IE 452/552: AGMDA

Course Project

March 16, 2020

1. Consider the vector dataset $\mathcal{D}$ given in the link https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones# with $|\mathcal{D}| = N$ such that each $v \in \mathcal{D}$ is embedded in a suitable $\mathbb{R}^D$ of minimum possible dimension D. Construct a suitable subspace $\mathcal{S} \subset \mathbb{R}^D$ of dimension at most $\sim O\left(\frac{\ln\left(\frac{N}{\sqrt{0.05}}\right)}{0.01}\right)$ such that at least 95% of the pairwise distances between the points in $\mathcal{D}$ and their corresponding projections to $\mathcal{S}$ do not differ by more than a factor of 0.1. Now produce the best-fit of $\mathcal{D}$ along this $\mathcal{S}$.

2. Construct the top k-SVD subspace $\mathcal{V}_k$ for $\mathcal{D}$ such that the ratio of fit of $\mathcal{D}$ along $\mathcal{V}_k$ to the fit of $\mathcal{D}$ along $\mathcal{V}$ (the full SVD-subspace) does not fall below 0.1. Having obtained this $\mathcal{V}_k$, compare this fit with the fit obtained in Part 1 above. Discuss the results.

3. Generate a dataset $\mathcal{D}'$ which has the same dimensions as the original dataset $\mathcal{D}$ such that each $v \in \mathcal{D}'$ is distributed $\mathcal{N}(0, \Sigma)$. Choose Σ such that it is non-zero in all its elements. Now find the probability of the following events:

   - $P\left[\frac{\sigma_{\max}(\mathcal{D}')}{\sqrt{|D|}} \geq 1.05\sigma_{\max}(\sqrt{\Sigma}) + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right]$
   - $P\left[\frac{\sigma_{\min}(\mathcal{D}')}{\sqrt{|D|}} \geq 0.95\sigma_{\min}(\sqrt{\Sigma}) - \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right]$

   by repeated generation of such a dataset under your same chosen Σ.