10-601 Review

Tom Mitchell and Aarti Singh

Machine Learning 10-601
Dec 8, 2011
Machine Learning Algorithm

Goal: Learn a rule $Z \rightarrow f(Z)$ that optimizes some objective – $\text{loss}(f(Z))$.

$Z$ can be $X$ or $(X,Y)$ modeled as a random variable, and we optimize $E_Z[\text{loss}(f(Z))]$

Training Data $D = \{Z_i\}_{i=1}^n$ $\xrightarrow{\text{Learning algorithm}}$ Rule $\hat{f}_n$

Why do we need training data?
Modeling Distributions

**Parametric:** $P_{\theta}(Z)$

- **Gaussian** – continuous random variables
  $\mu, \sigma$
- **Bernoulli** – binary/boolean random variable
  $\theta$
- **Binomial** – sum of binary/boolean random variables
  $n, \theta$
- **Multinomial** – sum of k-ary random variables
  $n, \theta_1, \theta_2, \ldots, \theta_k$
- **Beta, Dirichlet** (conjugate prior for binomial, multinomial), **Poisson**, ...

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

**Bayes Rule:** $P(\theta|Z) = \frac{P(Z|\theta) \ P(\theta)}{P(Z)}$  posterior
Modeling Distributions

Conditional independence assumptions for joint distributions:

- **Markov Models**

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

- **Hidden Markov Models**

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

- **Bayes Nets/Graphical models**

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

\[ \{X_i\}_{i=1}^n \rightarrow \text{Learning algorithm} \rightarrow \hat{f}_n \]

Documents

Model for word distribution OR Clustering of similar documents

Supervised Learning – Learning with a teacher

\[ \{(X_i, Y_i)\}_{i=1}^n \rightarrow \text{Learning algorithm} \rightarrow \hat{f}_n \]

Documents, topics

Mapping between Documents and topics
Semi-supervised & Active Learning

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

\[
\{(X_i, Y_i)\}_{i=1}^n, \quad \{X_j\}_{j=1}^m
\]

Documents, topics

\[
\{(X_i, Y_i)\}_{i=1}^n
\]

Documents, topics

\[
\{X_j\}_{j=1}^m
\]

Documents

\[
\text{Learning algorithm}
\]

\[
\hat{f}_{m,n}
\]

Mapping between Documents and topics

Active Learning – **selectively** labeled examples

\[
\{(X_i, Y_i)\}_{i=1}^n, \quad \{X_j\}_{j=1}^m
\]

Documents, topics

\[
\text{Selectively labeling}
\]

\[
\text{Learning algorithm}
\]

\[
\hat{f}_{m,n}
\]
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

Regression: (Continuous labels, Mean Square Error)

Optimal estimation rule

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

Linear Regression \( f(X) = Xw, \quad X = [x_1, x_2, \ldots, x_d] \)

Polynomial Regression \( X = [x_1^2, x_1x_2, x_2^2, \ldots] \)

Basis Regression \( X = [\phi_1(x), \phi_2(x), \ldots, \phi_d(x)] \)

Regularized versions (MAP)

Neural Networks \( f(X) = \text{nonlinear (combination of multiple logistic units)} \)

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

<table>
<thead>
<tr>
<th>Gen/Disc</th>
<th>K-NN</th>
<th>Gauss Naïve Bayes</th>
<th>Logistic Regression</th>
<th>Neural Nwks</th>
<th>HMM</th>
<th>Bayes Net</th>
<th>SVM</th>
<th>Boosting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss functions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision boundary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Output</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relation to others</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parametric/Nonparam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some Topics We’ve Covered (before Midterm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Decision trees</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>entropy, mutual info., overfitting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Probability basics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes rule, MLE, MAP, conditional indep.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Naïve Bayes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conditional independence,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># of parameters to estimate,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decision surface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Logistic regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>form of $P(Y</td>
<td>X)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>generative vs. discriminative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Linear Regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimizing sum sq. error (why?)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>regularization $\sim$ MAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sources of Error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unavoidable error, bias, variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overfitting, and Avoiding it</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>priors over $H$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cross validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAC theory: probabilistic bound on overfitting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bayesian Networks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>factored <em>representation</em> of joint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>distribution, conditional independence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>assumptions, D-separation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>inference</em> in Bayes nets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>learning</em> from fully/partly observed data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PAC Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>probabilistic bounds on $\text{error}<em>{\text{train}} - \text{error}</em>{\text{true}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC dimension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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: $\text{error}_{\text{true}} - \text{error}_{\text{train}}$
Bias and Variance of Estimators

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

the bias of estimator Y : \( E[Y] - \theta \)
the variance of estimator Y : \( E[(Y - E[Y])^2] \)

consider when

- \( \theta \) is the probability of “heads” for my coin
- \( Y \) = proportion of heads observed from 3 flips

consider when

- \( \theta \) 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 \rightarrow 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

Regular
• 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

New Spring 2012
• 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