Building Maximum Entropy Text Classifier Using Semi-supervised Learning

Zhang, Xinhua

For PhD Qualifying Exam Term Paper

Road map

- Introduction: background and application
- Semi-supervised learning, especially for text classification (survey)
- Maximum Entropy Models (survey)
- Combining semi-supervised learning and maximum entropy models (new)
- Summary

Road map

- Introduction: background and application
- Semi-supervised learning, esp. for text classification (survey)
- Maximum Entropy Models (survey)
- Combining semi-supervised learning and maximum entropy models (new)
- Summary

Introduction: Application of text classification

- Text classification is useful, widely applied:
  - cataloging news articles (Lewis & Gale, 1994; Joachims, 1998b);
  - classifying web pages into a symbolic ontology (Craven et al., 2000);
  - finding a person's homepage (Shavlik & Eliassi-Rad, 1998);
  - automatically learning the reading interests of users (Lang, 1995; Pazzani et al., 1996);
  - automatically threading and filtering email by content (Lewis & Knowles, 1997; Sahami et al., 1998);
  - book recommendation (Mooney & Roy, 2000).

Early ways of text classification

- Early days: manual construction of rule sets.
  (e.g., if *advertisement* appears, then filtered).

- Hand-coding text classifiers in a rule-based style is impractical. Also, inducing and formulating the rules from examples are time and labor consuming.

Supervised learning for text classification

- Using supervised learning
  - Require a large or prohibitive number of labeled examples, time/labor-consuming.
  - E.g., (Lang, 1995) after a person read and hand-labeled about 1000 articles, a learned classifier achieved an accuracy of about 50% when making predictions for only the top 10% of documents about which it was most confident.

What about using unlabeled data?

- Unlabeled data are abundant and easily available, may be useful to improve classification.
  - Published works prove that it helps.
- Why do unlabeled data help?
  - Co-occurrence might explain something.
  - Search on *Google*,
    - ‘*Sugar and sauce*’ returns 1,390,000 results
    - ‘*Sugar and math’* returns 191,000 results

    though *math* is a more popular word than *sauce*

Using co-occurrence and pitfalls

- Simple idea: when A often co-occurs with B (a fact that can be found by using unlabeled data) and we know articles containing A are often interesting, then probably articles containing B are also interesting.
- Problem:
  - Most current models using unlabeled data are based on problem-specific assumptions, which causes instability across tasks.

Road map

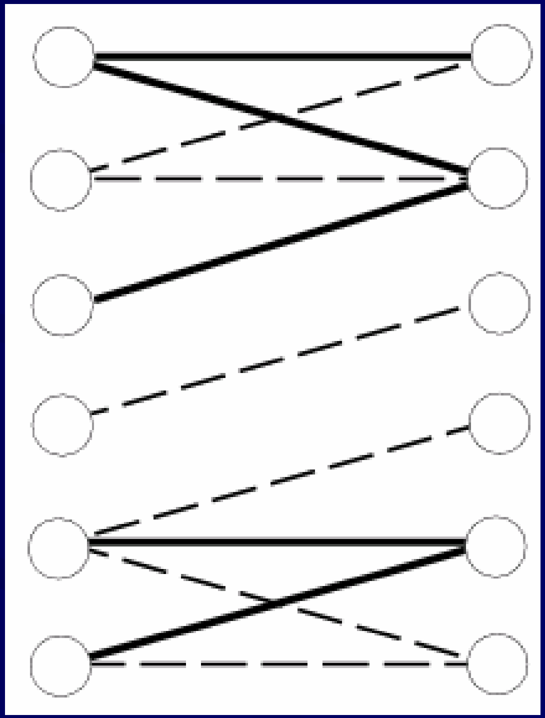
- Introduction: background and application
- **Semi-supervised learning, especially for text classification (survey)**
- Maximum Entropy Models (survey)
- Combining semi-supervised learning and maximum entropy models (new)
- Summary

Generative and discriminative semi-supervised learning models

- Generative semi-supervised learning (Nigam, 2001)
  - Expectation-maximization algorithm, which can fill the missing value using maximum likelihood
- Discriminative semi-supervised learning (Vapnik, 1998)
  - Transductive Support Vector Machine (TSVM)
    - finding the linear separator between the labeled examples of each class that maximizes the margin over both the labeled and unlabeled examples

