10-601 Review

Tom Mitchell and Aarti Singh

Machine Learning 10-601
Dec 8, 2011

MACHINE LEARNING DEPARTMENT

Carnegie Mellon
School of Computer Science

Machine Learning Algorithm

<u>Goal:</u> Learn a rule Z ➔ f(Z) that optimizes some objective – loss(f(Z)).

Z can be X or (X,Y)

modeled as a random variable, and we optimize E_Z[loss(f(Z))]

Training Data

D = $\{Z_i\}_{i=1}^n$ ⇨ **Learning algorithm** ⇨ Rule $\widehat{f}_n$

Why do we need training data?

Modeling Distributions

<u>Parametric</u>: $P_\theta(Z)$ θ

Gaussian – continuous random variables μ, σ

Bernoulli – binary/boolean random variable θ

Binomial – sum of binary/boolean random variables n, θ

Multinomial – sum of k-ary random variables $n, \theta_1, \theta_2, \ldots, \theta_k$

Beta, Dirichlet (conjugate prior for binomial, multinomial), Poisson, …

If θ is a random variable, $P_\theta(Z) = P(Z|\theta)$ likelihood

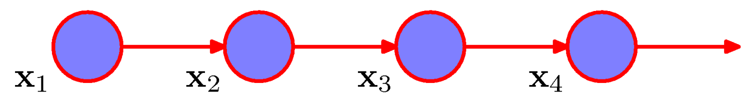
<u>Bayes Rule:</u> $P(\theta|Z) = \dfrac{P(Z|\theta)\, P(\theta)}{P(Z)}$ posterior

Modeling Distributions

Conditional independence assumptions for joint distributions:

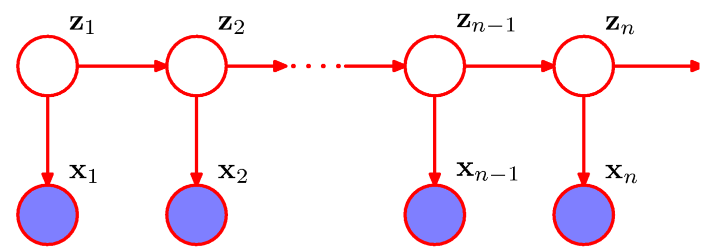
- Markov Models

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i | X_{i-1})$$



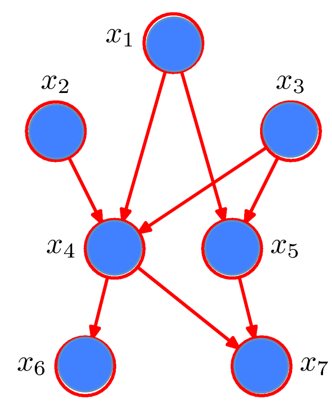
- Hidden Markov Models

$$p(\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} p(X_i | Z_i) \prod_{i=1}^{n} p(Z_i | Z_{i-1})$$



- Bayes Nets/Graphical models

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i | pa(X_i))$$



Machine Learning Problems

Broad categories -

- **Unsupervised learning**

  Density estimation, Clustering, Dimensionality reduction

- **Supervised learning**

  Classification, Regression

- **Semi-supervised learning**
- **Active learning**
- Many more …

Unsupervised & Supervised Learning

Unsupervised Learning – Learning without a teacher

$$\{X_i\}_{i=1}^n \Rightarrow \boxed{\text{Learning algorithm}} \Rightarrow \widehat{f}_n$$

Documents

Model for word distribution OR Clustering of similar documents

Supervised Learning – Learning with a teacher

$$\{(X_i, Y_i)\}_{i=1}^n \Rightarrow \boxed{\text{Learning algorithm}} \Rightarrow \widehat{f}_n$$

