Σ-Optimality for Active Learning on Gaussian Random Fields

Yifei Ma, Roman Garnett, Jeff Schneider.

Carnegie Mellon University

universitätbonn

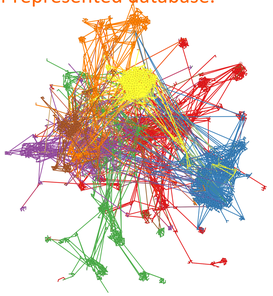
Motivation

We consider the problem of designing a good active learning strategy that, under labeling budget constraints, selects which instances to query for labels that are most helpful for classification/surveying on a graph-represented database.

❑ Examples of graphs

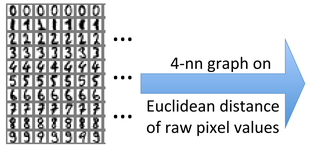
✧ Citation graph

Node: a paper.
Node label (y):
topic of the paper.
Edge (A_{ij}):
presence of a citation.

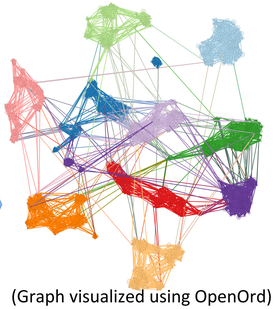


✧ K-nn graph on feature space

Node: a hand-written digit image.
Node label (y): 0-9, actual digit.
Edge (A_{ij}): 4-nearest-neighbors on Euclidean distance.



4-nn graph on
Euclidean distance
of raw pixel values



(Graph visualized using OpenOrd)

❑ Assume labels are connected to graph structure

✧ Specifically, assume Gaussian random field model (GRF)

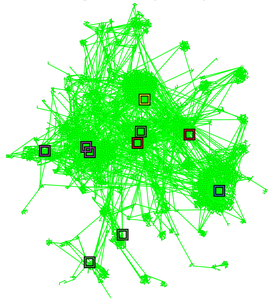
Bayesian generative model:

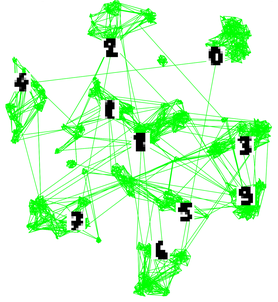
$$P(y) \propto \exp\left\{-\frac{1}{2}\sum_{i,j} A_{ij}(y_i - y_j)^2\right\} = \exp\left\{-\frac{1}{2}y^\top L y\right\}$$

where L is graph Laplacian:

$$L = \mathrm{diag}(\sum_j A_{ij}) - A$$

❑ The goal: query decisions (start from no label)





Approach

For GRFs, the distribution of labels on unlabeled nodes is the conditional Gaussian given known labels. We minimize its Bayesian risk, which is also the predictive variance. Σ-optimality can be viewed as a variant of V-optimality.

❑ Distribution of labels on unlabeled nodes

✧ Let ℓ: labeled, u: unlabeled. (u,ℓ): complementary. Split y, L.

$$P(y_u|y_\ell) \propto \mathcal{N}(y_u; \hat{y}_u, L_u^{-1}), \qquad \hat{y}_u = -L_u^{-1}L_{u\ell}y_\ell$$

❑ Bayesian risk minimization

✧ Greedy L2 risk minimization: V-optimality

$$R_V(\ell) = \mathbb{E}\left[\mathbb{E}\left[\sum_{u_i \in u}(y_{u_i} - \hat{y}_{u_i})^2 \middle| y_\ell\right]\right] = \mathrm{tr}(L_u^{-1})$$

✧ Greedy surveying risk minimization: Σ-optimality

$$R_\Sigma(\ell) = \mathbb{E}\left[\mathbb{E}\left[\left(\frac{y_u \cdot \mathbf{1}}{n} - \frac{\hat{y}_u \cdot \mathbf{1}}{n}\right)^2 \middle| y_\ell\right]\right] = \frac{1}{n^2}\mathbf{1}^\top L_u^{-1}\mathbf{1}$$

Surveying is to predict class proportions.