Other semi-supervised learning models

- Co-training
  (Blum & Mitchell, 1998)
- Active learning
  e.g., (Schohn & Cohn, 2000)
- Reduce overfitting
  e.g. (Schuurmans
  & Southey, 2000)



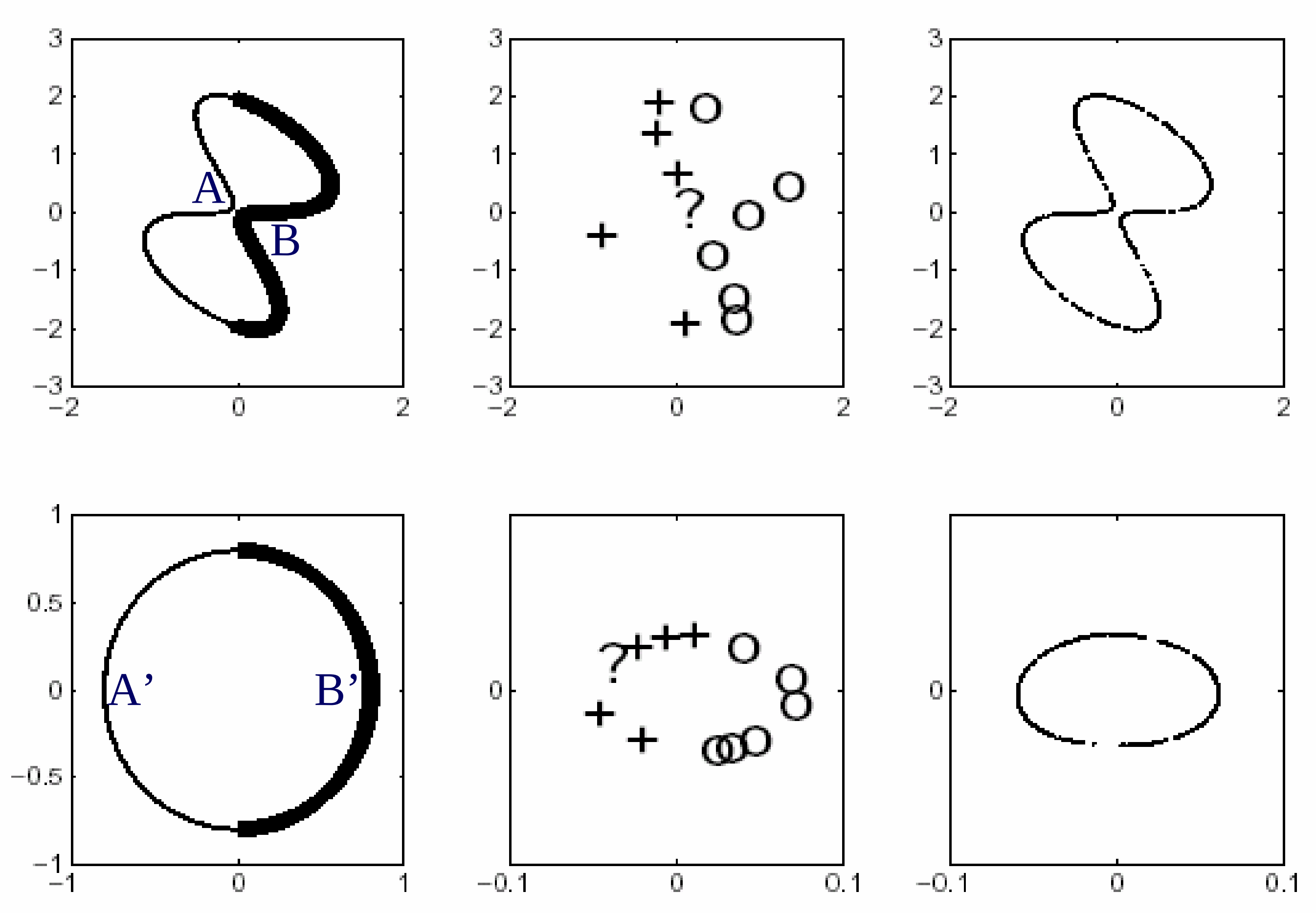
Theoretical value of unlabeled data

- Unlabeled data help in some cases, but not all.
- For class probability parameters estimation, labeled examples are exponentially more valuable than unlabeled examples, assuming the underlying component distributions are known and correct. (Castelli & Cover, 1996)
- Unlabeled data can degrade the performance of a classifier when there are incorrect model assumptions. (Cozman & Cohen, 2002)
- Value of unlabeled data for discriminative classifiers such as TSVMs and for active learning are questionable. (Zhang & Oles, 2000)

Models based on clustering assumption (1): Manifold

- Example: handwritten 0 as an ellipse (5-Dim)
- Classification functions are naturally defined only on the submanifold in question rather than the total ambient space.
- Classification will be improved if the convert the representation into submanifold.
  - Same idea as PCA, showing the use of unsupervised learning in semi-supervised learning
- Unlabeled data help to construct the submanifold.

Manifold, unlabeled data help



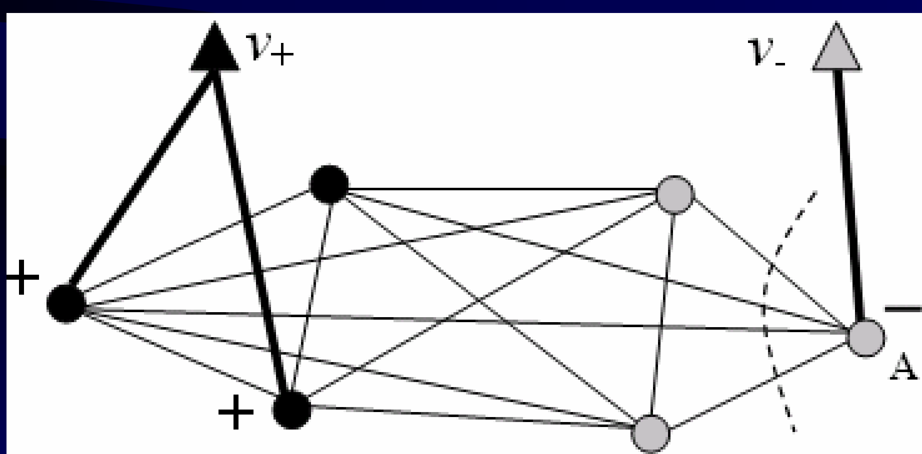
Belkin & Niyogi 2002

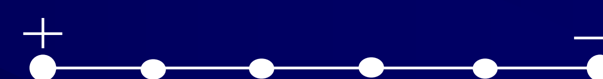
Models based on clustering assumption (2): Kernel methods

- Objective:
  - make the induced distance small for points in the same class and large for those in different classes
