

Algorithms For Distances Between Distributions

Suresh Venkatasubramanian

AT&T Labs – Research

Dagstuhl, Jul 19, 2005

Outline

- 1 Introduction
- 2 Two Applications
 - The Information Bottleneck
 - Nonparametric Change Detection
- 3 Algorithms
 - Representation And Approximation
 - Solutions

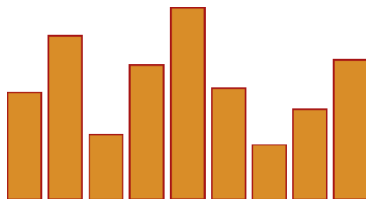
Distributions and Histograms

A family of distributions $p(\mathbf{x}; \theta)$ parametrized by θ

- The distribution could be described by a smooth density function:

$$p(\mathbf{x}; (A, \mu)) = C \exp((\mathbf{x} - \mu)^\top A (\mathbf{x} - \mu))$$

- Or it could be described by a set of buckets and counts:



$$p = p(\mathbf{x}, \{p_1, p_2, \dots, p_k\})$$

Distances between distributions

For two distributions p, q , the distance $d(p, q)$

- satisfies **reflexivity**: $d(p, p) = 0$
also will satisfy (in most cases) **isolation**: $d(p, q) = 0 \Rightarrow p = q$.
- Will occasionally be **symmetric** $d(p, q) = d(q, p)$.
- Will rarely satisfy the **triangle inequality**:

$$d(p, q) + d(q, r) \geq d(p, r)$$

but might satisfy a relaxed version:

$$d(p, q) + d(q, r) \geq c \cdot d(p, r), c < 1$$

A Rogues' Gallery

Kullback-Leibler Distance

$$\text{KL}(p, q) = \sum_i p_i \log \frac{p_i}{q_i}$$

The Jensen-Shannon Distance

$$\text{JS}_{\alpha, \beta}(p, q) = \alpha \text{KL}(p, m) + \beta \text{KL}(q, m), m = \alpha p + \beta q, \alpha + \beta = 1$$

χ^2 -Distance

$$\chi^2(p, q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$$

Δ -Distance

$$\Delta(p, q) = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i}$$

Hellinger-Matusita-Bhattacharya Distance

$$d_H(p, q) = \left[\sum_i (\sqrt{p_i} - \sqrt{q_i})^2 \right]^{\frac{1}{2}}$$

The Rogues' Club

f-divergence For convex $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(1) = 0$,

$$D_f(p, q) = \sum_i p_i f\left(\frac{q_i}{p_i}\right)$$

Bregman divergence For convex $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$

$$D_\phi(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \langle \nabla \phi(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle$$

α -divergence For $|\alpha| < 1$,

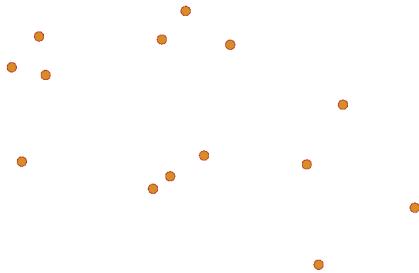
$$D_\alpha(p, q) = \frac{4}{1 - \alpha^2} \left[1 - \int p^{(1-\alpha)/2} q^{(1+\alpha)/2} \right]$$

$f = -\log x$, $\phi = x \log x$ all give $\text{KL}(p, q)$.

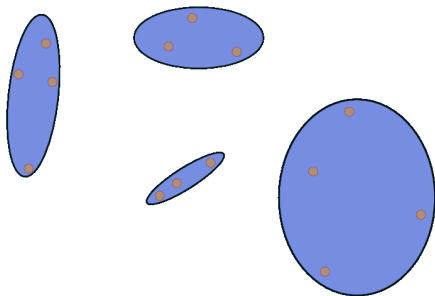
Outline

- 1 Introduction
- 2 Two Applications
 - The Information Bottleneck
 - Nonparametric Change Detection
- 3 Algorithms
 - Representation And Approximation
 - Solutions

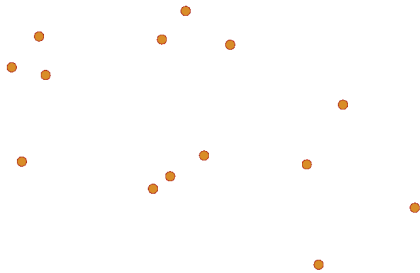
Soft Clustering



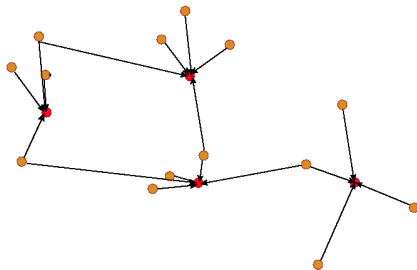
Soft Clustering



Soft Clustering



Soft Clustering



In soft clustering, a point may have multiple cluster memberships, but the “weight” of all memberships must sum to one.

