Robust hypothesis test using Wasserstein uncertainty sets

Yao Xie
Georgia Institute of Technology
Joint work with Rui Gao, Liyan Xie, Huan Xu
Classification with unbalanced data

- **Anomaly detection:**
  - self-driving car,
  - network intrusion detection,
  - credit fraud detection,
  - online detection with fewer samples

- **Health care:**
  - many negative samples,
  - not many positive samples
Non-parametric hypothesis test with **unbalanced and limited data**

- empirical distribution may **not** have common support

- no possible to use *likelihood ratio*: **optimal** by well-known Neyman-Pearson.
Hypothesis test using Wasserstein uncertainty sets

• Test two hypothesis
  \[ H_1 : \omega \sim P_1, \quad P_1 \in \mathcal{P}_1 \]
  \[ H_2 : \omega \sim P_2, \quad P_2 \in \mathcal{P}_2 \]

• **Wasserstein uncertainty sets** for distributional robustness

Wasserstein metrics can deal with distributions with different support, better than K-L divergence

• Goal: find optimal detector, minimizes worst-case type-I + type-II errors

\[
\inf_{\phi: \Omega \to \mathbb{R}} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \mathbb{E}_{P_1}[\ell \circ (-\phi)(\omega)] + \mathbb{E}_{P_2}[\ell \circ \phi(\omega)]
\]
Main results

Distributionally robust nearly-optimal detector

- **Theorem**: General distributionally robust detector has nearly-optimal detector has risk bounded by small constant
  \[ \psi(\epsilon) = \epsilon \]

<table>
<thead>
<tr>
<th>( t(t) )</th>
<th>( \psi(\epsilon) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \exp(t) )</td>
<td>( 2\sqrt{\epsilon(1-\epsilon)} )</td>
</tr>
<tr>
<td>( \log(1+\exp(t))/\log 2 )</td>
<td>( H(\epsilon) / \log 2 )</td>
</tr>
<tr>
<td>( (t + 1)^2 )</td>
<td>( 4\epsilon(1-\epsilon) )</td>
</tr>
<tr>
<td>( (t + 1)^2 )</td>
<td>( 2\epsilon )</td>
</tr>
</tbody>
</table>

Computationally efficient

- Tractable convex reformulation
- Complexity independent of dimensionality, scalable to large dataset

\[
O(\ln(n_1) + \ln(n_2))
\]

\[
\max_{p_1, p_2 \in \mathbb{R}_{+}^{n_1+n_2}} \sum_{l=1}^{n_1+n_2} (p_1^l + p_2^l) \psi\left(\frac{p_1^l}{p_1^l + p_2^l}\right)
\]

subject to

\[
\sum_{m=1}^{n_1+n_2} \gamma_k^{lm} \| \omega^l - \omega^m \| \leq \theta_k, \ k = 1, 2, \sum_{m=1}^{n_1+n_2} \gamma_k^{lm} = Q_{k}^{\omega^l}, \ 1 \leq l \leq n_1 + n_2, \ k = 1, 2, \sum_{l=1}^{n_1+n_2} \gamma_k^{lm} = p_k^m, \ 1 \leq m \leq n_1 + n_2, \ k = 1, 2
\]
Statistical interpretations

- Minimizes divergence between two distributions within two Wasserstein balls, centered around empirical distributions, and have common support on $n_1 + n_2$ data points

$$\inf_{\phi: \Omega \rightarrow \mathbb{R}} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \mathbb{E}_{P_1}[\ell \circ (-\phi)(\omega)] + \mathbb{E}_{P_2}[\ell \circ \phi(\omega)]$$

<table>
<thead>
<tr>
<th>Generating function</th>
<th>Auxiliary function</th>
<th>Optimal detector</th>
<th>Detector risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell(t)$</td>
<td>$\psi(p)$</td>
<td>$\phi^*$</td>
<td>$1 - 1/2 \inf_{\phi} \Phi(\phi; P_1, P_2)$</td>
</tr>
<tr>
<td>log(1 + exp(t))/log 2</td>
<td>$-H(p)/\log 2$</td>
<td>$\ln \sqrt{p_1/p_2}$</td>
<td>$H^2(P_1, P_2)$</td>
</tr>
<tr>
<td>$(t + 1)_+^2$</td>
<td>$4p(1-p)$</td>
<td>1 - $\frac{p_1}{p_1 + p_2}$</td>
<td>$JS(P_1, P_2)/\log 2$</td>
</tr>
<tr>
<td>$(t + 1)_+$</td>
<td>$2\min(p, 1-p)$</td>
<td>$\text{sgn}(p_1 - p_2)$</td>
<td>$\chi^2(P_1, P_2)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$TV(P_1, P_2)$</td>
</tr>
</tbody>
</table>

(Juditsky, Nemirovski, 2015)
Human activity detection

**Figure:** Jogging vs. Walking, the average is taken over 100 sequences of data.