Label Distribution Learning Forests

Wei Shen\(^1,2\), Kai Zhao\(^1\), Yilu Guo\(^1\), Alan Yuille\(^2\)

\(^1\)School of Communication and Information Engineering, Shanghai University
\(^2\)Department of Computer Science, Johns Hopkins University

Background

- What is label distribution learning (LDL)?
  - a set of outcomes \( y = \{y_1, y_2, \ldots, y_C\} \).
  - \( y \) is associated with a label distribution \( d = (d_1, d_2, \ldots, d_C) \).
  - \( d_C^x \) expresses the probability of the sample \( x \) having the \( c \)-th label \( y_c \).
  - LDL is suitable for many real-world problems like facial expression recognition and facial age estimation.

- How to address an LDL problem?
  - \( d_C^x = \frac{e^{f(x,y_c)}}{\sum_{j=1}^{C} e^{f(x,y_j)}} \).

- How to address LDL?
  - Decision trees have the potential to model any general world problem.

- Our Intuition
  - We propose LDL learning forests (LDLFs)
    - Our goal is to learn a mapping function \( g: x \rightarrow d \) between an input sample \( x \) and its corresponding label distribution \( d \).

Method

- Problem Formulation
  - We want to learn the mapping function \( g(x) \) by a decision tree based model \( T \).
  - \( f(x, \theta) \) is an index function to bring the \( \phi(n) \)-th output of function \( f(x, \theta) \) in correspondence with split node \( n \).
  - Learning a tree
    - The probability of the sample \( x \) falling into leaf node \( \ell \) is given by \( p(x; \theta, \ell) = \sum_{i=1}^{C} p(x_{i} \theta, \ell) \).
    - Loss function:
      \[ R(q_{\ell}, \theta, S) = -\frac{1}{N} \sum_{i=1}^{N} \delta_{\ell}(x_{i}) \log(q_{\ell}(x_{i}, \theta, S)) - \frac{1}{2} \sum_{i=1}^{N} \sum_{\ell \neq j} \delta_{\ell}(x_{i}) \log(\delta_{\ell}(x_{i}, \theta, S)) \]
    - An alternating optimization strategy to address \((\theta, q_{\ell})\) by minimizing \( R(q_{\ell}, \theta, S) \).
      - Learning Split Nodes
        \[ \frac{\partial R(q_{\ell}, \theta, S)}{\partial \theta_{j}} = \sum_{i=1}^{N} \sum_{\ell \neq j} \frac{d_{\ell}^{x_{i}} \log(q_{\ell}(x_{i}, \theta, S))}{\partial \theta_{j}} \delta_{\ell}(x_{i}) \delta_{\ell}(x_{i}, \theta, S) \]
        \[ \frac{\partial R(q_{\ell}, \theta, S)}{\partial \delta_{\ell}(x_{i}, \theta, S)} = \delta_{\ell}(x_{i}) \log((q_{\ell}(x_{i}, \theta, S)) \log(1 - q_{\ell}(x_{i}, \theta, S)) \delta_{\ell}(x_{i}) \delta_{\ell}(x_{i}, \theta, S)) \]
      - Learning Leaf Nodes
        \[ \min_{q_{\ell}} R(q_{\ell}, \theta, S), \text{ s.t. } \sum_{\ell} q_{\ell} = 1 \]
        - Constrained Convex Optimization Problem!
        - We propose to address this problem by Variational Bounding (VB) [2]
          \[ R(q_{\ell}, \theta, S) \leq -\frac{1}{N} \sum_{i=1}^{N} \sum_{\ell \neq j} \delta_{\ell}(x_{i}) \log(q_{\ell}(x_{i}, \theta, S)) \log(1 - q_{\ell}(x_{i}, \theta, S)) \]
        - Define
          \[ \phi(q, q) = \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell \neq j} \delta_{\ell}(x_{i}) \log(q_{\ell}(x_{i}, \theta, S)) \log(1 - q_{\ell}(x_{i}, \theta, S)) \]
        - \( \phi(q, q) \geq R(q_{\ell}, \theta, S) \) and \( \phi(q, q) = R(q_{\ell}, \theta, S) \) hold the conditions for VB
        - \[ \lambda_{\ell} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell \neq j} \delta_{\ell}(x_{i}) \log(q_{\ell}(x_{i}, \theta, S)) \log(1 - q_{\ell}(x_{i}, \theta, S)) \]
        - Learning a forest \( \mathcal{F} = \{T_1, \ldots, T_k\} \) \( g(x, \theta, \mathcal{F}) = \frac{1}{K} \sum_{k=1}^{K} g(x, \theta, T_k) \)
        \[ R_{\mathcal{F}} = \frac{1}{K} \sum_{k=1}^{K} R_{\mathcal{F}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \partial \phi_{\mathcal{F}}(x_{i}(x_{i}, \theta, \mathcal{F})) \frac{\partial \phi_{\mathcal{F}}(x_{i}(x_{i}, \theta, \mathcal{F}))}{\partial \theta} \]

Parameter Discussion

- Comparison of sLDLFs to Stand-alone LDL Methods
  - We can either use LDLF as a shallow stand-alone model (sLDLFs) or integrate it with any deep networks (dLDLFs)

Experimental Results

- Evaluation of dLDLFs on Facial Age Estimation

Reference