underset{q}{\text{argmax}}\; p(y=q|\vec{x})$

$\hat{y} = \int y\, p(y|\vec{x})\, dy$

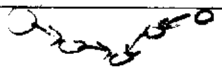
anomaly detection: $p(x) \geq$ threshold $\Longrightarrow$ true, else anomaly

**Graphical Model Rep:** nodes → variables

indep. ⇒ no link $\quad x \quad y$

depend ≡ link $\quad x \longrightarrow y$

acyclical



arrow ≡ parent-child (~causal)

$p(x_1, \dots, x_n) = \prod_i p(x_i \mid pa_i) = \prod_i p(x_i \mid \pi_i)$

↑ efficiency, tables, later on..

$k^n$ vars

chain of events

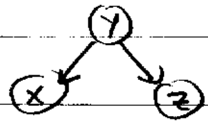
$x \longrightarrow y \longrightarrow z$ $\quad p(x)\,p(y|x)\,p(z|y)$

x = trip
y = fall down stairs
z = bruise

$x \longrightarrow y \longrightarrow z$ $\quad x \perp\!\!\!\perp z \mid y \quad p(x|z,y) = \frac{p(x,y,z)}{p(z,y)}$

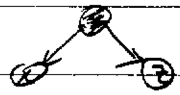
$= p(x|y)$

↑ shading

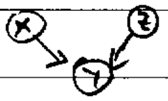



One cause* with 2 effects

[ Y = flu
x = sore throat
z = temperature

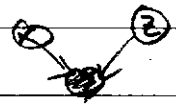



$x \perp\!\!\!\perp z \mid y$



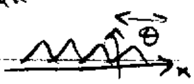

$x \perp\!\!\!\perp z$

2 causes*, 1 effect
explaining away




$x \not\perp\!\!\!\perp z \mid y$

[ x = rain
y = wet driveway
z = car oil leak




- How to get $p(x, y, \ldots)$ ? Parametric model $p_\theta(x, y, \ldots \mid \theta)$ ← model ~~choose~~ from

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(event) if never get data.

$p(\cancel{\theta})$, only 1 true model, not R.V.

$p_\theta(x,y)$ $\quad$ $p(x,y\cancel{|\theta})$

can't say prob. a coin will be 50% heads w/o observ.

plug in Single $\hat{\theta}$ is generated with estimator, MLE ⟷ ~~min variance, unbiased estimator~~

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

even on deterministic values, like speed of light
pure

Bayes rule: $p(\theta \mid X) = \dfrac{p(X \mid \theta)\, p(\theta)}{p(X)}$

You pick it
Simplicity
Ockham
1280-1349
mathematical convenience

distrib over $\hat{\theta}$

$\text{Posterior} = \dfrac{\text{likelihood prior}}{\text{evidence}}$

Bayes 1702-1761