❑ All active learning strategies considered

(at step k with u^k unlabeled)

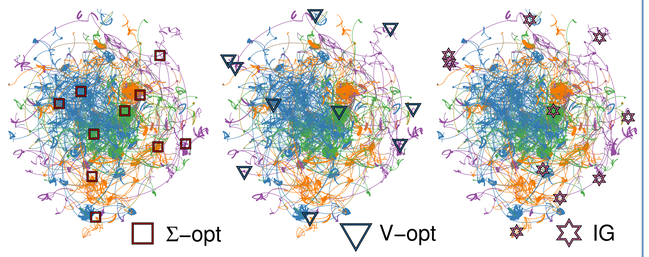
Σ-Optimality	$\min_{v'} \left(\mathbf{1}^\top (L_{u^k\setminus\{v'\}})^{-1}\mathbf{1}\right)$	ours
V-Optimality[2]	$\min_{v'} \mathrm{tr}\left((L_{u^k\setminus\{v'\}})^{-1}\right)$	Bayesian risk min
Info Gain (IG)[3]	$\max_{v'} \left(L_{u^k}^{-1}\right)_{v',v'}$	(same as det-opt)
Mutual (MIG)[3]	$\max_{v'} \left(L_{u^k}^{-1}\right)_{v',v'} \Big/ \left((L_{\ell^k\cup\{v'\}})^{-1}\right)_{v',v'}$	
Uncertainty[4]	$\min_{v'} \lvert\hat{y}_{v'}\rvert$	
E Error (EER)[4]	$\max_{v'} \mathbb{E}_{y_{v'}}\left[\left(\sum_{u_i\in u}\lvert\hat{y}_{u_i}\rvert \middle| y_{v'}\right) \middle| y_{\ell^k}\right]$	

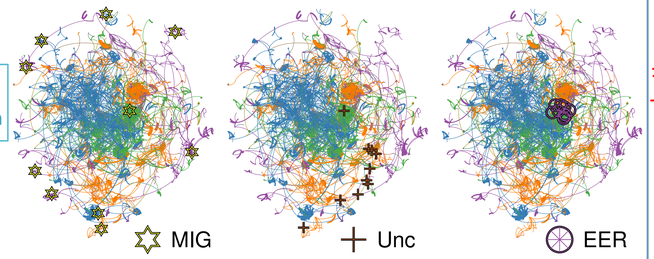
Observations

Let $(L_{u^k}^{-1})_{ij} = \Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ be the conditional covariance. Using rank one update, we have an equivalent selection rule:

$$v^{k+1} = \arg\max_v \begin{cases} \left(\sum_{t\in u}\rho_{vt}\sigma_t\right)^2 & \square \quad \text{(1-norm)} \\ \sum_{t\in u}(\rho_{vt}\sigma_t)^2 & \triangledown \quad \text{(2-norm)} \end{cases}$$

The idea: 1-norm might be less susceptible to outliers.





Theoretical Properties

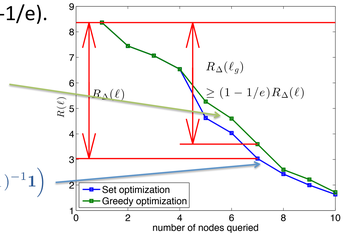
✧ Assume the graph Laplacian L is connected and diagonal dominant, then the risk reduction $R_\Delta(\ell) = R(\ell_0) - R(\ell_0 \cup \ell)$ is positive, monotone, and submodular.

✧ Greedy application of Bayesian risk minimization has bounded approximation ratio (1-1/e).

$$\begin{cases} \min_{v'} \left(\mathbf{1}^\top (L_{u^k\setminus\{v'\}})^{-1}\mathbf{1}\right) \\ \ell^{k+1} = \ell^k \cup \{v'_*\} \end{cases}$$

vs

$$\min_{\ell^{k+1}} R_\Sigma(\ell^{k+1}) = \left(\mathbf{1}^\top (L_{u^{k+1}})^{-1}\mathbf{1}\right)$$



✧ Proof:

- Suppose $L = D - W$ diagonal dominant ⇒

$$L^{-1} = (I - D^{-1}W)^{-1}D^{-1} = [I + D^{-1}W + (D^{-1}W)^2 + \cdots]D^{-1} \geq 0$$

$$\begin{pmatrix} L_{uu} & L_{u\ell} \\ L_{\ell u} & L_{\ell\ell} \end{pmatrix}^{-1} - \begin{pmatrix} L_{uu}^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -L_{uu}^{-1}L_{u\ell} \\ I \end{pmatrix} (L_{\ell\ell} - L_{\ell u}L_{uu}^{-1}L_{u\ell})^{-1} \begin{pmatrix} -L_{\ell u}L_{uu}^{-1}, I \end{pmatrix} \geq 0$$