Documents, topics

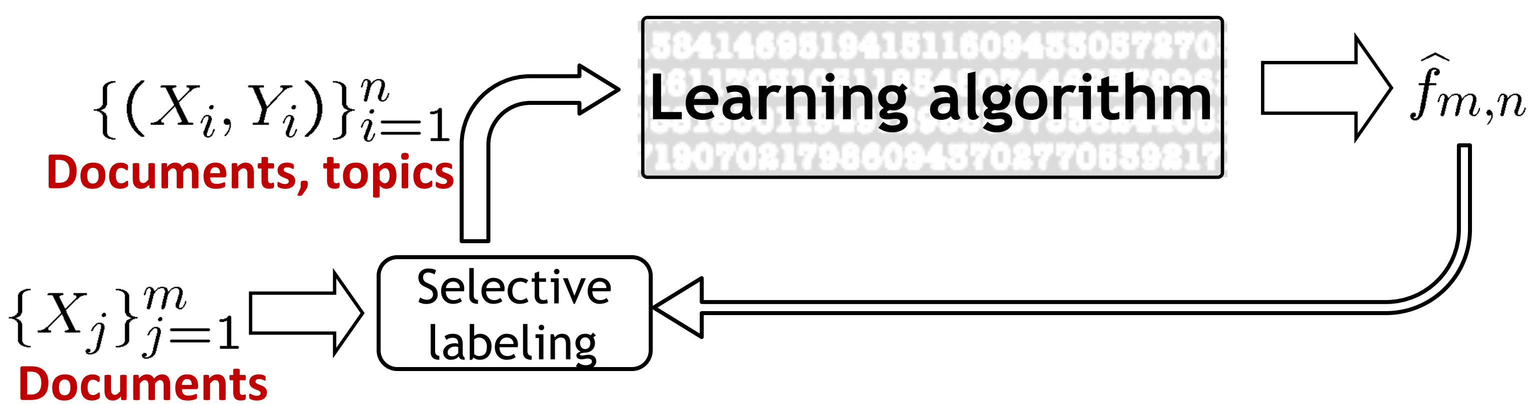
Mapping between Documents and topics

Semi-supervised & Active Learning

Semi-Supervised Learning – ***randomly*** labeled examples

Mapping between Documents and topics

Active Learning – ***selectively*** labeled examples



Unsupervised Learning

Density estimation:

Parametric (MLE, MAP)
Nonparametric (Histogram, Kernel)

Dimensionality reduction:

Feature Selection
Principal Component Analysis (PCA)
Laplacian Eigenmaps

Clustering:

Gaussian mixture models
k-means
spectral

Supervised Learning

<u>Regression:</u> (Continuous labels, Mean Square Error)

Optimal estimation rule

$f^*(X) = E[Y|X]$ MLE under $P(Y|X) = \mathcal{N}(f^*(X), \sigma^2)$

Linear Regression $f(X) = \mathbf{X}\, w$, $\mathbf{X} = [x_1, x_2, \ldots, x_d]$

Polynomial Regression $\mathbf{X} = [x_1^2, x_1x_2, x_2^2, \ldots]$

Basis Regression $\mathbf{X} = [\phi_1(x), \phi_2(x), \ldots, \phi_d(x)]$

Regularized versions (MAP)

Neural Networks $f(X) =$ nonlinear (combination of multiple logistic units)

Kernel (locally-weighted) - Weighted mean square error

Supervised Learning

Classification: (Discrete labels, Probability of error)

Bayes optimal classification rule

$$f^*(X) = \arg\max_{Y} P(Y|X)$$

plug-in MLE, MAP of distribution model

Naïve Bayes

Decision Trees

Logistic Regression

k-nearest neighbor

SVM

Boosting

Comparison Chart for Classification

	Decision Trees	K-NN	Gauss Naïve Bayes	Logistic regression	Neural Nwks	HMM	Bayes Net	SVM	Boosting
Gen/Disc									
Loss functions									
Decision boundary									
Output									
Algorithm									
Model Complexity									
Relation to others									
Parametric/ Nonparam									

Some Topics We've Covered (before Midterm)

Decision trees

entropy, mutual info., overfitting

Probability basics

Bayes rule, MLE, MAP,
conditional indep.

