Label Distribution Learning by Optimal Transport

Peng Zhao and Zhi-Hua Zhou
LAMDA Group
Nanjing University, China
2018.02.07
Outline

• Label Distribution Learning (LDL)

• Optimal Transport for LDL

• Theoretical Results

• Experiments

• Conclusions
Outline

• Label Distribution Learning (LDL)
• Optimal Transport for LDL
• Theoretical Results
• Experiments
• Conclusions
Multi-Label Learning

• An example of multi-label learning

<table>
<thead>
<tr>
<th>Sky</th>
<th>Lake</th>
<th>Road</th>
<th>Mountain</th>
<th>Bird</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Label Distribution Learning

- extension of multi-label learning
- care more about the relative importance of different labels in description of an instance
- label is given by a distribution

Text Emotional Analysis

Facial Emotional Prediction
Related Work

Plenty of algorithms are proposed,

• problem transformation
  - PT-SVM, PT-Bayes (Geng 2016)
• algorithm adaptation
  - AA-Bayes and AA-BP (Geng 2016)
• specialized algorithms
  - IIS-LLD (Geng, Yin, and Zhou 2013)

See the survey for more information,

Motivation

However, previous works ignore the label correlations.

For example:
correlations among different emotions
Outline

• Label Distribution Learning (LDL)

• Optimal Transport for LDL

• Theoretical Results

• Experiments

• Conclusions
Optimal Transport

We propose to use **optimal transport** to deal with LDL.

Optimal transport is defined as a distance measurement over two distributions with a pre-defined cost matrix,

**Definition 1.** *(Transport Polytope)* For two probability vectors $r$ and $c$ in the simplex $\Sigma_d$, we write $U(r, c)$ for the transport polytope of $r$ and $c$, namely the polyhedral set of $d \times d$ matrices,

$$U(r, c) := \{ P \in \mathbb{R}_+^{d \times d} | P1_d = r, P^T1_d = c \}. \quad (1)$$
Definition 2. (Optimal Transport) Given a $d \times d$ cost matrix $M$, the total cost of mapping from $r$ to $c$ using a transport matrix (or coupling probability) $P$ can be quantified as $\langle P, M \rangle$. The optimal transport (OT) problem is defined as,

$$d_M(r, c) := \min_{P \in U(r,c)} \langle P, M \rangle.$$  \hfill (2)

Property:

- $d_M$ is a distance whenever $M$ is a metric matrix.

Remarks:

- also known as the Earth Mover’s distance (Rubner, et al.; IJCV’00)
- similar to Wasserstein distance, when $M_{ij} = d^P_K(i, j)$. 

Optimal Transport

**Definition 2.** (Optimal Transport) Given a \(d \times d\) cost matrix \(M\), the total cost of mapping from \(r\) to \(c\) using a transport matrix (or coupling probability) \(P\) can be quantified as \(\langle P, M \rangle\). The optimal transport (OT) problem is defined as,

\[
d_M(r, c) := \min_{P \in U(r,c)} \langle P, M \rangle. \tag{2}
\]

**Solvers:**

- Classical LP problem \(\Rightarrow O(d^3 \log(d))\), too slow
- Sinkhorn entropic regularization (Cuturi et al.; NIPS’13)
- Bregman ADMM (Wang et al.; NIPS’14), Gibbs-OT (Ye et al.; ICML’17)
Optimal Transport

Definition 2. (Optimal Transport) Given a $d \times d$ cost matrix $M$, the total cost of mapping from $r$ to $c$ using a transport matrix (or coupling probability) $P$ can be quantified as $\langle P, M \rangle$. The optimal transport (OT) problem is defined as,

$$d_M(r, c) := \min_{P \in U(r,c)} \langle P, M \rangle. \tag{2}$$

Solvers:

- **Sinkhorn entropic regularization** (Cuturi et al.; NIPS’13)

Definition 3. (Sinkhorn Distance) Given a $d \times d$ cost matrix $M$, and marginal distributions $r, c \in \Sigma_d$. The Sinkhorn distance is defined as,

$$d_M^\lambda(r, c) := \langle P^\lambda, M \rangle, \quad P^\lambda = \arg \min_{P \in U(r,c)} \langle P, M \rangle - \frac{1}{\lambda} H(P), \tag{3}$$

where $H(P) = -\sum_{i=1}^d \sum_{j=1}^d p_{ij} \log p_{ij}$ is the entropy of $P$, and $\lambda > 0$ is entropic regularization coefficient.

