Seven Lectures on Statistical Parsing

Christopher Manning
LSA Linguistic Institute 2007
LSA 354
Lecture 7

The Classification Problem

- Given a training set of iid samples T={(X_1,Y_1) … (X_n,Y_n)} of input and class variables from an unknown distribution D(X,Y), estimate a function $\hat{h}(X)$ that predicts the class from the input variables
- The goal is to come up with a hypothesis $\hat{h}(X)$ with minimum expected loss

$$err(\hat{h}) = \sum_{<X,Y>\in\Omega} D(X,Y)\delta(Y \neq \hat{h}(X))$$

- Under 0-1 loss the hypothesis with minimum expected loss is the Bayes optimal classifier

$$h(X) = \arg\max_{Y\in\mathrm{Y}} D(Y \mid X)$$

Discriminative Parsing as a classification problem

- The observed X's are the sentences.
- The class Y of a sentence is its parse tree
- The model has a large (infinite!) space of variables, but we can still assign them probabilities
  - The way we can do this is by breaking whole parse trees into component parts

Approaches to Solving Classification Problems

1. Generative. Try to estimate the probability distribution of the data D(X,Y)
   - specify a parametric model family $\{P_\theta(X,Y) : \theta \in \Theta\}$
   - choose parameters $\hat{\theta}$ by maximum likelihood on training data

$$L(T \mid \theta) = \prod_{i=1}^{n} P_\theta(X_i, Y_i)$$

   - estimate conditional probabilities by Bayes rule
     - You use the generative model "backwards"
   - classify new instances to the most probable class Y according to

$$P_{\hat{\theta}}(Y \mid X) = P_{\hat{\theta}}(X,Y) / P_{\hat{\theta}}(X)$$
$$P_{\hat{\theta}}(Y \mid X)$$

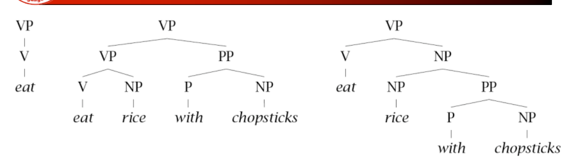
Approaches to Solving Classification Problems

2. Discriminative. Try to estimate the conditional distribution D(Y|X) from data.
   - specify a parametric model family $\{P_\theta(Y \mid X) : \theta \in \Theta\}$
   - estimate parameters $\hat{\theta}$ by maximum conditional likelihood of training data

$$CL(T \mid \theta, X) = \prod_{i=1}^{n} P_\theta(Y_i \mid X_i)$$

   - classify new instances to the most probable class Y according to $P_{\hat{\theta}}(Y \mid X)$
3. Discriminative. Distribution-free. Try to estimate $\hat{h}(X)$ directly from data so that its expected loss will be minimized

Motivating discriminative estimation (1)



100 6 2

A training corpus of 108 (imperative) sentences.

Follows an example by Mark Johnson

Motivating discriminative parsing (2)

- In discriminative models, it is easy to incorporate different kinds of features
  - Often just about anything that seems linguistically interesting
- In generative models, it's often difficult, and the model suffers because of false independence assumptions
- This ability to add informative features is the real power of discriminative models for NLP.

Discriminative Parsers

- Discriminative Dependency Parsing
  - Not as computationally hard (tiny grammar constant)
  - Explored considerably recently. E.g. McDonald et al. 2005
- Make parser action decisions discriminatively
  - E.g. with a shift-reduce parser
- Dynamic program Phrase Structure Parsing
  - Resource intensive! Most work on sentences of length <=15
  - The need to be able to dynamic program limits the feature types you can use
- Post-Processing: Parse reranking
  - Just work with output of k-best generative parser

1. Distribution-free methods
2. Probabilistic model methods

Discriminative models

Shift-reduce parser Ratnaparkhi (98)

- Learns a distribution P(T|S) of parse trees given sentences using the sequence of actions of a shift-reduce parser

$$P(T \mid S) = \prod_{i=1}^{n} P(a_i \mid a_1 \ldots a_{i-1} S)$$

- Uses a maximum entropy model to learn conditional distribution of parse action given history
- Suffers from independence assumptions that actions are independent of future observations as CMM
- Higher parameter estimation cost to learn local maximum entropy models
- **Lower** but still good accuracy 86% - 87% labeled precision/recall

Discriminative dynamic-programmed parsers

- Taskar et al. (2004 EMNLP) show how to do joint discriminative SVM-style ("max margin) parsing building a phrase structure tree also conditioned on words in O(n^3) time
  - In practice, totally impractically slow. Results were never demonstrated on sentences longer than 15 words
- Turian et al. (2006 NIPS) do a decision-tree based discriminative parser
- Research continues....

Discriminative Models – Distribution Free Re-ranking (Collins 2000)

- Represent sentence-parse tree pairs by a feature vector F(X,Y)
- Learn a linear ranking model with parameters $\overline{\alpha}$ using the boosting loss

Model	LP	LR
Collins 99 (Generative)	88.3%	88.1%
Collins 00 (BoostLoss)	89.9%	89.6%

13% error reduction

Still very close in accuracy to generative model [Charniak 2000]

Charniak and Johnson (2005 ACL): Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking

- Builds a maxent discriminative reranker over parses produced by (a slightly bugfixed and improved version of) Charniak (2000).
- Gets 50 best parses from Charniak (2000) parser
  - Doing this exploits the "coarse-to-fine" idea to heuristically find good candidates
- Maxent model for reranking uses heads, etc. as generative model, but also nice linguistic features:
  - Conjunct parallelism
  - Right branching preference
  - Heaviness (length) of constituents factored in
- Gets 91% LP/LR F1 (on *all* sentences! – up to 80 wd)