CS 194-10, Fall 2011
Assignment 6

1. *Density estimation in one dimension* (20)
   Assume we observe points $x_1, \ldots, x_N$ on the real line, and wish to estimate the underlying density.

   (a) (3) Let $K_b(d)$ be a kernel density, a function of distance d with width parameter b satisfying $K(d) \geq 0$ and $\int K(x)dx = 1$. The kernel density estimate at x is given by

   $$\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_b(d(x, x_i)) \ .$$

   Show that this estimator is a proper density.

   (b) (5) Let $d_k(x)$ be the distance from x to the kth-nearest neighbor of x. The k-NN density estimator is

   $$\hat{P}(x) = \frac{k}{2Nd_k(x)} \ .$$

   Show that this is *not* a proper density. [Hint: consider $N = 1$, and generalize.]

   (c) (5) The *generalized kernel density estimator* is given by

   $$\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{d_k(x)}(d(x, x_i)) \ ,$$

   i.e., the width parameter of each kernel is given by the k-NN distance of the query point. The idea is that only points that are among the k nearest neighbors and other nearby points have any significant influence on the density at x. For the Gaussian kernel, whose width parameter is σ, explore whether this estimator is a proper density.

   (d) (7) The *variable kernel density estimator* is given by

   $$\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{d_{ik}(x)}(d(x, x_i)) \ ,$$

   where $d_{ik}(x)$ is the distance from x_i to *its* kth-nearest neighbor among the other $N - 1$ points. (We assume $k \leq N - 1$.) Explain why this is a proper density and how it differs from the k-NN density estimator in the way it adapts the effective region size to the amount of data near the query point.

2. *EM* (20) Do Ex. 20.10 in Russell & Norvig, parts (a) and (b). For part (b), just consider $\theta_{F1}^{(1)}$ write out the expression in terms of the parameters, as is done for $\theta^{(1)}$ in the chapter, and calculate the value.

3. *Density estimation for seismic data* (60)
   The file `train.csv` contains a record of seismic events around the world over the last 10 years. The goal is to estimate a probability density for events as a function of location. (What we really care about is the *rate* at each point, but this is proportional to the probability density that the next event anywhere in the world will occur at any given point.) Points are defined by latitude and longitude; we will use the distance function provided for Assignment 1. For now, we will ignore magnitude. The data set is quite large, so you can put effort into making your code more efficient or you can use a subset of the data (at a cost of lower accuracy).

   (a) (50) For each of the four families of density estimators in question 1 above, measure its accuracy using fivefold cross-validation by calculating the log-likelihood of the held-out data. Use a Laplacian kernel and choose parameters (width, k, etc.) by cross-validation.

(b) (10) Investigate the accuracy of a density estimator comprised of a weighted mixture of two components: one is the best density estimator from part (a) and the second is a uniform density over the whole Earth. Determine the optimal weighting by cross-validation.

(c) (10 extra credit) According to geophysical theory, event magnitudes are distributed according to an exponential distribution, such that an event of magnitude m is ten times more likely than an event of magnitude $m+1$. So one might imagine that observation of a large event at a given point should result in a higher density estimate (or rate estimate) than observation of a small event. Investigate whether weighting the events by an appropriate function of their magnitude improves the quality of the density estimators.

Turn in all your code, organized into clearly marked sections according to the parts of the assignment. Supply documentation and explanations where appropriate; describe any methods (cross-validation, multiple trials, etc.) you used to evaluate your methods and get good results.

Submit your files collected together as `a6.tar.gz` using `submit a6` as described **here**