A Rate-Distortion One-Class Model and its Applications to Clustering

Koby Crammer Partha Pratim Talukdar Fernando Pereira[1]

University Of Pennsylvania

[1] Currently at Google, Inc.

One Class Prediction

- Problem Statement
  - Predict a coherent **superset** of a small set of positive instances.
- Applications
  - Document Retrieval
  - Information Extraction
  - Gene Expression
- Prefer high precision over high recall.

University of Pennsylvania

Previous Approaches

(Ester et al. 1996) : Density based non-exhaustive clustering algorithm. Unfortunately, density analysis is hard in high dimension.

(Tax & Duin 1999) : Find a small ball that contains as many of the seed examples as possible. Most of the points are considered relevant, a few outliers are dropped.

(Crammer & Chechik 2004) : Identify a small subset of relevant examples, leaving out most less relevant ones.

(Gupta & Ghosh 2006) : Modified version of (Crammer & Chechik 2004).

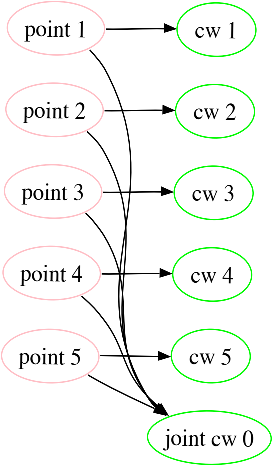
Our Approach: A Rate-Distortion One-Class Model

- Express the one-class problem as lossy coding of each instance into *instance-dependent* codewords (clusters).

- In contrast to previous methods, use more codewords than instances.

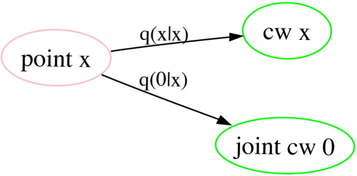
- Regularization via sparse coding: each instance has to be assigned to one of *only two* codewords.

Coding Scheme



- Instances can be coded as themselves, or as a shared codeword (“0”) represented by the vector $\mathbf{w}$.

Notation



$p(x)$ Prior on point x.

$q(0|x)$ Probability of x being encoded by the joint code ("0").

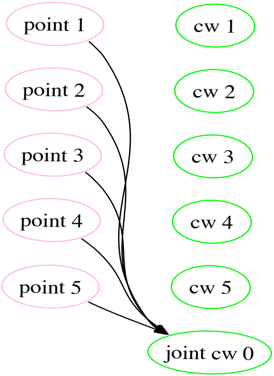
$q(x|x)$ Probability of self-coding point x.

$\mathbf{v}_x$ Vector representation of point x.

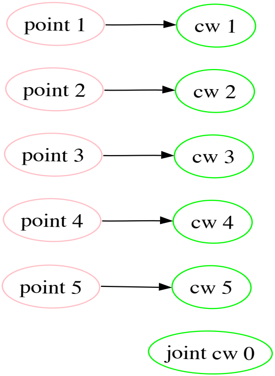
$\mathbf{w}$ Centroid vector of the single class.

$\mathcal{D}(\mathbf{v}_x \| \mathbf{w})$ Cost (distortion) suffered when point x is assigned to the one class whose centroid is $\mathbf{w}$.

Rate & Distorition Tradeoff



All in one
High Compression (Low Rate)
High Distortion



All alone
Low Compression (High Rate)
Low Distorition

Rate-Distortion Optimization

Random variables:

- X: instance to be coded;
- T: code for an instance, either $T = 0$ (shared codeword) or $T = x > 0$ (instance-specific codeword).

Rate: Amount of compression from the source X to the code T, measured by the mutual information $I(T; X)$

Distortion: How well on average the centroid $\mathbf{w}$ serves as a proxy to the instances $\mathbf{v}_x$.