can be solved by matrix scaling method, much faster than classical LP.
Optimal Transport for LDL

• Basic idea:

- Cost matrix
- OT distance
- Label correlations
- Loss functions
Optimal Transport for LDL

• Basic idea:

- cost matrix
- OT distance
- label correlations
- loss functions

• Challenges:

- have explicit label correlations
Optimal Transport for LDL

• Basic idea:

  - cost matrix
  - OT distance

  →

  - label correlations
  - loss functions

• Challenges:

  - have explicit label correlations ✓
  - no explicit label correlations ?
Formulation

- **Notations:**
  - Training dataset: \( S = \{ (x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m) \} \)
  - Feature matrix: \( X = [x_1, \cdots, x_m]^T \in \mathbb{R}^{m \times d} \)
  - Label matrix: \( Y = [y_1, \cdots, y_m]^T \), where \( y_i \in \sum_{L} \) and \( L \) is # of labels
Formulation

• Notations:
  - Training dataset: \( S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \)
  - feature matrix: \( X = [x_1, \ldots, x_m]^T \in \mathbb{R}^{m \times d} \)
  - label matrix: \( Y = [y_1, \ldots, y_m]^T \), where \( y_i \in \sum_L \) and \( L \) is # of labels

• Jointly learn transportation and ground metric

\[
\min_{K, h \in \mathcal{H}} \sum_{i=1}^{m} \langle P_i, M \rangle + \frac{C}{2} \| K - K_0 \|^2_F \\
\text{s.t} \quad P_i \in U(h(x_i), y_i) \\
\quad K \in S_+
\]  

(5)
Formulation

• Jointly learn **transportation and ground metric**

\[
\min_{K,h \in \mathcal{H}} \sum_{i=1}^{m} \langle P_i, M \rangle_i + \frac{C}{2} \left\| K - K_0 \right\|_F^2
\]

\[
\text{s.t. } P_i \in U(h(x_i), y_i) \\
K \in S_+
\]

• **kernel biased regularization**
  to learn the **kernel** instead of directly learning the ground metric

\[
M_{ij} = D_{\phi}^2(Y_{:,i}, Y_{:,j}) = \left\| \phi(Y_{:,i}) - \phi(Y_{:,j}) \right\|_2^2.
\]

\[
K_{ij} = K(Y_{:,i}, Y_{:,j}) = \phi(Y_{:,i})^T \phi(Y_{:,j}), \quad M_{ij} = K_{ii} - 2K_{ij} + K_{jj}.
\]
Formulation

• Jointly learn transportation and ground metric

\[
\min_{K, h \in \mathcal{H}} \sum_{i=1}^{m} \langle P_i, M \rangle_i + \frac{C}{2} \|K - K_0\|_F^2
\]

\[
\text{s.t} \quad P_i \in U(h(x_i), y_i) \\
K \in S_+ \\
\]

• kernel biased regularization
  - only needs a projection to positive semi-definite matrix cone
  - avoid the projection to metric space (very costly)
Optimization

- Alternative Optimization
  
  (i) fix $K$ to update $h$: learning the target mapping;
  (ii) fix $h$ to update $K$: learning the ground metric.
Optimization

• Alternative Optimization

(i) fix $K$ to update $h$: learning the target mapping;
(ii) fix $h$ to update $K$: learning the ground metric.

Learning the Target Mapping

\[
\min_{h \in \mathcal{H}} \sum_{i=1}^{m} \langle P_i, M \rangle \\
\text{s.t. } P_i \in \mathcal{U}(h(x_i), y_i).
\]

- Gradient descent, primal-dual method to compute gradient
- Sinkhorn Approximation to speed up
Optimization

• Alternative Optimization

(i) fix $K$ to update $h$: learning the target mapping;
(ii) fix $h$ to update $K$: learning the ground metric.

Learning the Ground Metric

$$\min_{K} \langle P, M \rangle + \frac{C}{2} \| K - K_0 \|_F^2$$

s.t. $K \in S_+$

$$M_{ij} = K_{ii} + K_{jj} - 2K_{ij},$$

where $P = \sum_{i=1}^{m} P_i$.