  - Example:
    - Generative: for a mixture of Gaussian (μ_k, Σ_k) one kernel can be defined as $K(x,y)=\sum_{k=1}^{q} P(k\,|\,x)P(k\,|\,y)x^T\Sigma_k^{-1}y$ (Tsuda et al., 2002)
    - Discriminative: RBF kernel matrix $K_{ij}=\exp\left(-\|x_i-x_j\|/\sigma^2\right)$
- Can unify the manifold approach

Models based on clustering assumption (3): Min-cut

- Express pair-wise relationship (similarity) between labeled/unlabeled data as a graph, and find a partitioning that minimizes the sum of similarity between differently labeled examples.





Min-cut family algorithm

- Problems with min-cut
  - Degenerative (unbalanced) cut
- Remedy
  - Randomness
  - Normalization, like Spectral Graph Partitioning
  - Principle:

Averages over examples (e.g., average margin, pos/neg ratio) should have the same expected value in the labeled and unlabeled data.

Road map

- Introduction: background and application
- Semi-supervised learning, esp. for text classification (survey)
- **Maximum Entropy Models (survey)**
- Combining semi-supervised learning and maximum entropy models (new)
- Summary

Overview: Maximum entropy models

- Advantage of maximum entropy model
  - Based on features, allows and supports feature induction and feature selection
  - offers a generic framework for incorporating unlabeled data
  - only makes weak assumptions
  - gives flexibility in incorporating side information
  - natural multi-class classification
- So maximum entropy model is worth further study.

