Structured Output Prediction

CS4780/5780 – Machine Learning
Fall 2013
Thorsten Joachims
Cornell University

Reading:

Discriminative vs. Generative

Bayes Decision Rule
- \( h_{bayes}(x) = \arg\max_{y} P(Y = y | X = x) \)
- \( = \arg\max_{y} \left[ P(X = x | Y = y) P(Y = y) \right] \)

Generative:
- Idea: Make assumptions about \( P(X = x | Y = y), P(Y = y) \)
- Method: Estimate parameters of the two distributions, then apply Bayes decision rule.

Discriminative:
- Idea: Define set of prediction rules (i.e. hypotheses) \( H \), then search for \( h \in H \) that best approximates
- Method: find \( h \in H \) that minimizes training error.

Question: Can we train HMM’s discriminately?

Idea for Discriminative Training of HMM

Start:
- \( h_{bayes}(x) = \arg\max_{y} \left[ P(Y = y | X = x) \right] \)
- \( = \arg\max_{y} \left[ P(X = x | Y = y) P(Y = y) \right] \)

Idea:
- Model \( P(Y = y | X = x) \) with \( \bar{w} \cdot \phi(x, y) \) so that
- \( \left( \arg\max_{y} \left[ P(Y = y | X = x) \right] \right) = \left( \arg\max_{y} \left[ P(Y = y | X = x) \right] \right) \)

Intuition:
- Tune \( \bar{w} \) so that correct \( y \) has the highest value of \( \bar{w} \cdot \phi(x, y) \)
- \( \phi(x, y) \) is a feature vector that describes the match between \( x \) and \( y \)

Training HMMs with Structural SVM

• Define \( \phi(x, y) \) so that model is isomorphic to HMM
  – One feature for each possible start state
  – One feature for each possible transition
  – One feature for each possible output in each possible state
  – Feature values are counts

Structural Support Vector Machine

• Joint features \( \phi(x, y) \) describe match between \( x \) and \( y \)
• Learn weights \( \bar{w} \) so that \( \bar{w} \cdot \phi(x, y) \) is max for correct \( y \)

Structural SVM Training Problem

Hard-margin optimization problem:
\[
\min_{\bar{w}} \frac{1}{2} \bar{w}^T \bar{w} \\
\text{s.t.} \quad y \in Y \setminus y_1 : \bar{w}^T \Phi(x_1, y_1) \geq \bar{w}^T \Phi(x_1, y) + 1 \\
... \\
\quad y \in Y \setminus y_n : \bar{w}^T \Phi(x_n, y_n) \geq \bar{w}^T \Phi(x_n, y) + 1
\]

• Training Set: \( (x_1, y_1), ..., (x_n, y_n) \)
• Prediction Rule: \( h_{svm}(x) = \arg\max_{y} \Phi(x, y) \)
• Optimization:
  – Correct label \( y \) must have higher value of \( \bar{w} \cdot \phi(x, y) \) than any incorrect label \( y \)
  – Find weight vector with smallest norm
Soft-Margin Structural SVM

• Loss function $\Delta(y_i, y)$ measures match between target and prediction.

![Soft-Margin Structural SVM Diagram]

Experiment: Part-of-Speech Tagging

• Task
  - Given a sequence of words $x$, predict sequence of tags $y$.
  - $\underbrace{x_1 \ x_2 \ x_3 \ y_1 \ y_2 \ y_3}_{\text{3 The dog chased the cat}} \underbrace{x_4 \ x_5 \ x_6 \ y_4 \ y_5 \ y_6}_{\text{Det N V Det N Det N}}$

• Model
  - Markov model with one state per tag and words as emissions
  - Each word described by ~250,000 dimensional feature vector (all word suffixes/prefixes, word length, capitalization ...)

• Experiment (by Dan Fleisher)
  - Train/test on 7966/1700 sentences from Penn Treebank

![Part-of-Speech Tagging Experiment]

NE Identification

• Identify all named locations, named persons, named organizations, dates, times, monetary

Cutting-Plane Algorithm for Structural SVM

• Input: $(x_1, y_1), \ldots, (x_n, y_n), C, \varepsilon$

  - $S \leftarrow \emptyset$, $\mathcal{W} \leftarrow 0$, $\xi \leftarrow 0$

  - REPEAT
    - FOR $i = 1, \ldots, n$
      - compute $\hat{y} = \arg \max_{y \in Y} \{\Delta(y, \hat{y}) + \omega^T \Phi(x_i, y)\}$
      - IF $\Delta(y_i, \hat{y}) > \omega^T \Phi(x_i, \hat{y}) - \xi$
        - $S \leftarrow S \cup \{y_i\}$
        - $S \leftarrow \{y_i\}$
        - $\Delta(y_i, \hat{y}) - \xi$
      - ENDIF
    - ENDFOR
  - UNTIL $S$ has not changed during iteration

Find most violated constraint

Violated by more than $\varepsilon$?