- has a close form solution
Outline

• Label Distribution Learning (LDL)

• Optimal Transport for LDL

• Theoretical Results

• Experiments

• Conclusions
Theoretical Results

**Theorem 1.** Let $\mathcal{H}$ be the family of hypothesis set, and denote the hypothesis returned by LALOT in Algorithm 1 as $\hat{h}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(\hat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + \frac{2 \log L}{\lambda} + \|M\|_{\infty} \left(16L\mathcal{R}_m(\mathcal{H}) + \frac{2 \log \frac{1}{\delta}}{m}\right),$$

where $R(h)$ is the Bayes risk minimized by the Sinkhorn algorithm, $\mathcal{R}_m(\mathcal{H})$ is the generalization error given the algorithm output $\hat{h}$, and $\lambda$ is the regularization parameter.

**Remarks:**
- second term introduced by Sinkhorn approximation
- standard convergence rate $O(1/\sqrt{m})$ as $\lambda \to \infty$
- risk bound shows a trade-off between accuracy and efficiency.
Theoretical Results

**Theorem 1.** Let $\mathcal{H}$ be the family of hypothesis set, and denote the hypothesis returned by LALOT in Algorithm 1 as $\hat{h}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(\hat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + \frac{2\log L}{\lambda} + \|M\|_{\infty} \left(16L\mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{2\log \frac{1}{\delta}}{m}}\right).$$

where $\mathcal{R}_m(\mathcal{H})$ is Rademacher complexity of hypothesis class $\mathcal{H}$, and $\|M\|_{\infty} = \max_{i,j} M_{ij}$.

**Proof Sketch:**

by uniform generalization bounds, concentration of measure, with Rademacher vector contraction inequality applied on Lipschitz loss.
Outline

• Label Distribution Learning (LDL)

• Optimal Transport for LDL

• Theoretical Results

• Experiments

• Conclusions
Experiments

• Datasets

5 datasets cover fields of biological information classification, natural scene recognition, emotional analysis and so on

• Baselines
  - problem transformation methods:
    PT-Bayes and PT-SVM (Geng2016)
  - algorithm adaptation methods
    AA-KNN and AA-BP (Geng 2016)
  - specialized algorithm maximizing entropy
    IIS-LLD (Geng, Yin, and Zhou 2013).

• Evaluation
  - measure distance of two vectors (↓):
    Chebyshev, Clark, Canberra and KL divergence
  - measure similarity of two vectors (↑):
    Cosine and Intersection

<table>
<thead>
<tr>
<th>Table 1: Statistics of 15 real-world datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
</tbody>
</table>
Table 2: Experimental results on LDL datasets. Each row corresponds to a data set. On each dataset, 10 test runs were conducted and the average performance as well as standard deviation are presented, - indicates numerical limits or errors. Besides, ● (○) indicates that LALOT is significantly better (worse) than the compared method (paired t-tests at 95% significance level).

