CoConut: Co-Classification with Output Space Regularization

Sameh Khamis  Christoph H. Lampert

Motivation
- In many real-world applications the samples to be classified occur in batches
  - Words in a document
  - Images in a photo collection
  - Stocks in a portfolio
- Can we do better at test-time by exploiting this fact and without re-training our classifiers?

Co-Classification
- We are interested in the task of jointly classifying multiple, otherwise independent, data samples
- Consider the situation of a linear classifier
  - Efficient to train
  - Generalizes well
  - Has a decision hypersurface that might not reflect the class boundaries
- Given a trained classifier and enough test samples, we can modulate its decision surface at test-time so that, for example, it does not cross high density regions

Approach
- Co-Classification with Output Space Regularization
  - Formulated as regularized risk minimization
  - Does not require classifier re-training
  - Can handle test-time additional data modalities
  - Formally:
  \[
  y^* = \arg\min_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(x_i)} w_{ij} \delta_{ij}(y) + \lambda \Omega(y)
  \]
  - The regularizer \( \Omega \) penalizes undesirable label combinations and \( \lambda \) controls its strength
  - In our choice of the regularizer we encode the inductive bias we have about the problem
    - Cluster assumption
    \[
    \Omega_\delta(g) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}(x_i)} w_{ij} \delta_{ij}(g)
    \]
    where \( \delta_{ij}(g) = \|g(x_i) - g(x_j)\| \) indicate whether the label changed between two neighboring samples, where the neighborhood structure can be constructed:
    - With respect to the original features
    - With respect to an additional modality
    - Through a prior or side information
  - Class label distribution
  \[
  \Omega_\pi(g) = \frac{1}{n} \sum_{i=1}^{n} \left| p_i(g) - Q_i \right|
  \]
  where \( p_i(g) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}[g(x_j) = l] \) is the label proportion that should match the target value \( Q_i \)
- The resulting labeling problem is optimized using a combination of discrete optimization and Lagrangian Relaxation

Experiments
- We evaluated CoConut on six different datasets: four image and two network datasets
- We used the features provided by the original authors
- CoConut improves over the baselines, whether through an additional data modality or a prior on class label distribution
- Unsurprisingly, with more data at test-time the improvement is generally more consistent