A convex relaxation for weakly supervised classifiers

Armand Joulin & Francis Bach
Presented by Eunsu Ryu
Introduction

Classifiers learn the relationship between input variables and certain labels

• **Supervised classifiers**: labeling of the training data must be complete, but detailed annotation is very hard

• **Weakly supervised classifiers**: small number of points are labeled, and use unlabeled points to make better classifier
  • Goal: construct a classifier based observable partial labels
  • Cost function is non-convex. Optimization is done using greedy method or coordinate descent (e.g. EM) and result in local minimum.
Weekly supervised classifiers

- **Multiple-instance learning (MIG)** groups data into bags, and each data point has many potential labels
  - Inference based on discriminative classifiers like SVM (hinge loss)
- **Semi-supervised learning (SSL)** uses both labeled and unlabeled data for training
  - Makes assumptions on smoothness or clustering properties
- **Discriminative clustering** uses existing supervised algorithms to explicitly estimate latent labels
  - Fails in multiclass case; perfect separation is reached by assigning all points to same label
  - Must add linear constraints on the size of clusters

Cost functions for weakly supervised classifiers are in general non-convex
Contributions

New framework for weakly supervised problems
• Solution to the convexity problem of the cost functions using a smart relaxation
• Design of a convex cost function for weakly supervised and unsupervised problems
• Efficient optimization procedure
Problem construction

- Group data into $I$ bags. There are $N$ data points.
- $\mathcal{N}_i$: set of all instances in bag $i$. $|\mathcal{N}_i| = N_i$, $N = \sum_i N_i$.
- Each instance $n \in \mathcal{N}_i$ in bag $i$ has:
  - Feature $x_n \in \mathcal{X}$
  - Label $y_n \in \mathcal{Y}$, $|\mathcal{Y}| = Y$. This label is common to all instances of the same bag.
  - Latent label $z_n \in \mathcal{P}_{y_n} \subset \mathcal{P}$, associated with label $y_n$ of the bag.
- Canonical representation: $z \in \{0,1\}^{N \times P}$, $z_{np} = 1$ only if $n$ has latent label $p$. 
Problem construction

- Group data into $I$ bags. There are $N$ data points.
- $\mathcal{N}_i$: set of all instances in bag $i$. $|\mathcal{N}_i| = N_i, N = \sum_i N_i$.
- Each instance $n \in \mathcal{N}_i$ in bag $i$ has:
  - Feature $x_n \in \mathcal{X}$
  - Label $y_n \in \mathcal{Y}, |\mathcal{Y}| = Y$. This label is common to all instances of the same bag.
  - Latent label $z_n \in \mathcal{P}_{y_n} \subset \mathcal{P}$, associated with label $y_n$ of the bag.
- Canonical representation: $z \in \{0,1\}^{N \times P}$, $z_{np} = 1$ only if $n$ has latent label $p$. 
Problem construction

• **Goal:** given a feature map $\phi: \mathcal{X} \to \mathbb{R}^d$, and the loss function $\ell: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, find $w \in \mathbb{R}^{d \times P}$ and $b \in \mathbb{R}^P$ that minimizes

$$L(z, w, b) = \sum_{n=1}^{N} \pi_n \ell(z_n, w^T \phi(x_n) + b)$$

where $\pi_n \geq 0$ is the importance of $n$ relative to the others. By design, $\sum_n \pi_n = 1$.

• Choose $\ell$ to be **soft-max** so that $\ell(z_n, w^T \phi(x_n) + b)$ is

$$- \sum_{l \in \mathcal{L}} y_{nl} \sum_{p \in \mathcal{P}_l} z_{np} \log \left( \frac{\exp(w_p^T \phi(x_n) + b_p)}{\sum_{k \in \mathcal{P}_l} \exp(w_k^T \phi(x_n) + b_k)} \right)$$

• Cluster size balancing term:

$$H(z) = \sum_{i \in I} h \left( \sum_{n \in \mathcal{N}_i} \pi_n z_n \right) \quad h(v) = - \sum_k v_k \log(v_k)$$
Objective function

**Objective:** find \((z, w, b)\) that gives

\[
\forall n \leq N, \ z_n \in S_{Py_n} \quad \min_{w \in \mathbb{R}^{d \times P}, \ b \in \mathbb{R}^P} f(z, w, b)
\]

where

\[
f(z, w, b) = L(z, w, b) - H(z) + \frac{\lambda}{2P} \|w\|_F^2
\]

and \(S_P = \{t \in \mathbb{R}^P \mid t \geq 0, \ t^T 1_P = 1\}\) (simplex on \(P\)
Convex relaxation

