High-Dimensional Clustering with Sparse Mixtures of Gaussians

Alakshay Krishnamurthy
Carnegie Mellon University

Motivation

- The Curse of Dimensionality - Many clustering algorithms perform poorly in high-dimensions.
- Distances become more uniform, so distance-based clustering algorithms cannot identify clusters.
- For GMMs, covariance matrices overfit samples ⇒ EM does not accurately cluster data.
- $\ell_1$ regularization is a popular strategy for addressing overfitting in classification, regression, and model selection. Can we use it for clustering as well?

For each $\lambda$, train sparse GMM on $X_{train}$.
- Choose $\lambda$ that minimizes $\mathcal{L}(X_{validate}(\theta_{K,\lambda}))$.
- Choose number of clusters $K$ with AIC model selection.
- For each $K$, train sparse GMM and compute AIC score $\mathcal{L}(X_{train}(\theta_{K,\lambda})) - \mathcal{L}(X_{train}(\theta_{K/2}))$.  
- For large $p$, penalty outweighs $\mathcal{L}$ and $K = 2$ is typically chosen.

Experimental Results (Simulations)

- AIC scores for various choices of $K$.
- Correct number of clusters is $K = 5$, $p = 20$.

Training error vs. distance between means ($d$) for $p = 100$. Center: Training error $\times$ vs. $p$ for $d = 10$. Right: Training error $\times$ vs. $p$ for $d = 50$.

Graphical Lasso can be solved with iterative Lasso regressions. 1

A Bayesian Viewpoint

- $C_{i,j,k} \sim \text{Laplace}(\lambda)$ (entrywise)
- $\psi(\cdot)$ $\sim \text{Multinomial}(\pi)$
- $X(\cdot)$ $\sim \mathcal{N}(\mu(\cdot), C_{\cdot,\cdot}^{-1})$
- Using this equivalent formulation we can prove that EM will converge to a local optimum
- Lagrangian:
  $$\sum_{i}^{n} \log \left( \sum_{j}^{K} z_{i,j} \mathcal{N}(X(i) | \mu_{j}, C_{j,j}^{-1}) \right) - \lambda \|C_{\cdot,\cdot}\|_{1}$$
- Use Jensen’s Inequality to only lower bound the first term.
- $z_{i,j}$ learned in the E-step specify a distribution that minimizes KL-divergence to the conditional distribution $P(\psi(\cdot) | X(i), \theta_{K,\lambda})$.
- M-step just maximizes this lower bound w.r.t parameter values.

Practical Issues

- Cross Validation
  - Partition data into subsets $X_{train}$ and $X_{validate}$.
  - For each $\lambda$, train sparse GMM on $X_{train}$.
  - Choose $\lambda$ that minimizes $\mathcal{L}(X_{validate}(\theta_{\lambda,\lambda}))$.

Experimental Results (MNIST)

- Left: Ability to recover number of clusters $K$ with $p = 10, 20, 30, 40, 60, 80$, evaluated by training error. In high dimension AIC penalty is too harsh, resulting in high error.
- Related Work