(a) Performance Measure: Chebyshev ↓

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IIS-LLD</th>
<th>PT-Bayes</th>
<th>PT-SVM</th>
<th>AA-BP</th>
<th>AA-KNN</th>
<th>LALOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>.6749 ± .0065</td>
<td>.6682 ± .0103</td>
<td>.7305 ± .0465</td>
<td>.5863 ± .0370</td>
<td>.5130 ± .0074</td>
<td>.6850 ± .0073</td>
</tr>
<tr>
<td>JAFFE-5</td>
<td>.1396 ± .0046</td>
<td>.4251 ± .0270</td>
<td>.1430 ± .0053</td>
<td>.1490 ± .0073</td>
<td>.1444 ± .0059</td>
<td>.1394 ± .0044</td>
</tr>
<tr>
<td>JAFFE-6</td>
<td>.1207 ± .0060</td>
<td>.3674 ± .0296</td>
<td>.1220 ± .0057</td>
<td>.1265 ± .0085</td>
<td>.1321 ± .0042</td>
<td>.1206 ± .0049</td>
</tr>
<tr>
<td>Emotions</td>
<td>.4429 ± .0138</td>
<td>.6659 ± .0312</td>
<td>.5526 ± .0565</td>
<td>.4087 ± .0140</td>
<td>.3989 ± .0155</td>
<td>.4476 ± .0153</td>
</tr>
<tr>
<td>Yeast-alpha</td>
<td>.0201 ± .0002</td>
<td>.1093 ± .0085</td>
<td>.0139 ± .0003</td>
<td>.0358 ± .0022</td>
<td>.0147 ± .0002</td>
<td>.0136 ± .0002</td>
</tr>
<tr>
<td>Yeast-cdc</td>
<td>.0232 ± .0005</td>
<td>.1211 ± .0080</td>
<td>.0170 ± .0005</td>
<td>.0370 ± .0019</td>
<td>.0176 ± .0003</td>
<td>.0168 ± .0003</td>
</tr>
<tr>
<td>Yeast-cold</td>
<td>.0618 ± .0007</td>
<td>.2060 ± .0155</td>
<td>.0565 ± .0033</td>
<td>.0574 ± .0024</td>
<td>.0553 ± .0009</td>
<td>.0541 ± .0009</td>
</tr>
<tr>
<td>Yeast-diau</td>
<td>.0452 ± .0007</td>
<td>.1793 ± .0126</td>
<td>.0439 ± .0026</td>
<td>.0471 ± .0015</td>
<td>.0395 ± .0006</td>
<td>.0418 ± .0007</td>
</tr>
<tr>
<td>Yeast-dtt</td>
<td>.0491 ± .0010</td>
<td>.2019 ± .0188</td>
<td>.0380 ± .0016</td>
<td>.0443 ± .0022</td>
<td>.0391 ± .0007</td>
<td>.0370 ± .0006</td>
</tr>
<tr>
<td>Yeast-elu</td>
<td>.0239 ± .0004</td>
<td>.1254 ± .0076</td>
<td>.0171 ± .0004</td>
<td>.0363 ± .0015</td>
<td>.0177 ± .0002</td>
<td>.0167 ± .0002</td>
</tr>
<tr>
<td>Yeast-heat</td>
<td>.0526 ± .0006</td>
<td>.1942 ± .0086</td>
<td>.0441 ± .0009</td>
<td>.0520 ± .0013</td>
<td>.0453 ± .0003</td>
<td>.0435 ± .0005</td>
</tr>
<tr>
<td>Yeast-spo</td>
<td>.0653 ± .0010</td>
<td>.1855 ± .0112</td>
<td>.0625 ± .0019</td>
<td>.0664 ± .0035</td>
<td>.0636 ± .0008</td>
<td>.0603 ± .0012</td>
</tr>
<tr>
<td>Yeast-spo5</td>
<td>.0958 ± .0022</td>
<td>.2209 ± .0159</td>
<td>.0923 ± .0025</td>
<td>.0927 ± .0021</td>
<td>.0956 ± .0014</td>
<td>.0908 ± .0015</td>
</tr>
<tr>
<td>Yeast-spoem</td>
<td>.0930 ± .0019</td>
<td>.1909 ± .0144</td>
<td>.0916 ± .0022</td>
<td>.0892 ± .0050</td>
<td>.0919 ± .0022</td>
<td>.0887 ± .0013</td>
</tr>
<tr>
<td>Human Gene</td>
<td>.0535 ± .0007</td>
<td>.1826 ± .0198</td>
<td>.0540 ± .0040</td>
<td>.0602 ± .0009</td>
<td>.0647 ± .0007</td>
<td>.0532 ± .0007</td>
</tr>
</tbody>
</table>

