

Histograms, Information Distances And “Nonparametric” Inference

Suresh Venkatasubramanian
AT&T Labs – Research
<http://www.research.att.com/~suresh>

Mathematics of Visual Analysis, MSRI 2006

Analysis And Exploration

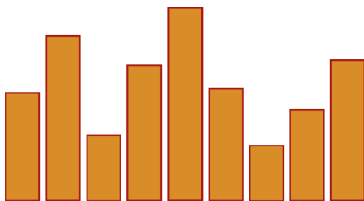
- We need to know what the data looks like in order to analyze it
- We need to analyze the data to see what it looks like.

Assume data follows some known distribution:

- Estimate parameters (not always trivial!)
- Model data accordingly.

Parametric vs "Nonparametric"

Data doesn't usually behave nicely ! Can we use "non-parametric" inference ?



- Represent data via histograms
- Do inference on histograms
- Get first approximation to interesting structure.

It's important that this is a first approximation only: the goal is to generate something that's fast, which allows us to prune out or select regions of interest for further analysis.

The Geometry of Information

Kullback-Leibler Distance

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

The Jensen-Shannon Distance

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

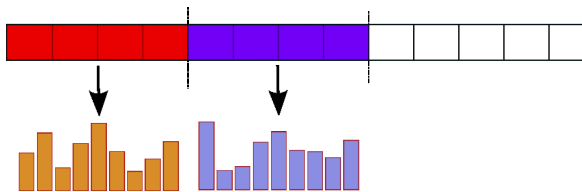
Bregman divergence For convex $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$

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

Using the right space and the right measures of similarity informs the way we manipulate, view and understand data.

Some Examples

Example: Change Detection



- Checking for signals in neuronal pathways
- Anomaly detection in network monitoring
- “Data cleaning”: Have billing records changed (unexpectedly) from yesterday to today ?

Using the KL distance

- If p, q are Gaussians with same variance, then $KL(p, q) = (\mu_p - \mu_q)^2$, which gives us the t -test

Theorem (Neyman-Pearson, Stein's Lemma)

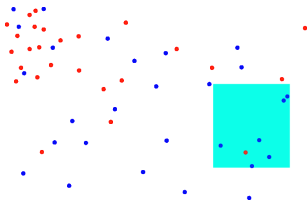
If null hypothesis is "no change occurred", then measuring KL distance in this way is the asymptotically optimal test in terms of minimizing Type I/II error. Moreover, the KL distance is the exponent of the misclassification error.

- Using bootstrapping, we can obtain high confidence bounds on the quality of the alert generated, for different data and different domains. [DKVY06]
- Once the interesting events are generated, further analysis (time-series, Fourier analysis) can be performed on them.

Example: Finding Hotspots

Rather than finding a region in time that's unusually different, find a region in *space*.

Given a set of points, a set of regions \mathcal{R} and a discrepancy function f



Find the region $R \in \mathcal{R}$ that maximizes the discrepancy function $f(R)$

- Disease outbreak patterns and the CDC; anomalies in billing records.

Bump Hunting [APV06,AMPVZ06]

- Problem reduces to finding the KL distance between the histogram inside a region and outside, and maximizing this.
- For different underlying data densities, different information distances are needed
- Good algorithms need to take into account the *geometry* of these distances.
- The basic data type is a histogram
- inferences are based on computing information distances.

Example: Finding Heterogeneity In Data

- Data cleaning is a big problem with large data sets; How do you tell if tables have been merged incorrectly ?
- Can we use a tool for finding “dirty” data to navigate tables and diagnose problems in a database ?

Clustering Histograms

Being generic is key:

- You can always come up with case-by-case methods to analyze databases, but the list keeps growing.
- Methods should work on generic data (strings are a good choice)

How do we cluster histograms ?

- Distance measures are strange (information distances)
- Underlying domain is strange (histograms lie on the simplex)
- Need new concepts: what is a cluster ? what is cluster membership ?

A Connection To Information Theory

An information-theoretic treatment of data clustering yields interesting results:

	“Normal” clustering	Information-theoretic clustering
Points	Vectors in d dimensions	$p(y x), p(x)$
Clusters	Partition of points	$p(t x), t \in T$
Quality Measure	k -means	$I(T; Y)$
Space measure	k	$I(T; X)$
Distance function	ℓ_2, ℓ_2^2	relative entropy
Objective	Fix k , minimize error	Minimize $I(T; X) - \beta I(T; Y)$

Using information-theoretic clustering techniques allows us to cluster different kinds of data generically and accurately [DKOSV06].

Main Thesis

- Generic histogram-based methods allow first-pass quick analysis of data for interesting patterns.
- They create and inhabit different kinds of data spaces (going beyond simple Euclidean spaces) that can be more informative, (and visualized?)
- Beautiful mathematics lurks underneath, and needs to be made effective.