The Information Bottleneck Method [TPB]

X is a collection of “points” defined as vectors over a set Y . In practice, we are given an array of the joint distribution $p(x, y)$.

X	Y
Documents	Words
Hi-D points	Dimensions
Strings	q -grams

A *cluster* is a distribution over Y , described by

Variable	Meaning	Equivalent in “hard” clustering
$p(t x)$	Cluster memberships for $x \in X$	Subsets of X
$p(t)$	Total mass of cluster	Size of cluster
$p(y t)$	Description of cluster center	Coordinates of cluster center

Defining Cluster Quality

Minimize cluster “description length” vs Maximize quality (or minimize error)

Idea: Use **Mutual Information**

$$\begin{aligned} I(X; Y) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \end{aligned}$$

$I(X; T)$ small $\Rightarrow T$ **compresses** X very well.

$I(T; Y)$ large $\Rightarrow T$ is a good representation of Y .

Definition (Information Bottleneck)

Given X, Y , minimize the functional

$$F = I(X; T) - \beta I(T; Y)$$

Solutions

Solution: $p(t|x) \propto p(t)e^{-\beta \text{KL}(p(y|x), p(y|t))}$

Other variables solved consistently.

Hard clustering limit ($\beta \rightarrow \infty$): Probabilistic k -means:

Cluster a collection of distributions into k clusters such that

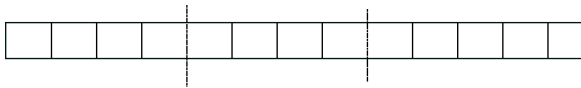
$$\sum_i p(C_i) \sum_{p \in C_i} \text{KL}(p, c_i)$$

is minimized (c_i is center of C_i)

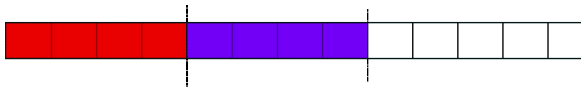
- KL-distance analogous to ℓ_2^2
- Generalization to Bregman distances; lots of interest in Bregman clustering.

Main Question: Compute distances efficiently, or compute “sketches”?

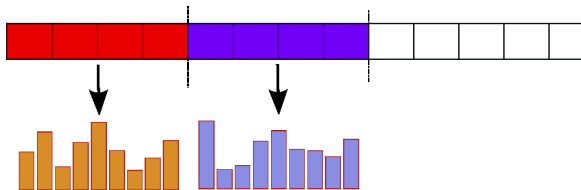
“Nonparametric” Change Detection



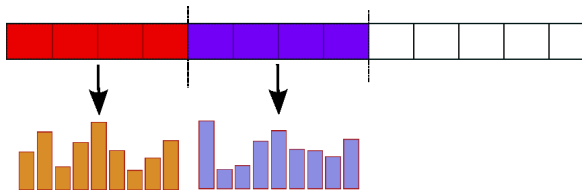
“Nonparametric” Change Detection



“Nonparametric” Change Detection



“Nonparametric” Change Detection



- Checking for signals in neuronal pathways
- Anomaly detection in network monitoring
- “Data cleaning”: Have billing records changed (unexpectedly) from yesterday to today ? [Dasu-Krishnan-V-Yi,Krishnamurthy-Madhyastha-V]

A General Approach

- Could try fitting a model
 - Determine parameters for model from both windows, and compute the difference.
 - How do we know what family to use ?
- More general: Build a distribution out of data, and compare distributions using a general distance.

But which one ?

Using the KL distance

The KL distance generalizes known statistical tests of difference

- If p, q are Gaussians with same variance, then $KL(p, q) = (\mu_p - \mu_q)^2$, which gives us the t -test
- The KL distance behaves (to first order) like a χ^2 -function.

The KL distance relates to optimal hypothesis testing:

Theorem (Neyman-Pearson, Stein's Lemma)

If null hypothesis is "no change occurred", then measuring KL distance in this way is the asymptotically optimal test in terms of minimizing Type I/II error. Moreover, the KL distance is the exponent of the misclassification error.