Feature in MaxEnt

- Indicate the strength of certain aspects in the event
  - *e.g.*, $f_t(x, y) = 1$ if and only if the current word, which is part of document x, is “back” and the class y is verb. Otherwise, $f_t(x, y) = 0$.
- Contributes to the flexibility of MaxEnt

$\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i)$

Standard MaxEnt Formulation

maximize $\quad -\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i)$

s.t. $\quad \sum_i p(x_i)\sum_k p(y_k \mid x_i)\, f_t(x_i, y_k)$

$$= \sum_i \tilde{p}(x_i)\sum_k \tilde{p}(y_k \mid x_i)\, f_t(x_i, y_k) \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

The dual problem is just the *maximum likelihood* problem.

Smoothing techniques (1)

- Gaussian prior (MAP)

$$\text{maximize} \quad -\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i) + \sum_t \frac{\sigma_t^2}{2}\delta_t^2$$

$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) = \delta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Smoothing techniques (2)

- Laplacian prior (Inequality MaxEnt)

maximize $\quad -\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$

s.t. $\quad E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t \quad$ for all t

$$-\sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) - E_{\tilde{p}}[f_t] \le B_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Extra strength: feature selection.

MaxEnt parameter estimation

- Convex optimization ☺
- Gradient descent, (conjugate) gradient descent
- Generalized Iterative Scaling (GIS)
- Improved Iterative Scaling (IIS)
- Limited memory variable metric (LMVM)
- Sequential update algorithm

Road map

- Introduction: background and application
- Semi-supervised learning, esp. for text classification (survey)
- Maximum Entropy Models (survey)
- **Combining semi-supervised learning and maximum entropy models (new)**
- Summary

Semi-supervised MaxEnt

- Why do we choose MaxEnt?
  - 1st reason: simple extension to semi-supervised learning

$$\text{maximize } -\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i)$$

$$s.t.\ \ E_{\tilde{p}}[f_t]-\sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k)=0 \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i)=1 \quad \text{for all } i$$

$$\text{where} \qquad E_{\tilde{p}}[f_t]=\sum_i \tilde{p}(x_i)\sum_k \tilde{p}(y_k \mid x_i) f_t(x_i, y_k)$$

  - 2nd reason: weak assumption

Estimation error bounds

- 3rd reason: estimation error bounds in theory

$$\text{maximize} \quad -\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i)$$

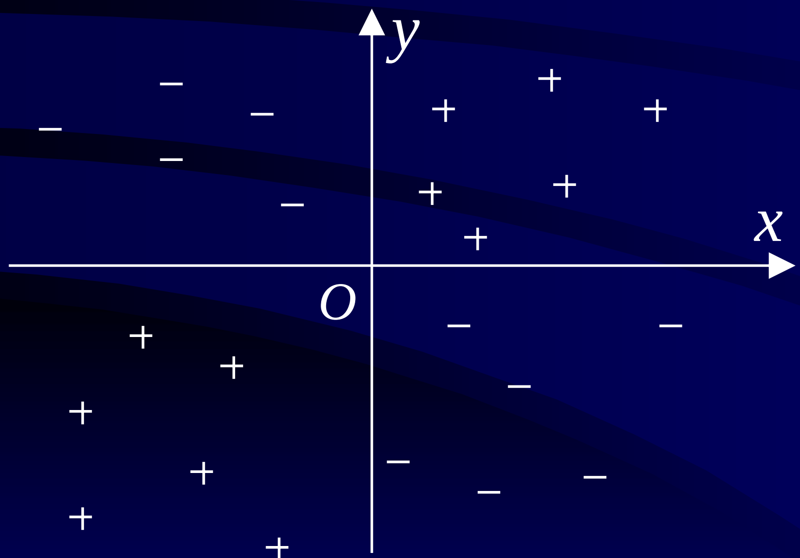
$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t \quad \text{for all } t$$

$$\sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) - E_{\tilde{p}}[f_t] \le B_t \quad \text{for all } t$$

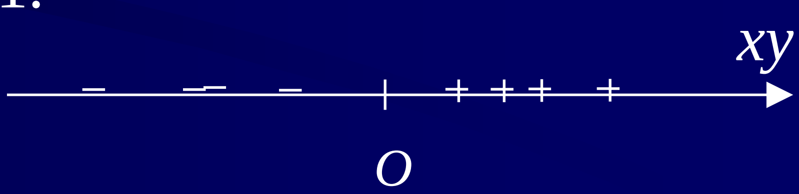
$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Side Information

- Only assumptions over the accuracy of empirical evaluation of sufficient statistics is not enough



1.



