Regularization and partially labeled data

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

in collaboration with
Adrian Corduneanu
Martin Szummer
Example problem

- A few labeled examples (e.g., documents)
  
  "0"
  This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled
  
  "1"
  We consider the problem of labeled sample to boosting algorithm when only a small set of labeled examples is available. In a problem setting motivated by the task in particular, we consider performance
  
  "1"
  Most computational models of supervised learning rely only on labeled training examples, and ignore the possible role of unlabeled data. This is true both for cognitive science models of learning

- Many unlabeled examples or marginal distribution \( p(x) \)
  
  This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled
  
  \( \ldots \)
  Most computational models of supervised learning rely only on labeled training examples, and ignore the possible role of unlabeled data. This is true both for cognitive science models of learning
Assumptions and approaches

- Parametric family of joint distributions
  - e.g., Naive Bayes and EM

- Constraints within and across examples
  - redundantly sufficient features in co-training
  - explicit constraints across examples

- Neighborhood relations
  - mincut, diffusion processes

- Large margin separation
  - transduction, maximum entropy discrimination
Outline

- A few labeled examples, \((x_1, y_1), \ldots, (x_n, y_n)\), and a large number of unlabeled examples or the marginal \(P(x)\)

- Three scenarios:
  1. Assume a family of joint distributions \(Q(y, x) \in Q\)
  2. Assume a family of conditionals \(Q(y|x) \in Q'\)
  3. no parametric constraints on \(Q(y|x)\)
Parametric case: example

- A Naive Bayes model:

  Binary labels $y \in \{0, 1\}$, $\mathbf{x} = [x_1, \ldots, x_m]^T \in \mathcal{X}$ feature representation of documents

  In this case the parametric family $Q$ is defined by

  $$Q \in Q : \quad Q(y, \mathbf{x}) = Q(y) \prod_{i=1}^{m} Q(x_i | y)$$
Parametric case cont’d

- $y \in \{0, 1\}$, $x \in \mathcal{X}$, parametric family $Q(y, x) \in Q$ such as the Naive Bayes model

- Training error: $D(\hat{P}_{y,x}||Q_{y,x})$, where $\hat{P}(y, x)$ is the empirical distribution based on a few available labeled examples
Parametric case cont’d

• $y \in \{0, 1\}$, $x \in \mathcal{X}$, parametric family $Q(y, x) \in Q$ such as the Naive Bayes model

• Training error: $D(\hat{P}_{y,x} \| Q_{y,x})$, where $\hat{P}(y, x)$ is the empirical distribution based on a few available labeled examples

• Regularization penalty: $D(P_{x} \| Q_{x})$, where $P(x)$ is known or empirical from unlabeled examples
Parametric case cont’d

• \( y \in \{0, 1\}, \ x \in \mathcal{X} \), parametric family \( Q(y, x) \in Q \) such as the Naive Bayes model

• Training error: \( D(\hat{P}_{y,x} \| Q_{y,x}) \), where \( \hat{P}(y, x) \) is the empirical distribution based on a few available labeled examples

• Regularization penalty: \( D(P_x \| Q_x) \), where \( P(x) \) is known or empirical from unlabeled examples

• Regularization problem:

\[
\min_{Q \in Q} \left\{ D(P_x \| Q_x) \right\} \quad \text{subject to} \quad D(\hat{P}_{y,x} \| Q_{y,x}) \leq \eta
\]
Parametric case cont’d

- $y \in \{0, 1\}$, $x \in \mathcal{X}$, parametric family $Q(y, x) \in Q$ such as the Naive Bayes model

- Training error: $D(\hat{P}_{y,x} \| Q_{y,x})$, where $\hat{P}(y, x)$ is the empirical distribution based on a few available labeled examples

- Regularization penalty: $D(P_x \| Q_x)$, where $P(x)$ is known or empirical from unlabeled examples