Polynomial Time Algorithm (SVM-struct)
General Problem: Predict Complex Outputs

• Supervised Learning from Examples
  – Find function from input space $X$ to output space $Y$
    $$h : X \rightarrow Y$$
  such that the prediction error is low.

• Typical
  – Output space is just a single number
    • Classification: $-1,+1$
    • Regression: some real number

• General
  – Predict outputs that are complex objects

Examples of Complex Output Spaces

• Natural Language Parsing
  – Given a sequence of words $x$, predict the parse tree $y$.
  – Dependencies from structural constraints, since $y$ has to be a tree.

Examples of Complex Output Spaces

• Multi-Label Classification
  – Given a (bag-of-words) document $x$, predict a set of labels $y$.
  – Dependencies between labels from correlations between labels ("iraq" and "oil" in newswire corpus)

Examples of Complex Output Spaces

• Noun-Phrase Co-reference
  – Given a set of noun phrases $x$, predict a clustering $y$.
  – Structural dependencies, since prediction has to be an equivalence relation.
  – Correlation dependencies from interactions.

Examples of Complex Output Spaces

• Scene Recognition
  – Given a 3D point cloud with RGB from Kinect camera
  – Segment into volumes
  – Geometric dependencies between segments (e.g. monitor usually close to keyboard)

Wrap-Up
# Classification

- **Discriminative**
  - Decision Trees
  - Perceptron
  - Linear SVMs
  - Kernel SVMs

- **Generative**
  - Multinomial Naïve Bayes
  - Multivariate Naïve Bayes
  - Less Naïve Bayes
  - Linear Discriminant
  - Nearest Neighbor

- **Other Methods**
  - Logical rule learning
  - Online Learning
  - Logistic Regression
  - Neural Networks
  - RBF Networks
  - Boosting
  - Bagging
  - Parametric (Graphical) Models
  - Non-Parametric Models
  - *-Regression
  - *-Multiclass

# Structured Prediction

- **Discriminative**
  - Structural SVMs

- **Generative**
  - Hidden Markov Model

- **Other Methods**
  - Maximum Margin Markov Networks
  - Conditional Random Fields
  - Markov Random Fields
  - Bayesian Networks
  - Statistical Relational Learning

# Unsupervised Learning

- **Clustering**
  - Hierarchical Agglomerative Clustering
  - K-Means
  - Mixture of Gaussians and EM-Algorithm

- **Other Methods**
  - Spectral Clustering
  - Latent Dirichlet Allocation
  - Latent Semantic Analysis
  - Multi-Dimensional Scaling

- **Other Tasks**
  - Outlier Detection
  - Novelty Detection
  - Dimensionality Reduction
  - Non-Linear Manifold Detection

# Other Learning Problems and Applications

- **Recommender Systems, Search Ranking, etc.**

- **Reinforcement Learning and Markov Decision Processes**
  - CS4758 Robot Learning

- **Computer Vision**
  - CS4670 Intro Computer Vision

- **Natural Language Processing**
  - CS4740 Intro Natural Language Processing

# Other Machine Learning Courses at Cornell

- INFO 3300 - New course by David Mimno
- CS 4700 - Introduction to Artificial Intelligence
- CS 4780/5780 - Machine Learning
- CS 4758 - Robot Learning
- CS 4782 - Probabilistic Graphical Models
- OR 4740 - Statistical Data Mining
- OR 4740 - Statistical Data Mining
- CS 6756 - Advanced Topics in Robot Learning: 3D Perception
- CS 6780 - Advanced Machine Learning
- CS 6784 - Advanced Topics in Machine Learning
- ORIE 6740 - Statistical Learning Theory for Data Mining
- ORIE 6750 - Optimal learning
- ORIE 6780 - Bayesian Statistics and Data Analysis
- ORIE 6127 - Computational Issues in Large Scale Data-Driven Models
- BTRY 6502 - Computationally Intensive Statistical Inference
- MATH 7740 - Statistical Learning Theory