Using Multiple Samples to Learn Mixture Models

Jason Lee, Ran Gilad-Bachrach, and Rich Caruana

Stanford University
Microsoft Research

NIPS 2013

Sat30

7:00pm on Saturday December 7

A Tale of Two Cities

(a) Stanford

(b) Beijing

Figure: Consider sore throat patients at a hospital in Stanford/Beijing. The two hospitals observe different proportions of causes of sore throat.

1. $D_{\text{Stanford}} = .8\mu_{\text{cold}} + .2\mu_{\text{pollution}}$
2. $D_{\text{Beijing}} = .3\mu_{\text{cold}} + .7\mu_{\text{pollution}}$

Multiple Samples from Mixture Models

- **Example 1**: Cluster medical records to identify different subtypes of heart disease. Records are from different hospitals, that represent different communities of patients, so the disease subtypes appear in different proportions.
- **Example 2**: Query categorization. We have queries from different geographical locations, so the query categories appear in different proportions. For example, the Yankees are queried in NYC and the Lakers are queried in California.

Two Algorithms

- **Multi-sample Projection.** Projects the mixtures to a low dimensional space while preserving the separation distance of the means of the distribution.
- **Double Sample Clustering**. Finds the supports of the distributions θ_i, by building a tree of classifiers. The clustering is done by recursively solving classification problems.

Sat30, 7:00pm on Saturday December 7