- Regularization problem:

\[
\min_{Q \in \mathcal{Q}} \left\{ D(P_x \| Q_x) \right\} \quad \text{subject to} \quad D(\hat{P}_{y,x} \| Q_{y,x}) \leq \eta
\]

or

\[
\min_{Q \in \mathcal{Q}} \left\{ (1 - \lambda)D(\hat{P}_{y,x} \| Q_{y,x}) + \lambda D(P_x \| Q_x) \right\}
\]
Estimation, fixed points

- The EM algorithm

E-step: \( \tilde{Q}(x, y) \leftarrow (1 - \lambda) \hat{P}(x, y) + \lambda Q(y|x)P(x) \)

M-step: \( Q(x_i, y) \leftarrow \sum_{x \setminus x_i} \tilde{Q}(x, y) \)
Estimation, fixed points

- The EM algorithm
  
  **E-step:** \( \tilde{Q}(x, y) \leftarrow (1 - \lambda) \hat{P}(x, y) + \lambda Q(y|x)P(x) \)
  
  **M-step:** \( Q(x_i, y) \leftarrow \sum_{x \neq x_i} \tilde{Q}(x, y) \)

- Solution as a fixed point

\[
Q = EM_\lambda(Q)
\]

A number of possible fixed points \((Q, \lambda) \in Q \times [0, 1]\)
Performance and fixed points

- the weighting $\lambda$ can substantially affect the results
The nature of $\text{EM}_\lambda$ fixed points

- loops
- curves starting and ending at $\lambda = 1$
- single curve associated with the labeled only solution $\hat{Q}$
Justification

Theorem: If the Jacobian of \([Q - \text{EM}_\lambda(Q)]\) has full rank, then \(Q = \text{EM}_\lambda(Q)\) defines a smooth 1-dim manifold in \(Q \times [0, 1]\).
Fixed points and instability

- At “critical” values of $\lambda$, the EM algorithm has to find a substantially different fixed point.
Tracing the unique path of fixed points

Numerically stable following of the EM fixed points:

\[
\begin{bmatrix}
I - \nabla_Q \text{EM}_\lambda(Q) - \frac{\partial}{\partial \lambda} \text{EM}_\lambda(Q)
\end{bmatrix}
\begin{bmatrix}
dQ/ds \\
d\lambda/ds
\end{bmatrix} = 0
\]

where \( s \) parameterizes the curve. \( \lambda \) starts to decrease at the “critical” point.
Example results

\[ \hat{Q} \]

\[ Q \]

\[ \lambda = 0 \]

labeled only

\[ \lambda = 1 \]

unlabeled only

<table>
<thead>
<tr>
<th></th>
<th>critical runs</th>
<th>all runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>labeled only</td>
<td>35.8%</td>
<td>35.7%</td>
</tr>
<tr>
<td>EM</td>
<td>28.0%</td>
<td>27.7%</td>
</tr>
<tr>
<td>continuation</td>
<td>20.4%</td>
<td>21.4%</td>
</tr>
</tbody>
</table>

Tommi Jaakkola, MIT AI Lab
Brief discussion

• Regularization in the parametric case leads to the selection of fixed points

• Avoiding critical points with continuation can substantially improve the classification accuracy

• Stronger performance guarantees can be found
  – sensitivity analysis corresponding to the initial labeled solution assuming $P(x)$ is a marginal of some $Q \in Q$
  – extendable to the case of an incorrect model family
Outline cont’d

• A few labeled examples, \((x_1, y_1), \ldots, (x_n, y_n)\), and a large number of unlabeled examples or the marginal \(P(x)\)

• Three scenarios:
  1. Assume a family of joint distributions \(Q(y, x) \in \mathcal{Q}\)
  2. Assume a family of conditionals \(Q(y|x) \in \mathcal{Q}'\)
  3. no parametric constraints on \(Q(y|x)\)
Estimation with known marginal $P(x)$

- What is a sensible general way of relating the conditional $Q(y|x)$ to the marginal distribution $P(x)$?
Estimation with known marginal $P(x)$

- What is a sensible general way of relating the conditional $Q(y|x)$ to the marginal distribution $P(x)$?

