Discriminative Clustering by Regularized Information Maximization

Ryan Gomes, Andreas Krause, and Pietro Perona

Caltech

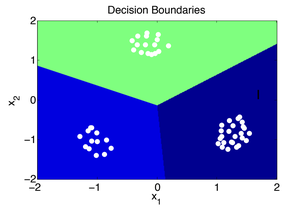
gomes@vision.caltech.edu

The Problem

- Unsupervised discriminative learning

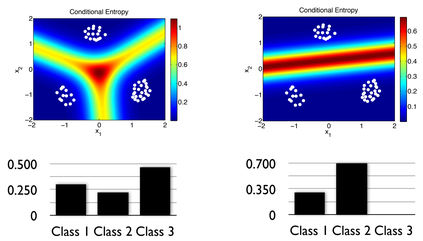
$$p(y|\mathbf{x}, \mathbf{W})$$

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$



$$\arg\max_k p(y = k|\mathbf{x}, \mathbf{W})$$

Class Balance



$$\hat{p}(y; \mathbf{W}) = \frac{1}{N} \sum_i p(y|\mathbf{x}_i, \mathbf{W})$$

$$\arg\max_W H\{\hat{p}(y; \mathbf{W})\} - \frac{1}{N} \sum_i H\{p(y|\mathbf{x}_i, \mathbf{W})\}$$

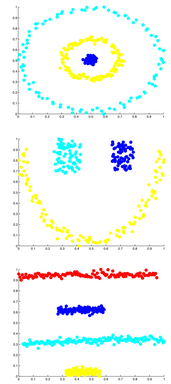
Unsupervised Logistic Regression

$$p(y = k|\mathbf{x}, \mathbf{W}) \propto \exp(\mathbf{w}_k^T \mathbf{x} + b_k)$$

$$R(\mathbf{W}; \lambda) = \lambda \sum_k \mathbf{w_k^T w_k}$$

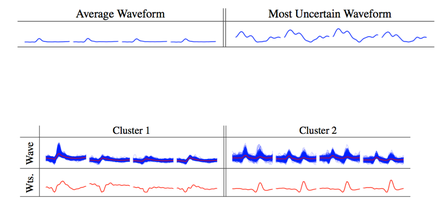
Optimize with L-BFGS quasi-newton algorithm

More Toy Examples



Diffusion Kernel

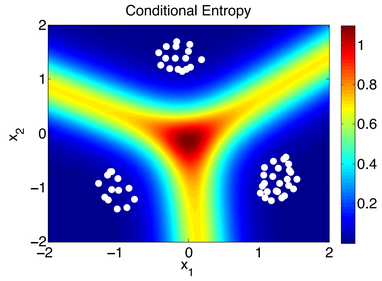
Neural Tetrode Recordings



Linear algorithm, > 300,000 152-dimensional data vectors

Data courtesy of Siapas Lab, Caltech

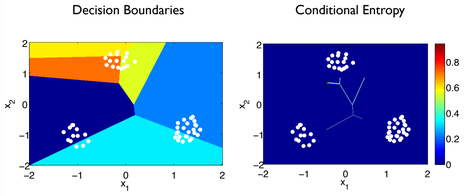
The Cluster Assumption



$$H\{p(y|\mathbf{x}, \mathbf{W})\}$$

Semi-supervised learning by entropy minimization. Grandvalet and Bengio. NIPS 2004.

Information Maximization



$$\operatorname{argmax}_{\mathbf{W}} H\{\hat{p}(y; \mathbf{W})\} - \frac{1}{N} \sum_i H\{p(y|\mathbf{x}_i, \mathbf{W})\}$$

$$= \operatorname{argmax}_{\mathbf{W}} I_{\mathbf{W}}\{\mathbf{x}; y\}$$

Unsupervised classifiers, mutual information and 'phantom targets'. Bridle et al. NIPS 1992.