- **Use Fenchel conjugate** \( \log \left( \sum_{p=1}^{P} \exp(t_p) \right) = \max_{v \in S_P} \sum_{p=1}^{P} v_p t_p + h(v) \)** to rewrite the objective function as

\[
\min_{z \in S_P^N} \max_{q \in S_P^N} \sum_{i \in I} \sum_{n \in \mathcal{N}_i} \pi_n h(q_n) - H(z) + g(z, q)
\]

where \( q \in \mathbb{R}^{N \times P} (q 1_K = 1_N) \) and \( g(z, q) \) is

\[
\min_{\substack{w \in \mathbb{R}^{P \times d} \\ b \in \mathbb{R}^P}} \sum_{i \in I} \sum_{n \in \mathcal{N}_i} \pi_n (q_n - z_n)^T (w^T \phi(x_n) + b) + \frac{\lambda}{2P} \|w\|^2_F
\]

- **Minimizing wrt \( b \):** \((q - z)^T \pi = 0 \) (intercept constraint)
- **Minimizing wrt \( w \):**

\[
g(z, q) = -\frac{P}{2\lambda} \text{tr}((q - z)(q - z)^T K)
\]

with \( N \times N \) matrix \( K \) with entries

\[
K_{nm} = \pi_n \phi(x_n)^T \phi(x_m) \pi_m
\]
Convex relaxation

- **Reparametrization in** $q$: write $q = \Omega z$, for $\Omega \in \mathbb{R}^{N \times N}$. Constraints now become $\Omega^T \pi = \pi$, $\Omega 1_N = 1_N$.

- **Tight upper bound on entropy**

  $$\sum_{i \in I} \sum_{n \in N_i} \pi_n h(q_n) \leq -\sum_n \pi_n h(\Omega_n) + H(z) + C_0$$

  leads to the minimization of

  $$\max_{\Omega \in \mathcal{O}} -\frac{P}{2\lambda} \text{tr}(zz^T(I - \Omega)^T K(I - \Omega)) - \sum_n \pi_n h(\Omega_n)$$

  where $\mathcal{O} = \{\Omega \mid \Omega 1_N = 1_N, \Omega^T \pi = \pi, \Omega \geq 0\}$.

- **Constraints on** $Z = zz^T$: $\mathcal{E}_N = \{Z \in \mathbb{R}^{N \times N} \mid \text{diag}(Z) = 1_N, Z \succeq 0\}$

- **Relaxed form**: minimize

  $$\max_{\Omega \in \mathcal{O}} -\frac{P}{2\lambda} \text{tr}(ZR(I - \Omega)^T K(I - \Omega)R^T) - \sum_n \pi_n h(\Omega_n)$$
Optimization

- **Inner loop (maximization):** use proximal method with a reweighted KL divergence. Given a point $\Omega_0$, the update is the maximum of

$$l_D(\Omega) = \text{tr}(\Omega^T \nabla T(\Omega^0)) - \sum_n \pi_n h(\Omega_n) - LD_\pi(\Omega \parallel \Omega^0)$$

where $T(\Omega) = -\frac{1}{2\lambda} \text{tr}((I - \Omega)R^TZR(I - \Omega)^TK)$, and

$$D_\pi(\Omega \parallel \Omega^0) = \sum_i \sum_{n \in N_i} \pi_n \sum_{m=1}^N \Omega_{nm} \log \left( \frac{\Omega_{nm}}{\Omega^0_{nm}} \right)$$
Optimization

- **Outer loop (minimization):** simplify the projection on $\mathcal{E}_{NR}$
  - Replace $g(Z)$ with $g_R(Z) = g(\text{diag}(Z)^{-1/2} Z \text{diag}(Z)^{-1/2})$
  - Minimize $g_R(Z)$ over $\mathcal{E}_{NR}$ using a proximal method with Bregman divergence to guarantee updates that stay in the feasible set
  - Use KL divergence as Bregman divergence
  - Given a point $Z_0$, update using $V \text{diag} \left( \exp \left( \text{diag}(E/t) \right) \right) V^T$, where $V$ and $E$ are eigenvectors and eigenvalues of $-\nabla g_R(Z_0) + t \log(Z_0)$ and $t$ is a step size computed using a line-search with backtracking
# Complexity analysis

<table>
<thead>
<tr>
<th>step</th>
<th>Inner loop update</th>
<th>Inner loop duality gap</th>
<th>Outer loop proximal</th>
<th>Outer loop duality gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>complexity</td>
<td>$O(N^2)$</td>
<td>$O(N^2)$</td>
<td>$O(N^3)$</td>
<td>$O(N)$</td>
</tr>
</tbody>
</table>

*Figure 1. Complexity of the different steps in our algorithm.*
Results: clustering

- **Proof of concept:** $N=500$, want to find 3 (top) and 5 (bottom) clusters, linear kernels

Figure 2. (a) The clustering problem, (b) the given kernel matrix $K = xx^T$, (c) the matrix $Z$ obtained with (Bach & Harchaoui, 2007), (d) the matrix $Z$ obtained with no intercept and (e) our method (best seen in color).
Results: clustering

- **Proof of concept:** three-cluster problem, \( N = 300 \), RBF kernel

---

*Figure 3.* (a) The matrix obtained with our method and (b) its corresponding clusters. (c) Comparison with k-means on noise robustness \((P = 3, N = 300)\).
## Results: MIL