- Regularization based on the information induced by the conditional
Estimation with known marginal $P(x)$

- same global information between $x$ and $y$
Estimation with known marginal $P(x)$

- same global information between $x$ and $y$
- locally very different
Covers and local regularization

- A regular cover

- The regularization penalty associated with each region $V \in \mathcal{V}$ in the cover is $p(V) \cdot I_V(x; y)$, where

$$I_V(x; y) = \int_V p_V(x) \sum_y Q(y|x) \log \frac{Q(y|x)}{Q_V(y)} dx$$
Optimization criterion

- We minimize either
  a) maximum local information

\[
\max_{V \in \mathcal{V}} \left\{ p(V) \cdot I_V(x; y) \right\}
\]

b) average local information

\[
\sum_{V \in \mathcal{V}} p(V) \cdot I_V(x; y)
\]

subject to the classification constraints (here log-loss)

\[
\frac{1}{n} \sum_{i=1}^{n} (- \log Q(y|x_i)) \leq c
\]
Solution properties

- The solution $Q(y|x)$ is constant within each atomic region; complexity arises from a large number of atomic regions in high dimensions.
- Atomic regions without samples become irrelevant.
Regularization penalty: continuum limits

- One dimensional case:

\[
\int p(x) \sum_y Q(y|x) \left[ \frac{\partial}{\partial x} \log Q(y|x) \right]^2 dx
\]

- General case:

\[
\int \text{trace}\{F(x)\} P(x) dx,
\]

where \( F(x) \) is the Fisher information matrix corresponding to \( Q(y|x) \) parameterized by \( x \).
Tikhonov style regularization

\[
\frac{1}{n} \sum_{i=1}^{n} (-\log Q(y|x_i)) + \lambda \int \text{trace}\{F(x)\} P(x) dx,
\]

- regularization penalty defines the appropriate interpolation
- solution is continuous, piece-wise differentiable, and the interpolating conditional has a closed-form expression in 1-dim.
Examples

\[ Q(y=1|x) \]
\[ p(x) \]
Regularization with specific conditional families

- We can always solve the regularization problem

\[
\frac{1}{n} \sum_{i=1}^{n} (-\log Q(y|x_i)) + \lambda \int \text{trace}\{F(x)\} P(x) dx,
\]

within any chosen family of conditionals \( Q \in Q' \) such as the logistic regression models.
Performance guarantees

- We derive generalization guarantees for conditionals with bounded information regularizer: \( Q \in Q_{\gamma} \), where

\[
Q_{\gamma} = \left\{ Q : \int \text{trace}\{F(x)\} P(x) dx \leq \gamma \right\}
\]

- not uniform Glivenko-Cantelli
- we rely on the specific properties of the marginal
Learnability

- We base the guarantees on two new quantities
  - $m_p(\alpha)$, the probability mass on $\{x : P(x) \leq \alpha\}$
  - $c_p(\alpha)$, the number of regions induced by $\alpha$

![Learnable vs Not Learnable](image_url)
Preliminary result

- **Theorem:** For any $\epsilon, \delta > 0$

$$\Pr \left\{ \forall Q \in Q_\gamma : \left| \hat{E}\{L(x, y, Q)\} - E\{L(x, y, Q)\} \right| < \epsilon \right\} > 1 - \delta$$

whenever the sample size is greater than

$$O \left( \frac{1}{\epsilon^4} \left( \log \frac{1}{\epsilon} \right) \left[ \log \frac{1}{\delta} + c_p(m_p^{-1}(\epsilon^2)) + \frac{\gamma}{(m_p^{-1}(\epsilon^2))^2} \right] \right)$$

Simplifying assumptions:
- $x \in \mathcal{R}$ (extendable to multiple dimensions)
- squared classification loss $L(y, x, Q) = (1 - Q(y|x))^2$
  (bounded log-loss would work as well)
Summary

- Information regularization provides a general link between the marginal and the conditional.
- The resulting regularization problems can be solved within any chosen family of conditionals or without such restrictions.
- Generalization guarantees exist for conditionals with bounded information regularizer, where the guarantees are expressed in terms of properties of the marginal.