2. Use distance/similarity info

Source of side information

- Instance similarity.
  - neighboring relationship between different instances
  - redundant description
  - tracking the same object
- Class similarity, using information on related classification tasks
  - combining different datasets (different distributions) which are for the same classification task;
  - hierarchical classes;
  - structured class relationships (such as trees or other generic graphic models)

Incorporate similarity information: flexibility of MaxEnt framework

- Add assumption that the class probability of x_i, x_j is similar if the distance in one metric is small between x_i, x_j.
- Use the distance metric to build a minimum spanning tree and add side info to MaxEnt. Maximize:

$$-\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i) - \sum_{k,(i,j)\in E} w_{k,(i,j)}\varepsilon_{i,j,k}^2$$

$$E_{\tilde{p}}[f_t] = \sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

$$p(y_k \mid x_i) - p(y_k \mid x_j) = \varepsilon_{i,j,k} \quad \text{for all } k \text{ and } (i,j)\in E$$

$w_{k,(i,j)} \triangleq C_s / w_{(i,j)}$ where $w_{(i,j)}$ is the true distance between (x_i, x_j)

Connection with Min-cut family

- Spectral Graph Partitioning

$$\max_{\vec{y}} \frac{cut(G^+, G^-)}{\left|\{i \mid y_i = 1\}\right| \cdot \left|\{i \mid y_i = -1\}\right|}$$

$$s.t. \quad \begin{array}{l} y_i = +1 \quad \text{if } x_i \text{ is positively labeled} \\ y_i = -1 \quad \text{if } x_i \text{ is negatively labeled} \\ \vec{y} \in \{+1, -1\}^n \end{array}$$

Harmonic function
(Zhu et al. 2003)

minimize

$$\frac{1}{2} \sum_{i,j} w_{ij} \left(P(y_i = 1) - P(y_j = 1) \right)^2$$

$$\text{maximize} \quad -\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i) - \sum_{k,(i,j)\in E} w_{k,(i,j)} \varepsilon_{i,j,k}^2 \qquad |\varepsilon_{i,j,k}| \ ?$$

$$E_{\tilde{p}}[f_t] = \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

$$p(y_k \mid x_i) - p(y_k \mid x_j) = \varepsilon_{i,j,k} \quad \text{for all } k \text{ and } (i,j) \in E$$

Miscellaneous promising research openings (1)

- Feature selection
  - Greedy algorithm to incrementally add feature to the random field by selecting the feature which maximally reduces the objective function.
- Feature induction
  - If *IBM* appears in labeled data while *Apple* does not, then using '*IBM* or *Apple*' as feature can help (though costly).

Miscellaneous promising research openings (2)

- Interval estimation

$$\text{minimize} \quad \sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$$

$$s.t. \quad -B_t \le E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

– How should we set the A_t and B_t? Whole bunch of results in statistics. W/S LLN, Hoeffding's inequality

$$P\left(\left|E_p[f_t] - E_{\tilde{p}}[f_t]\right| > \beta\right) \le \exp(-2\beta^2 m)$$

or using more advanced concepts in statistical learning theory, e.g., VC-dimension of feature class

Miscellaneous promising research openings (3)

- Re-weighting
  - In view that the empirical estimation of statistics is inaccurate, we add more weight to the labeled data, which may be more reliable than unlabeled data.

$$\text{minimize} \quad \sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i) + \sum_t \frac{\sigma_t^2}{2} \delta_t^2$$

$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) = \delta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Re-weighting