- Monotonicity of $\ell \subset \ell'$ ⇒
L_u "$\supset$" $L_{u'}$ ⇒ $tr(L_u^{-1}) \geq tr(L_{u'}^{-1})$ ⇒ $R_\Delta(\ell) \leq R_\Delta(\ell')$
- Diminishing marginal reward of a singleton ℓ

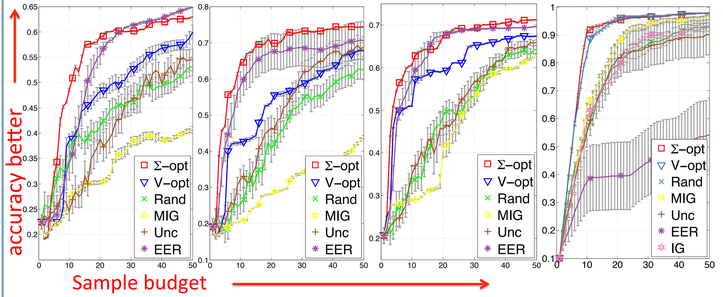
$$\Delta(\ell) = tr(L^{-1}) - tr(L_{uu}^{-1}) = \frac{\left\| \begin{pmatrix} -L_{\ell u}L_{uu}^{-1}, I \end{pmatrix}^T \right\|_2^2}{L_{\ell\ell} - (-L_{\ell u})L_{uu}^{-1}(-L_{u\ell})}$$

It decreases as L shrinks in size nestingly, i.e., as ℓ is queried later.

✧ Bonus: The GRF is a subclass of suppressor-free GPs. Knowing more decouples the unknown. It extends the conditional independence idea. Suppressor-free is desirable.

Empirical Results

❑ Active classification



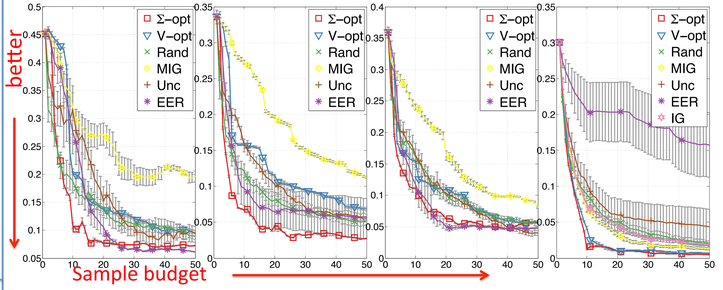
DBLP coauthorship graph
1711 nodes (scholars). Labels: Machine learning / data mining / information retrieval / database
2898 edges: Coauthorship
Used in Ji & Han 2012

Cora citation graph
2485 nodes (papers). Labels: Case based / Genetic algorithms / Neural networks / Probabilistic methods / Reinforcement learning / Rule learning / Theory
5069 edges: Undirected citation
Created by McCallum et al 2000

Citeseer citation graph
2109 nodes (papers) labels: Agents / AI / DB / IR / ML / HCI
3665 edges: Presence of citation
Used in Sen et al 2008

4-nn handwritten digits
1797 images of 8x8 resolution
4-nearest-neighbors graph of Euclidean distances between concatenations of raw pixels.
Labels: actual digit in images
Random subsample 70% digits
From UCI Repository 1998

❑ Active surveying



❑ Active regression (omitted): V-opt worked better

Reference

Methods

1. Zhu, Xiaojin, Lafferty, John, and Ghahramani, Zoubin. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.
2. Ji, Ming and Han, Jiawei. A variance minimization criterion to active learning on graphs. In *AISTAT*, 2012.
3. Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 2008.
4. Settles, Burr. Active learning literature survey. 2010.

Theory

1. Das, Abhimanyu and Kempe, David. Algorithms for subset selection in linear regression. *ACM symposium on Theory of computing*, 2008.
2. Friedland, S and Gaubert, S. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 2011.

Analysis

1. Garnett, Roman, Krishnamurthy, Yamuna, Xiong, Xuehan, Schneider, Jeff, and Mann, Richard. Bayesian optimal active search and surveying. In *ICML*, 2012.
2. Martin, Shawn, Brown, W Michael, Klavans, Richard, and Boyack, Kevin W. Openord: an open-source toolbox for large graph layout. In *IS&T/SPIE Electronic Imaging*, 2011.