Main Question: Sample enough data points to estimate distance accurately

Outline

- 1 Introduction
- 2 Two Applications
 - The Information Bottleneck
 - Nonparametric Change Detection
- 3 Algorithms
 - Representation And Approximation
 - Solutions

Representation and Approximation

Problem

Can we compute distributional distances efficiently ?

This can be done easily in time $O(d)$, if the vectors are d -dimensional. But what if d is large ?

Problem

Can we reduce dimensionality while preserving distances approximately ?

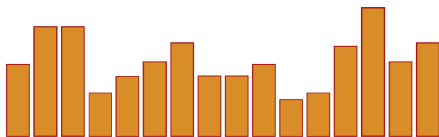
This is a big can of worms ! Lots of work on combining distributions, classification, independence, *etc.*

Histograms: A Simple Example

Extensive study of histograms in database/algorithms community. Histograms are used to maintain approximate profiles of a database so that query optimization can be performed effectively.

Problem (Histogram construction)

Given a function defined on n data points and an error function, find a piece-wise constant function with B pieces that minimizes error to the original function.

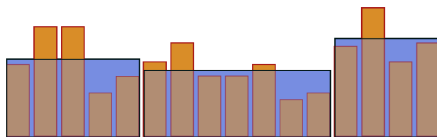


Histograms: A Simple Example

Extensive study of histograms in database/algorithms community. Histograms are used to maintain approximate profiles of a database so that query optimization can be performed effectively.

Problem (Histogram construction)

Given a function defined on n data points and an error function, find a piece-wise constant function with B pieces that minimizes error to the original function.



Maintaining small size histograms over streams

Two step process:

- 1 Determine compact “representative” for an interval
- 2 Determine compact representation of error term when using this representative.

Run an evolving dynamic program over the stream, using the compact representations to store optimal representation of $[x_1, \dots, x_j]$ in $b \leq B$ buckets.

Lemma (GKS01)

If compact representations are provided, the stream algorithm to compute a $(1 + \epsilon)$ -approximate B -bucket histogram runs in space $\frac{B^2}{\epsilon}$ polylogn.

Can we compute histograms for distributional distances?

Representatives

Setup: Given x_1, \dots, x_n , distance function $d(p, q)$, determine “representative” c such that

- ① $\sum_i D(c, x_i)$ is minimized
- ② $\sum_i D(x_i, c)$ is minimized.

Bregman distances:

$$D_\phi(\mathbf{p}, \mathbf{q}) = \phi(\mathbf{p}) - \phi(\mathbf{q}) - \langle \nabla \phi(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle$$

Problem 1 is easy !

Lemma

For *any* Bregman distance, $c = (1/n) \sum_i x_i$ minimizes $\sum_i D(x_i, c)$.

Problem 2 is not so hard either !

Lemma (GV05)

For *any decomposable* Bregman distance, let $c = \phi'^{-1}((1/n) \sum_i \phi'(x_i))$. Then c minimizes $\sum_i D(c, x_i)$.

α -divergences

For $|\alpha| < 1$,

$$D_\alpha(p, q) = \frac{4}{1 - \alpha^2} \left[1 - \int p^{(1-\alpha)/2} q^{(1+\alpha)/2} \right]$$

- Minimizer for Problem 1: $\ell^{1-\alpha}$ norm.
- Minimizer for Problem 1: ℓ^α norm.

α -divergences are “universal”: all “distance-preserving” distribution distances are equivalent to α -divergences with a scale factor.

f-divergences For convex $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(1) = 0$,

$$D_f(p, q) = \sum_i p_i f\left(\frac{q_i}{p_i}\right)$$

f -divergence maps to α -divergence for $\alpha = 3 + 2f'''(1)/f''(1)$.

Duality

In a sense, $D(p, q)$ and $D(q, p)$ are dual to each other.

Theorem (LPP)

For Bregman distance D and representatives p_0, q_0 ,

$$\min_{p \in \mathcal{L}} D(p, q_0) = \min_{q \in \text{dual of } \mathcal{L}} D(p_0, q)$$

- If \mathcal{L} and its dual space are both convex, this yields a primal-dual iterative scheme that converges to optimal solution (think k -means).
- For α -divergences, the spaces that ℓ^α and $\ell^{1-\alpha}$ inhabit are dual (in a differential geometric sense).
- For f -divergences, natural “dual” corresponds to $g(x) = xf(1/x)$.

Fast Estimation, in space and time

Problem

How many samples of p and q do I need to make to estimate the distance between them accurately ?

Problem

How much space do I need to accurately measure (within a $(1 + \epsilon)$ factor) the distance between distributions p and q , when they are presented as a stream of samples from the distribution?

Coming up next...