- Originally, n_1 labeled examples and n_2 unlabeled examples

$$\underbrace{x_1^l, x_2^l, \ldots, x_{n_1}^l}_{n_1 \text{ labeled data}}, \underbrace{x_1^u, x_2^u, \ldots, x_{n_2}^u}_{n_2 \text{ unlabeled data}} \Rightarrow \underbrace{\overbrace{x_1^l, \ldots x_1^l}^{\beta \text{ copies of } x_1^l}, \ldots, \overbrace{x_{n_1}^l, \ldots x_{n_1}^l}^{\beta \text{ copies of } x_{n_1}^l}}_{\beta n_1 \text{ labeled data}}, \underbrace{x_1^u, x_2^u, \ldots, x_{n_2}^u}_{n_2 \text{ unlabeled data}}$$

Then $p(x)$ for labeled data: $\frac{1}{n_1 + n_2} \Rightarrow \frac{\beta}{\beta n_1 + n_2}$

$p(x)$ for unlabeled data: $\frac{1}{n_1 + n_2} \Rightarrow \frac{1}{\beta n_1 + n_2}$

All equations before keep unchanged!

Initial experimental results

- Dataset: optical digits from UCI
  - 64 input attributes ranging in [0, 16], 10 classes
- Algorithms tested
  - MST MaxEnt with re-weight
  - Gaussian Prior MaxEnt, Inequality MaxEnt, TSVM (linear and polynomial kernel, one-against-all)
- Testing strategy
  - Report the results for the parameter setting with the best performance on the test set

Initial experiment result

No. of Labeled data	No. of Unlabeled Data	Re-weighted MST MaxEnt Accuracy	Gaussian MaxEnt Accuracy	Inequality MaxEnt Accuracy	TSVM Result
39	3894	93.8352	59.57	58.80	73.444
39	0	85.5957	77.53	77.59	
78	3855	94.7244	79.37	70.59	84.766
78	0	88.2039	87.72	86.01	
117	3816	94.8429	84.77	77.53	84.766
117	0	91.8791	89.27	88.20	
156	3777	95.3764	87.31	81.21	90.279
156	0	92.4718	90.10	89.27	
196	3737	96.5027	89.27	85.12	89.627
196	0	91.7012	92.95	90.75	

Summary

- Maximum Entropy model is promising for semi-supervised learning.
- Side information is important and can be flexibly incorporated into MaxEnt model.
- Future research can be done in the area pointed out (feature selection/induction, interval estimation, side information formulation, re-weighting, etc).

Question and Answer Session

Questions
are
welcomed.

GIS

- Iterative update rule for unconditional probability:

$$\lambda_t^{(s+1)} = \lambda_t^{(s)} + \log\left(\frac{E_{\tilde{p}}[f_t]}{E_{p^{(s)}}[f_t]}\right) \qquad p^{(s+1)}(x_i) = p^{(s)}(x_i)\prod_t \left(\frac{\sum_j \tilde{p}(x_j) f_t(x_j)}{\sum_j p^{(s)}(x_j) f_t(x_j)}\right)^{f_t(x_i)}$$

- GIS for conditional probability

$$\lambda_t^{(s+1)} = \lambda_t^{(s)} + \eta \log\left(\frac{E_{\tilde{p}}[f_t]}{\sum_i \tilde{p}(x_i)\sum_k p(y_k \mid x_i, \lambda^{(s)}) f_t(x_i, y_k)}\right)$$

$$= \lambda_t^{(s)} + \eta \log\left(\frac{E_{\tilde{p}}[f_t]}{E_{p^{(s)}}[f_t]}\right)$$

IIS

- Characteristic:
  - monotonic decrease of MaxEnt objective function
  - each update depends only on the computation of expected values $E_{p^{(s)}}$, not requiring the gradient or higher derivatives
- Update rule for unconditional probability:
  - $\Delta\lambda_t$ is the solution to:

$$E_{\tilde{p}}[f_t] = \sum_i p^{(s)}(x_i) f_t(x_i) \exp\left(\Delta\lambda_t \sum_j f_j(x_i)\right) \quad \text{for all } t$$

  - $\Delta\lambda_t$ are decoupled and solved individually
  - Monte Carlo methods are to be used if the number of possible x_i is too large

GIS

- Characteristics:
  - converges to the unique optimal value of λ
  - parallel update, i.e., $\lambda_t^{(s)}$ are updated synchronously
  - slow convergence
- prerequisite of original GIS
  - for all training examples x_i: $f_t(x_i) \geq 0$ and $\sum_t f_t(x_i) = 1$
  - relaxing prerequisite

    if $\sum_t f_t(x_i) = C$ then define $f_t' = f_t / C$

  If not all training data have summed feature equaling C, then set C sufficiently large and incorporate a 'correction feature'.