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Musk1</th>
<th>Tiger</th>
<th>Elephant</th>
<th>Fox</th>
<th>Trec1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation k-NN (Wang &amp; Zucker, 2000)</td>
<td><strong>91.3</strong></td>
<td>78.0</td>
<td>80.5</td>
<td><strong>60.0</strong></td>
<td>87.0</td>
</tr>
<tr>
<td>EM-DD (Zhang &amp; Goldman, 2001)</td>
<td>84.8</td>
<td>72.1</td>
<td>78.3</td>
<td>56.1</td>
<td>85.8</td>
</tr>
<tr>
<td>mi-SVM (Andrews et al., 2003)</td>
<td><strong>87.4</strong></td>
<td>78.9</td>
<td>82.0</td>
<td><strong>58.2</strong></td>
<td><strong>93.6</strong></td>
</tr>
<tr>
<td>MI-SVM (Andrews et al., 2003)</td>
<td>77.9</td>
<td><strong>84.0</strong></td>
<td>81.4</td>
<td><strong>59.4</strong></td>
<td><strong>93.9</strong></td>
</tr>
<tr>
<td>PPMM Kernel (Wang et al., 2008)</td>
<td><strong>95.6</strong></td>
<td>80.2</td>
<td>82.4</td>
<td><strong>60.3</strong></td>
<td><strong>93.3</strong></td>
</tr>
<tr>
<td>Random init / Uniform</td>
<td>71.1</td>
<td>69.0</td>
<td>74.5</td>
<td>61.0</td>
<td>81.3</td>
</tr>
<tr>
<td>Tandom init / Weight</td>
<td>76.6</td>
<td>71.0</td>
<td>74.5</td>
<td><strong>59.0</strong></td>
<td>84.4</td>
</tr>
<tr>
<td>No intercept / Uniform</td>
<td>75.0 ± 19.5</td>
<td>67.8 ± 10.4</td>
<td>77.3 ± 9.2</td>
<td>51.3 ± 6.4</td>
<td>87.5 ± 5.2</td>
</tr>
<tr>
<td>No intercept / Weight</td>
<td>77.8 ± 15.7</td>
<td>71.0 ± 10.8</td>
<td>78.9 ± 9.8</td>
<td>52.1 ± 5.0</td>
<td>87.3 ± 5.6</td>
</tr>
<tr>
<td><strong>Ours / Uniform</strong></td>
<td><strong>84.4 ± 14.0</strong></td>
<td>73.0 ± 8.2</td>
<td><strong>86.7 ± 3.5</strong></td>
<td><strong>57.5 ± 5.9</strong></td>
<td><strong>93.0 ± 4.7</strong></td>
</tr>
<tr>
<td><strong>Ours / Weight</strong></td>
<td><strong>87.7 ± 13.3</strong></td>
<td>78.0 ± 5.4</td>
<td><strong>83.9 ± 4.2</strong></td>
<td><strong>62.5 ± 6.4</strong></td>
<td><strong>89.0 ± 6.2</strong></td>
</tr>
</tbody>
</table>

*Figure 4. Accuracy of our approach and of standard methods for MIL. We evaluate our method with and without the intercept and with two types of weights. In bold, the significantly best performances.*
### Results: SSL

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Linear</th>
<th>Nonlinear</th>
<th>Entropy-Reg.</th>
<th>Ours (Linear)</th>
<th>Ours (Nonlinear)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit1</td>
<td>79.41</td>
<td>82.23</td>
<td>75.56</td>
<td><strong>84.57 ± 0.67</strong></td>
<td>75.45 ± 2.88</td>
</tr>
<tr>
<td>BCI</td>
<td>49.96</td>
<td>50.85</td>
<td>52.29</td>
<td><strong>52.22 ± 1.13</strong></td>
<td>50.21 ± 1.09</td>
</tr>
<tr>
<td>g241c</td>
<td>79.05</td>
<td>75.29</td>
<td>52.64</td>
<td><strong>87.15 ± 0.21</strong></td>
<td>87.29 ± 0.42</td>
</tr>
<tr>
<td>g241d</td>
<td><strong>53.65</strong></td>
<td><strong>49.92</strong></td>
<td><strong>54.19</strong></td>
<td>54.44 ± 9.09</td>
<td>53.15 ± 10.09</td>
</tr>
<tr>
<td>USPS</td>
<td>69.34</td>
<td>74.80</td>
<td><strong>79.75</strong></td>
<td>57.08 ± 13.34</td>
<td><strong>79.48 ± 0.50</strong></td>
</tr>
</tbody>
</table>

**Figure 5.** Comparison in accuracy on SSL databases with methods proposed in (Chapelle et al., 2006). In bold, the significantly best performances.
Summary

• Proposed a simple framework for weakly supervised problems
• Convex relaxation of cost functions based on soft-max with intercept