LALOT W/ T/ L: 10/ 5/ 0 14/ 1/ 0 13/ 2/ 0 13/ 0/ 2 12/ 0/ 3 rank first 12/15
Table 2: Experimental results on LDL datasets. Each row corresponds to a data set. On each dataset, 10 test runs were conducted and the average performance as well as standard deviation are presented, - indicates numerical limits or errors. Besides, (○) indicates that LALOT is significantly better (worse) than the compared method (paired t-tests at 95% significance level).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IIS-LLD</th>
<th>PT-Bayes</th>
<th>PT-SVM</th>
<th>AA-BP</th>
<th>AA-KNN</th>
<th>LALOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>.5154 ± .0036</td>
<td>.4882 ± .0064</td>
<td>.3618 ± .0684</td>
<td><strong>.6261 ± .0520</strong></td>
<td>.6220 ± .0100</td>
<td>.4908 ± .0027</td>
</tr>
<tr>
<td>JAFFE-5</td>
<td>.9229 ± .0035</td>
<td>.6331 ± .0303</td>
<td>.9155 ± .0053</td>
<td>.9080 ± .0096</td>
<td>.9122 ± .0072</td>
<td><strong>.9304 ± .0034</strong></td>
</tr>
<tr>
<td>JAFFE-6</td>
<td>.9306 ± .0048</td>
<td>.6561 ± .0229</td>
<td>.9274 ± .0056</td>
<td>.9203 ± .0084</td>
<td>.9124 ± .0035</td>
<td><strong>.9307 ± .0042</strong></td>
</tr>
<tr>
<td>Emotions</td>
<td>.6253 ± .0121</td>
<td>.4352 ± .0361</td>
<td>.3893 ± .1232</td>
<td><strong>.6979 ± .0184</strong></td>
<td>.6892 ± .0194</td>
<td>.5513 ± .0065</td>
</tr>
<tr>
<td>Yeast_cdc</td>
<td>.9872 ± .0004</td>
<td>.8336 ± .0099</td>
<td>.9927 ± .0003</td>
<td>.9598 ± .0038</td>
<td>.9920 ± .0002</td>
<td><strong>.9928 ± .0002</strong></td>
</tr>
<tr>
<td>Yeast_cold</td>
<td>.9883 ± .0004</td>
<td>.8745 ± .0112</td>
<td>.9863 ± .0013</td>
<td>.9857 ± .0012</td>
<td>.9868 ± .0005</td>
<td><strong>.9873 ± .0005</strong></td>
</tr>
<tr>
<td>Yeast_diau</td>
<td>.9822 ± .0004</td>
<td>.8435 ± .0131</td>
<td>.9836 ± .0014</td>
<td>.9800 ± .0017</td>
<td>.9860 ± .0003</td>
<td><strong>.9860 ± .0003</strong></td>
</tr>
<tr>
<td>Yeast_dtt</td>
<td>.9892 ± .0004</td>
<td>.8785 ± .0149</td>
<td>.9935 ± .0004</td>
<td>.9910 ± .0010</td>
<td>.9931 ± .0002</td>
<td><strong>.9938 ± .0002</strong></td>
</tr>
<tr>
<td>Yeast_elu</td>
<td>.9877 ± .0003</td>
<td>.8359 ± .0089</td>
<td>.9932 ± .0003</td>
<td>.9643 ± .0029</td>
<td>.9929 ± .0002</td>
<td><strong>.9935 ± .0002</strong></td>
</tr>
<tr>
<td>Yeast_heat</td>
<td>.9813 ± .0003</td>
<td>.8468 ± .0065</td>
<td>.9868 ± .0007</td>
<td>.9810 ± .0010</td>
<td>.9860 ± .0002</td>
<td><strong>.9872 ± .0003</strong></td>
</tr>
<tr>
<td>Yeast_spo</td>
<td>.9715 ± .0007</td>
<td>.8503 ± .0094</td>
<td>.9728 ± .0020</td>
<td>.9698 ± .0035</td>
<td>.9720 ± .0007</td>
<td><strong>.9746 ± .0008</strong></td>
</tr>
<tr>
<td>Yeast_spo5</td>
<td>.9714 ± .0011</td>
<td>.8832 ± .0120</td>
<td>.9734 ± .0013</td>
<td>.9735 ± .0010</td>
<td>.9709 ± .0009</td>
<td><strong>.9741 ± .0007</strong></td>
</tr>
<tr>
<td>Yeast_spoem</td>
<td>.9762 ± .0008</td>
<td>.9132 ± .0099</td>
<td>.9750 ± .0024</td>
<td>-</td>
<td>.9755 ± .0011</td>
<td><strong>.9771 ± .0007</strong></td>
</tr>
<tr>
<td>Human Gene</td>
<td>.8332 ± .0018</td>
<td>.4597 ± .0403</td>
<td>.8320 ± .0110</td>
<td>.7205 ± .0051</td>
<td>.7694 ± .0021</td>
<td><strong>.8333 ± .0018</strong></td>
</tr>
</tbody>
</table>

**Rank first 12/15**
Experiments

• Label Correlations Exploration

In Figure 1(a), the cost between (desert, sea) ranks the top indicating a very small correlation.

In Figure 1(b), the cost between (amazed, calm) and (happy, sad) rank the top and cost between (quiet, sad) and (quiet, calm) are very small.
Outline

• Label Distribution Learning (LDL)

• Optimal Transport for LDL

• Theoretical Results

• Experiments

• Conclusions
Conclusions

• We propose LALOT to learn the label distribution based on optimal transport theory.

• Incorporate the label correlations into LDL.

• Provide the first data-dependent risk bound analysis for label distribution learning.

• Experiments show the effectiveness of LALOT.
Q & A

Thanks!