Other standard optimization algorithms

- Gradient descent

$$\lambda_t^{(s+1)} = \lambda_t^{(s)} + \eta \frac{\partial L}{\partial \lambda_t}\bigg|_{\lambda = \lambda^{(s)}}$$

- *Conjugate gradient* methods, such as *Fletcher-Reeves* and *Polak-Ribière-Positive* algorithm
- *limited memory variable metric, quasi-Newton* methods: approximate Hessian using successive evaluations of gradient

Sequential updating algorithm

- For a very large (or infinite) number of features, parallel algorithms will be too resource consuming to be feasible.
- Sequential update: A style of coordinate-wise descent, modifies one parameter at a time.
- Converges to the same optimum as parallel update.

$\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$

Dual Problem of Standard MaxEnt

minimize $$\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$$

$$E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) = 0 \quad \text{for all } t$$

$\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Dual problem: $$L(p_{\min}, \lambda) = -\sum_t \lambda_t E_{\tilde{p}}[f_t] + \sum_i p(x_i) \log Z_i$$

where $$Z_i = \sum_k \exp\left(\sum_t \lambda_t f_t(x_i, y_k) \right)$$

$\sum_i p(x_i)\log p(x_i) + \sum_i \lambda_i E_{\tilde{p}}[f_i] - \sum_i p(x_i)\log Z_i$

Relationship with maximum likelihood

Suppose
$$p(y_k \mid x_i) = \frac{1}{Z_i}\exp\left(\sum_t \lambda_t f_t(x_i, y_k)\right)$$

where
$$Z_i = \sum_k \exp\left(\sum_t \lambda_t f_t(x_i, y_k)\right)$$

$$L(\lambda) = \sum_i \sum_k \tilde{p}(x_i, y_k)\log p(x_i, y_k) \qquad \leftarrow \text{maximize}$$

$$= \sum_i \tilde{p}(x_i)\log \tilde{p}(x_i) + \sum_t \lambda_t E_{\tilde{p}}[f_t] - \sum_i p(x_i)\log Z_i$$

Dual of MaxEnt:
$$L(p_{\min}, \lambda) = -\sum_t \lambda_t E_{\tilde{p}}[f_t] + \sum_i p(x_i)\log Z_i \qquad \leftarrow \text{minimize}$$

Smoothing techniques (2)

- Exponential prior

minimize $$\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$$

$Z_i = \sum \exp\left(\sum \lambda_t f_t(x_i, y_k)\right)$

s.t. $$E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Dual problem: minimize

$$L(\lambda) = -\sum_t \lambda_t E_{\tilde{p}}[f_t] + \sum_i p(x_i) \log Z_i + \sum_t A_t \lambda_t$$

$$Z_i = \sum_k \exp\left(\sum_t \lambda_t f_t(x_i, y_k)\right)$$

Equivalent To maximize

$$\prod_i p(\tilde{y}_i \mid x_i) \times \prod_t A_t \exp(-A_t \lambda_t)$$

Smoothing techniques (1)

- Gaussian prior (MAP)

$$\text{minimize} \quad \sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i) + \sum_t \frac{\sigma_t^2}{2} \delta_t^2$$

$Z_i = \sum_k \exp\left(\sum_t \lambda_t f_t(x_i, y_k) \right)$

$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) = \delta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Dual problem: minimize

$$L(\lambda) = -\sum_t \lambda_t E_{\tilde{p}}[f_t] + \sum_i \tilde{p}(x_i) \log(Z_i) + \sum_t \frac{\lambda_t^2}{2\sigma_t^2}$$

$$Z_i = \sum_k \exp\left(\sum_t \lambda_t f_t(x_i, y_k) \right)$$

Smoothing techniques (3)