Objective ($\beta > 0$ tradeoff parameter):

$$\min_{\mathbf{w}, \{q(0|x)\}} \quad \textbf{Rate} + \beta \times \textbf{Distortion}$$

Self-Consistent Equations

Solving the Rate-Distortion optimization in the OC setting, we get the following three self-consistent equations, as in IB.

$$q(0) = \sum_x p(x) q(0|x) \tag{1}$$

$$q(0|x) = \min\left\{ q(0) \frac{e^{-\beta \mathcal{D}(\mathbf{v}_x \| \mathbf{w})}}{p(x)}, 1 \right\} \tag{2}$$

$$\mathbf{w} = \sum_x q(x|0) \mathbf{v}_x \tag{3}$$

One Class Rate Distortion Algorithm (OCRD)

We optimize the rate-distortion tradeoff following the Blahut-Arimoto and Information bottleneck (IB) algorithms, alternating between the following two steps:

1. Compute the centroid location $\mathbf{w}$ as the weighted average of instances $\mathbf{v}_x$ with weights proportional to $q(0|x)p(x)$

$$\mathbf{w} = \sum_x q(x|0)\mathbf{v}_x$$

2. Fix $\mathbf{w}$ and optimize for the coding policy $q(0|x), q(0)$

Step 2: Finding a Coding Policy

Let $\mathcal{C} = \{x : q(0|x) = 1\}$ be the set of points assigned to the one class.

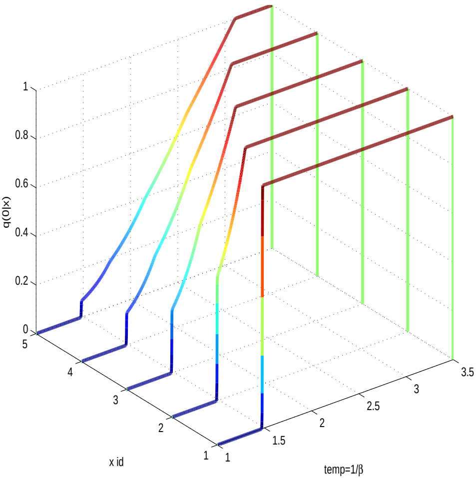
Lemma

$$\textit{Let} \quad s(x) = \beta d_x + \log(p(x))$$

$$\textit{then there is } \theta \textit{ such that } x \in \mathcal{C} \textit{ if and only if } s(x) < \theta$$

The lemma allows us to develop a deterministic algorithm to solve for $q(0|x)$ for $x = 1, \ldots, m$ simultaneously in time complexity $\mathcal{O}(m \log m)$

Phase Transitions in the Optimal Solution



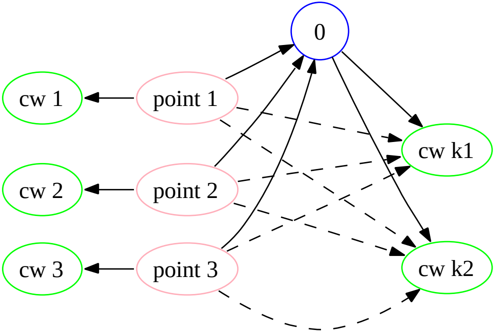
Multiclass Extension

Multiclass Coding Scheme

- We have m points and k centroids. The natural extension doesn't work because $1 - q(x|x)$ does not specify which centroid x should be assigned to.

Multiclass Coding Scheme

- We have m points and k centroids. The natural extension doesn't work because $1 - q(x|x)$ does not specify which centroid x should be assigned to.
- Our Multiclass Coding Scheme:



Multiclass Rate-Distortion Algorithm (MCRD)

MCRD alternates between the following two steps:

1. Use the OCRD algorithm to decide whether we want to self-code a point or not.

Multiclass Rate-Distortion Algorithm (MCRD)

MCRD alternates between the following two steps:

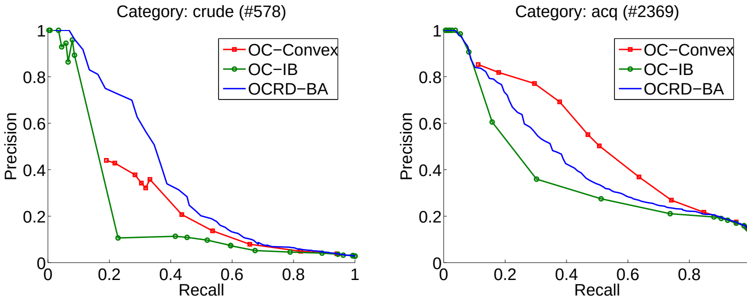
1. Use the OCRD algorithm to decide whether we want to self-code a point or not.

2. Use a hard clustering algorithm (sIB) to clusters the points which we decided not to self-code in the first step. Then iterate.

Experimental Results

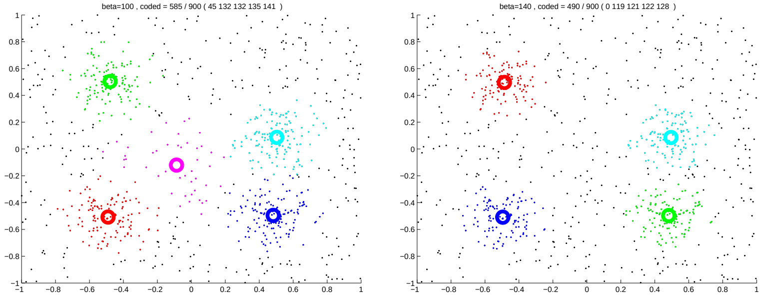
1. One Class Document Classification.
2. Multiclass Clustering of synthetic data.
3. Multiclass Clustering of real-world data.

One Class Document Classification



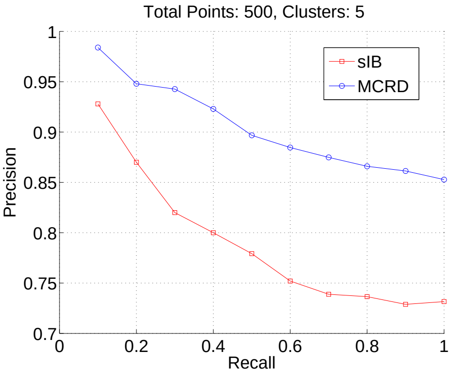
PR plots for two categories of the Reuters-21678 data set using OCRD and two previously proposed methods (OC-IB & OC-Convex). During training, each of the algorithms searched for a meaningful subset of the training data and generated a centroid. The centroid was then used to label the test data, and to compute recall and precision.

Multiclass: Synthetic Data Clustering



Clusterings produced by MCRD on a synthetic data set for two values of β with $k = 5$. There were 900 points, 400 sampled from four Gaussian distributions, 500 sampled from a uniform distribution. Self-coded points are marked by black dots, coded points by colored dots and cluster centroids by bold circles.

Multiclass: Unsupervised Document Clustering



PR plots for sIB and MCRD ($\beta = 1.6$) on the *Multi5_1* dataset (2000 word vocabulary). These plots show that better clustering can be obtained if the algorithm is allowed to selectively leave out data points (through self-coding).

Conclusion

- We have cast the problem of identifying a small coherent subset of data as an optimization problem that trades off between class size (compression) and accuracy (distortion).

Conclusion

- We have cast the problem of identifying a small coherent subset of data as an optimization problem that trades off between class size (compression) and accuracy (distortion).
- We also show that our method allows us to move from one-class to standard clustering, but with background noise left out (the ability to "give up" some points).

Conclusion

- We have cast the problem of identifying a small coherent subset of data as an optimization problem that trades off between class size (compression) and accuracy (distortion).
- We also show that our method allows us to move from one-class to standard clustering, but with background noise left out (the ability to "give up" some points).
- Extend to more general instance spaces and distortions: graphs, manifolds.

UNIVERSITY of PENNSYLVANIA

THANKS!