Unsupervised Kernel Logistic Regression

Stationary condition: $\mathbf{w}_k = \sum_i \alpha_{ki} \mathbf{x}_i$

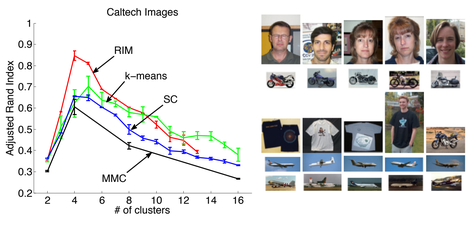
"Kernel Trick": $\mathbf{x}^T \mathbf{x}_i \rightarrow K(\mathbf{x}, \mathbf{x}_i)$

$$p(y = k|\mathbf{x}, \alpha, \mathbf{b}) \propto \exp(\sum_i \alpha_{ki} K(\mathbf{x}, \mathbf{x}_i) + b_k)$$

$$R(\alpha; \lambda) = \lambda \sum_k \sum_{ij} \alpha_{ki} \alpha_{kj} K(\mathbf{x}_i, \mathbf{x}_j)$$

Optimize with L-BFGS quasi-newton algorithm

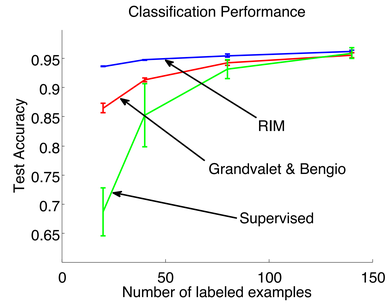
Caltech Images



Spatial Pyramid Kernel

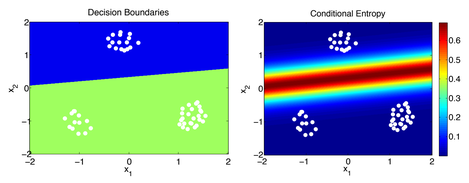
Beyond bags of features. Lazebnik et al. CVPR 2006.

Semi-supervised Classification



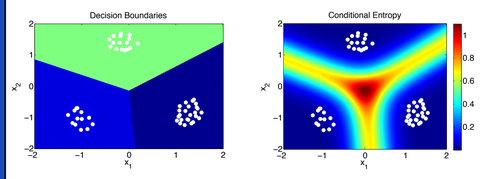
$$S(\mathbf{W}) = \tau F(\mathbf{W}; \mathbf{X}^U, \lambda) + \sum_i \log p(y_i|\mathbf{x}_i^L, \mathbf{W})$$

Degeneracy



$$\arg\min_{\mathbf{W}} \frac{1}{N} \sum_i H\{p(y|\mathbf{x}_i, \mathbf{W})\}$$

Regularized Information Maximization

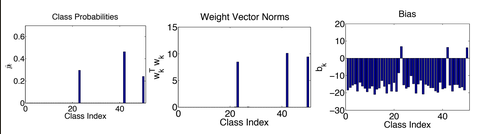


$$\operatorname{argmax}_{\mathbf{W}} I_{\mathbf{W}}\{\mathbf{x}; y\} - R(\mathbf{W}; \lambda)$$

"Model Selection"

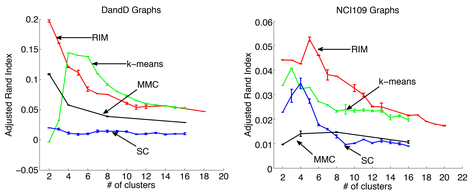
$$p(y = k|\mathbf{x}, \mathbf{W}) \propto \exp(\mathbf{w}_k^T \mathbf{x} + b_k)$$

$$R(\mathbf{W}; \lambda) = \lambda \sum_k \mathbf{w_k^T w_k}$$



If $\hat{p}(y = k; \mathbf{W}) \rightarrow 0$ then $\mathbf{w}_k^T \mathbf{w}_k \rightarrow 0$

Chemical Structure Graphs



Subtree Match Kernel

Fast subtree kernels on graphs. Shervashidze et al. NIPS 2009.

Summary

- Flexible data representation (kernels)
- Rich cluster representation
- Semi-supervised extension
- Model Selection
- Extension to priors on class sizes