- Laplacian prior (Inequality MaxEnt)

minimize $\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i)$

$s.t.$ $-B_t \le E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t$ for all t

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Dual problem: minimize

$$L(\alpha, \beta) = -\sum_t (\alpha_t - \beta_t) E_{\tilde{p}}[f_t] + \sum_i p(x_i) \log Z_i + \sum_t A_t \alpha_t + \sum_t B_t \beta_t \qquad \alpha_t \ge 0, \beta_t \ge 0$$

where $Z_i = \sum_k \exp(\sum_t (\alpha_t - \beta_t) f_t(x_i, y_k))$

Smoothing techniques (4)

- Inequality with 2-norm Penalty

minimize

$$\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i) + C_1\sum_t \delta_t^2 + C_2\sum_t \zeta_t^2$$

$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t + \delta_t \quad \text{for all } t$$

$$\sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i, y_k) - E_{\tilde{p}}[f_t] \le B_t + \zeta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

Smoothing techniques (5)

- Inequality with 1-norm Penalty

minimize

$$\sum_i p(x_i) \sum_k p(y_k \mid x_i) \log p(y_k \mid x_i) + C_1 \sum_t \delta_t + C_2 \sum_t \zeta_t$$

$$s.t. \quad E_{\tilde{p}}[f_t] - \sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) \le A_t + \delta_t \quad \text{for all } t$$

$$\sum_i p(x_i) \sum_k p(y_k \mid x_i) f_t(x_i, y_k) - E_{\tilde{p}}[f_t] \le B_t + \zeta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

$$\delta_t \ge 0, \zeta_t \ge 0 \quad \text{for all } t$$

Using MaxEnt as Smoothing

- Add maximum entropy term into the target function of other models, using MaxEnt's preference of uniform distribution

maximize

$$\sum_{k=1}^{L}\sum_{d=1}^{C}\frac{1}{N_{C(k)}}\gamma_{kd}$$

s.t.

$$P_{post}(y=\tilde{y}_k \mid k) \ge P_{post}(y=d \mid k)+\gamma_{kd} \quad \forall k \in 1\ldots L,\ \forall d \in 1\ldots C$$

$$\sum_{c=1}^{C} P(y=c \mid i)=1 \ \text{ and } 0 \le P(y \mid i) \le 1 \quad \forall i \in 1\ldots N$$

minimize

$$\sum_i p(x_i)\sum_k p(y_k \mid x_i)\log p(y_k \mid x_i) + \sum_t \frac{\sigma_t^2}{2}\delta_t^2 - C_1\sum_{i=1}^{L}\sum_{k=1}^{C}\frac{1}{N_{C(i)}}\gamma_{ik}$$

s.t.

$$E_{\tilde{p}}[f_t] - \sum_i p(x_i)\sum_k p(y_k \mid x_i) f_t(x_i,y_k) = \delta_t \quad \text{for all } t$$

$$\sum_k p(y_k \mid x_i) = 1 \quad \text{for all } i$$

$$P_{post}(y=\tilde{y}_i \mid i) \ge P_{post}(y=y_k \mid i)+\gamma_{ik} \quad \forall i \in 1\ldots L,\ \forall k \in 1\ldots C$$

$$\sum_{c=1}^{C} P(y=y_k \mid i)=1 \ \text{ and } 0 \le P(y=y_k \mid i) \le 1 \quad \forall i \in 1..N$$

Bounded error

- Correct distribution $p^C(x_i)$

$$E_p^C[f_t] = \sum_i p^C(x_i) \sum_k p^C(y_k \mid x_i) f_t(x_i, y_k)$$

$$L_p^C(\lambda) = -\sum_t \lambda_t E_p^C[f_t] + \sum_i p(x_i) \log Z_i$$

- Conclusion:

$$\hat{\lambda} = \arg\min_{\lambda} L_{\tilde{p}}^{A,B}(\lambda) \qquad \lambda^* = \arg\min_{\lambda} L_p^C(\lambda)$$

then

$$L_p^C(\hat{\lambda}) \le L_p^C(\lambda^*) + \sum_t \left| \lambda_t^* \right| (A_t + B_t)$$