Semi-supervised Learning and Label Diffusion on a Graph

Paper review from X. Zhu, Z. Ghahramani, J. Lafferty : "Semi-supervised Learning Using Gaussian Fields and Harmonic Functions", 2003

Sara Santos, Martin Josifoski, Antony Doukhan

Outline

Motivation

- Labeled training examples are often costly to obtain

$\longrightarrow$ Leveraging unlabeled data in learning alleviates this issue

- Data often lives in a complex, difficult to capture, manifold

$\longrightarrow$ Imposing a network topology, based on the similarity between features

Theoretical Framework

- l labeled points $(x_1, y_1), \ldots, (x_l, y_l) \in \mathbb{R}^m \times \{0, 1\}$
- u unlabeled points, with known features $x_{l+1}, \ldots, x_{l+u} \in \mathbb{R}^m$
- Underlying graph structure $G = L \cup U$ connecting the n nodes, fully described by a weight matrix W.

Example of a weight matrix:

- RBF: $W_{ij} = \exp\left\{-\sum_{d=1}^{m} \frac{(x_d^{(i)} - x_d^{(j)})^2}{\sigma_d^2}\right\}$, σ_d scale hyper-parameters.

Objective

Given a small number of known labels ($l \ll u$) we want to predict the labels of other nodes.

- Consider $f : G \to R$ and assign labels according to value of f
- Nearby points on the graph are assigned similar labels.
- Loss function: $E(f) = \frac{1}{2} \sum_{i,j} w_{ij}(f(i) - f(j))^2$.

$$\hat{f} = \text{argmin}_{f|_L = y} E(f)$$

Laplacian on the graph

- Cominatorial Laplacian: $\Delta = D - W, \quad d_i = \sum_j w_{ij}$.
- $E(f) = \boldsymbol{f}^T \Delta \boldsymbol{f} \Longrightarrow f$ satisfies $\Delta f = 0$ on U and is unique.
- $f(j) = \frac{1}{d_j} \sum_{i \sim j} w_{ij} f(i) \Longleftrightarrow \boldsymbol{f} = P\boldsymbol{f}, \quad P = D^{-1}W$.

In vector notation, we have :

$$\boldsymbol{f_u} = (D_{uu} - W_{uu})^{-1} W_{ul} \boldsymbol{f_l} = (I - P_{uu})^{-1} P_{ul} \boldsymbol{f_l}$$

Representer Theorem and RKHS

- **Other approach**: consider $\mathcal{H}$ the space of real-valued functions on G.
- $E(f) = \langle f, \Delta f \rangle_{\mathcal{H}} \triangleq \|f\|^2_{\mathcal{H}_K}$
  It can be seen as a regularization term that quantifies the smoothness of f on G.
- Setting K to be the Green operator (inverse Laplacian) on $\mathcal{H}$ and using the Representer theorem :

$$f = \sum_{k \in U} \beta_k K(k, \cdot), \quad \beta_k = \sum_{i \in L} y_i w_{ik}$$

- Can be generalized to other regularizations involving the Laplacian

IMDB Dataset

- The IMDB is a sentiment analysis dataset, comprising of 50K movie reviews
- Each review is associated with a positive or negative label
- For the purposes of this project we assume that, we have access to 5k randomly sampled reviews for training and 2k for testing

Methodology | Network Construction

Starting from the dataset, the network is built as follows:

- Limit the vocabulary to the most frequent 20k words on the web
- Represent the reviews as TF-IDF vectors
- Construct a network where the edge weight between two nodes is equal to

Methodology | Semi-Supervised Classification

- Given the score assigned by

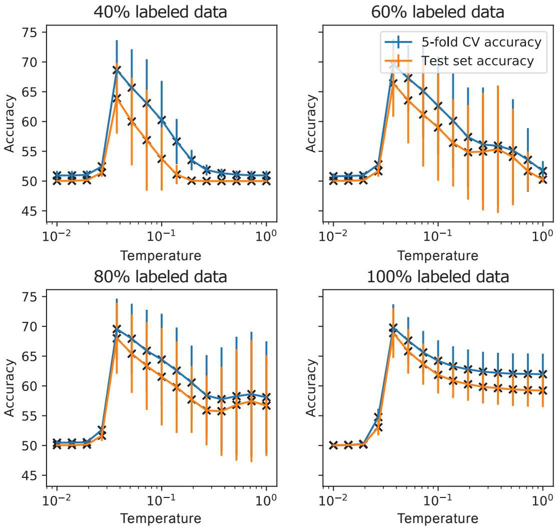
$$\boldsymbol{f_u} = (D_{uu} - W_{uu})^{-1} W_{ul} \boldsymbol{f_l} = (I - P_{uu})^{-1} P_{ul} \boldsymbol{f_l}$$

  we classify reviews with values higher than a threshold $\tau = 0.5$ as positive, while the ones below it as negative
- For unbalanced classification problems the scoring can be adjusted to account for the imbalance

