STA561/CS571 — Fall 2013 **Homework 9** Due: November 25, 2013

Instructions: Please put all answers in a single PDF with your name and NetID and upload to SAKAI before class on the due date (there is a LaTeX template on the course web site for you to use). Definitely consider working in a group; please include the names of the people in your group and write up your solutions separately. If you look at any references (even wikipedia), cite them. If you happen to track the number of hours you spent on the homework, it would be great if you could put that at the top of your homework to give us an indication of how difficult it was.

Problem 1

Adaptive K-means Revisited:

Statement of the problem: Given a data set $X \in \mathbb{R}^{1000\times 1}$ with an unknown number of cluster components, propose an algorithm which simultaneously identifies the number of cluster components and clusters the data.

In Problem 3 of Homework 4, you thought about how to construct an algorithm based on K-means that would identify clusters in a data with an *a priori* unknown number of cluster components. You saw that the number of clusters grows quickly and is on the order of the number of data points. In order to prevent overfitting and encourage clustering behavior, many of you proposed various penalty functions to remove or otherwise discourage small clusters.

Now that we know about Dirichlet processes, we will put a DP prior on the partition of our data into clusters. Let us put this model in a probabilistic framework: let each cluster represent a univariate Gaussian distribution; this model will then represent an infinite Gaussian mixture model. In this mapping to our Chinese Restaurant Process metaphor, a table at a restaurant is a cluster, and a data point is a customer.

(a) Write out the generative model. What is a simple choice of base distribution?

(b) Write out the equation that you would use in your Gibbs sampler for cluster assignment. Exploit exchangeability (i.e. treat each data point as if it were the last to arrive at the restaurant, then remove it from it's current cluster assignment, and reassign it to a cluster using this equation).

(c) How does this algorithm handle empty clusters?

(d) Explain how this model addresses the problems we encountered in Homework 4, including how this approach differs from specifying a penalty to cluster size of the form: 'remove a cluster if it has fewer than γ points assigned to it'.

Problem 2

Project Proposal Update:

You should now have dived into the analysis of your data and chosen a modeling framework and an validation approach. An important step in the research process is to look at what other ideas researchers have used to approach a related problem. In addition to making progress on your project this week, you will perform a literature search to see what other researchers have worked with similar data or related problems. What relevant approaches, feature sets, or kernels have they developed that might be useful in your own analysis? What modeling approaches and simplifying assumptions worked in this related work, and what didn't work? What can you take from the scientific literature to your project? Write this up as a few paragraphs that will (as with last week's homework) be included in your final project writeup; include citations.