Naïve Bayes

conditional independence,
# of parameters to estimate,
decision surface

Logistic regression

form of P(Y|X)
generative vs. discriminative

Linear Regression

minimizing sum sq. error (why?)
regularization ~ MAP

Sources of Error

unavoidable error, bias, variance

Overfitting, and Avoiding it

priors over H

cross validation

PAC theory: probabilistic bound on overfitting

Bayesian Networks

factored *representation* of joint
distribution, conditional independence
assumptions, D-separation
inference in Bayes nets
learning from fully/partly observed data

PAC Learning

sample complexity
probabilistic bounds on $error_{train} - error_{true}$
VC dimension

Some Topics We've Covered (after Midterm)

Hidden Markov Models
- time-series/sequential modeling
- representation, parameters
- *evaluate* prob of output sequence
- *decode* hidden states
- *learning* parameters

Neural Networks
- nonlinear classifier
- layers of multiple logistic units
- training – backpropagation
- local minimum

Dimensionality reduction
- feature selection
- PCA –linear, directions of max variance, SVD
- Laplacian Eigenmaps – nonlinear

Clustering
- k-means – isotropic, convex
- spectral - connectivity based

Nonparametric methods
- histogram, kernel density est
- kernel regression
- k-NN classifier

Support Vector Machines
- hard-margin, soft-margin
- support vectors
- dual formulation, kernel trick

Boosting
- weak base classifiers trained on re-weighted data
- Adaboost algorithm, exp loss

Four Fundamentals for ML

1. Learning is an optimization problem
   - many algorithms are best understood as optimization algs
   - what objective do they optimize, and how? Local minima?
   - gradient descent/ascent as general fallback approach

Four Fundamentals for ML

1. Learning is an optimization problem
   - many algorithms are best understood as optimization algs
   - what objective do they optimize, and how?
2. Learning is a parameter estimation problem
   - the more training data, the more accurate the statistical estimates
   - MLE, MAP, M(Conditional)LE, …
   - to measure accuracy of learned model, we must use test (not train) data

Four Fundamentals for ML

1. Learning is an optimization problem
   - many algorithms are best understood as optimization algs
   - what objective do they optimize, and how?
2. Learning is a parameter estimation problem
   - the more training data, the more accurate the estimates
   - MLE, MAP, M(Conditional)LE, …
   - to measure accuracy of learned model, we must use test (not train) data
3. Error arises from three sources
   - unavoidable error, bias, variance
   - PAC learning theory: probabilistic bound on overfitting: $error_{true}$ - $error_{train}$

Bias and Variance of Estimators

given some estimator Y for some parameter θ, we note Y is a random variable (why?)

the <u>bias</u> of estimator Y : $E[Y] - \theta$

the <u>variance</u> of estimator Y $E[(Y - E[Y])^2]$

consider when

- θ is the probability of "heads" for my coin
- Y = proportion of heads observed from 3 flips

consider when

- θ is the vector of correct parameters for learner
- Y = parameters output by learning algorithm

Four Fundamentals for ML

1. Learning is an optimization problem
   - many algorithms are best understood as optimization algs
   - what objective do they optimize, and how?
2. Learning is a parameter estimation problem
   - the more training data, the more accurate the estimates
   - MLE, MAP, M(Conditional)LE, …
   - to measure accuracy of learned model, we must use test (not train) data
3. Error arises from three sources
   - unavoidable error, bias, variance
   - PAC learning theory: probabilistic bound on overfitting: error_{true} - error_{train}
4. Practical learning requires making assumptions
   - Why?
   - form of the f:X → Y, or P(Y|X) to be learned
   - priors on parameters: MAP, regularization
   - Conditional independence: Naive Bayes, Bayes nets, HMM's

Other interesting ML topics

- Reinforcement learning
- Transfer learning
- Multi-task learning
- Online learning, …

Useful tools:

- Matrix factorization
- Matrix completion
- Random projections
- Compressed sensing, …

Related courses

<u>Regular</u>

- Machine Learning Theory (15-859 B) - Avrim Blum
- Statistical Machine Learning (10-702) – Larry Wasserman
- Adaptive Control and Reinforcement Learning (16-899 C) - Drew Bagnell
- Probabilistic Graphical Models (10-708) – various instructors

<u>New Spring 2012</u>

- Information Processing and Learning (10-704) – Aarti Singh
- Machine Learning with Large Datasets (10-605) - William Cohen

ML PhD Thesis topics 2010

- Coupled Semi-Supervised Learning – Andrew Carlson
- Rare Category Analysis - Jingrui He
- Tractable Algorithms for Proximity Search on Large Graphs - Purnamrita Sarkar
- Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy - Brian D. Ziebart
- Structural Analysis of Large Networks: Observations and Applications - Mary McGlohon
- Nonparametric Learning in High Dimensions - Han Liu