CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 5

Due: Thursday, May 16, 2013

This assignment consists of three problems. *Complete only two of the three problems.*

1. *k-Means Business* (50 points)

   (a) The optimal population quantization C^* minimizes $R(C)$. Show that $R(\widehat{C})$ being close to $R(C^*)$ does not imply that $\widehat{C}$ is close to C^*.

   (b) Let $R^{(k)}$ denote the minimal risk among all possible clusterings with k clusters. Show that $R^{(k)}$ is nonincreasing in k.

   (c) Show that, under appropriate conditions on the distribution P from which the random variables X_i are drawn, the optimal k-means risk satisfies $R^{(k)} \to 0$ as $k \to \infty$.

2. *No k-Means Feat* (50 points)

   For LSD Project 1 you extracted a set of 4,000 "tinyimages" and labeled 100 in each category ("cats" or "maps"). You then computed the gist descriptors of these images and computed pairwise distances and the nearest neighbors. These are all of the main ingredients needed for k-means clustering. Building on what you did for LSD 1, now carry out k-means clustering on these data. Try different values of k, display a subset of the images in each cluster, and describe the results. How sensitive are the clusterings to the initial choice of cluster centers?

   Pass in your code in a iPython notebook named `assignment5_prob2.ipynb` in your `submissions/assn5` directory.

3. *Practice Makes Perfect* (50 points)

   Complete the practice exam (which will be posted by Monday, May 13).