Methodology | Evaluation

- We keep the evaluation dataset of 2k reviews held-out only for testing
- The hyper-parameter tuning is done using 5-fold stratified CV
- We run experiments with different portion of the 5k reviews being labeled
- We repeat each run with 5 random seeds that affect the data that is sampled for training and testing

Results | Semi-Supervised Classification

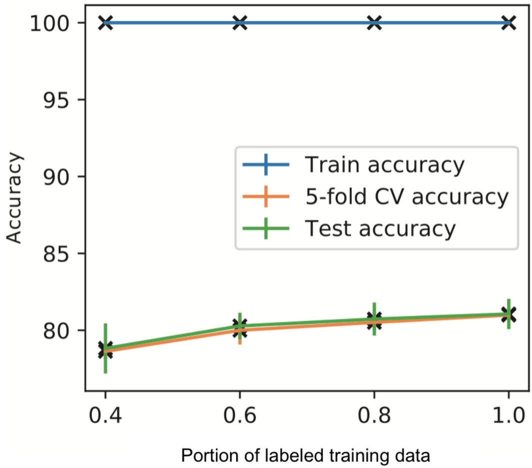


Methodology | External Classifier

To improve the performance we introduce an external classifier:

- Represent each data point by its vector representation
- Train a Random Forest Classifier using the labeled data

Results | Random Forest Classifier

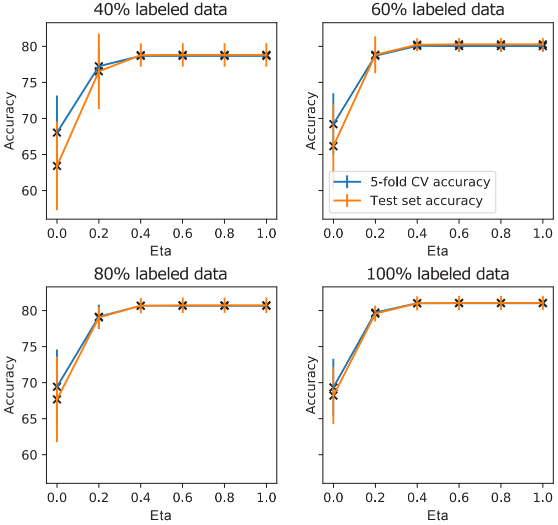


Methodology | Incorporating the External Classifier

- Predict the labels $\hat{\boldsymbol{f}}_{\boldsymbol{u}}$ of the unlabeled samples using the external classifier
- Incorporate the predictions as

$$\boldsymbol{f}_{\boldsymbol{u}} = (I - (1-\eta)P_{uu})^{-1}\left((1-\eta)\,P_{ul}\boldsymbol{f}_{\boldsymbol{l}} + \eta\hat{\boldsymbol{f}}_{\boldsymbol{u}}\right)$$

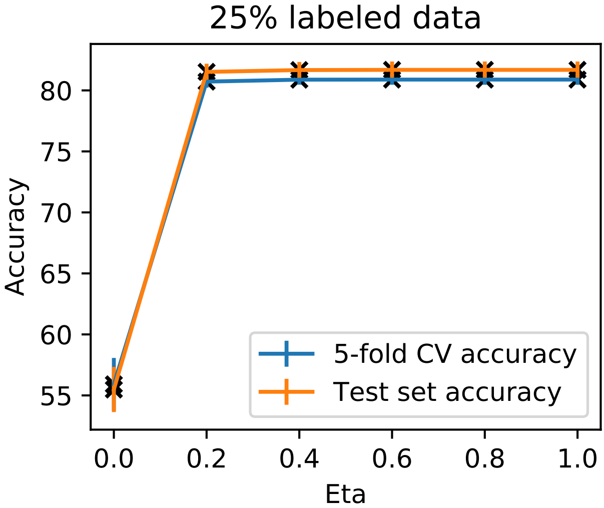
Results | Joint Classification



Discussion

- Marginal improvements over "vanilla" classifier
- Implicit assumption that W captures the structure of the manifold where the data lives
  - Strong assumption in practice – experiments where we use pre-trained word embedding to generate the reviews' vector representations perform even worst than TF-IDF
- Practical implications
  - Memory requirements – addressed by keeping the weight matrix sparse
  - Computation and Optimization – stable multiplication with inverse cast as linear system tackled using iterative solvers (like CG method)

Results | Plethora of unlabeled data (25k)



Bibliography

[1] X. Zhu, Z. Gharamani, J. Laferty (2003). Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. *International Conference on Machine Learning*.

[2] F. Chung (1997). Lectures on Spectral Graph Theory. *University of Pennsylvania*

[3] O. Chapelle, B. Shölkopf, A. Zien (2005). Semi-Supervised Learning. *Massachusetts Institute of Technology*

[4] A.J. Smola, R. Kondor (2003). Kernels and Regularization on Graphs.

[5] C. Hongler (2019). Lattice Models Course, *EPFL*