CS 194-10, Fall 2011
Assignment 6

1. *Density estimation using k-NN*
   To show that a density estimator $\hat{P}$ is a proper density function we have to show that (1) $\hat{P}(x) \geq 0$ and that (2) $\int \hat{P}(x)\mathrm{d}x = 1$.

   (a) Consider the estimator $\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_b(d(x, x_i))$. First, note that because $K_b(d(\cdot, x_i))$ are proper density functions we have $K_b(d(x, x_i) \geq 0$ and therefore $\hat{P}(x)$ is an average of non-negative numbers and hence non-negative. Second, note that

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

   (b) We show that $\hat{P}(x) = \frac{k}{2Nd_k(x)}$ is *not* a proper density function. Assume ordinary Euclidean distance, $k=1$, and a single observed point x_1. Then $N=1$ and the density estimate simplifies to $\hat{P}(x) = \frac{1}{2|x-x_1|}$. This estimate is not a proper density function:

   $$\begin{aligned}\int \hat{P}(x)\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{2|x-x_1|}\mathrm{d}x = \int_{-\infty}^{x_1} \frac{1}{2(x_1-x)}\mathrm{d}x + \int_{x_1}^{\infty} \frac{1}{2(x-x_1)}\mathrm{d}x \\ &= \int_{-\infty}^{0} \frac{-1}{2u}\mathrm{d}u + \int_{0}^{\infty} \frac{1}{2u}\mathrm{d}u \qquad \text{by appropriate changes of variable} \\ &= \int_{0}^{\infty} \frac{1}{u}\mathrm{d}u = [\log u]_0^{\infty}\end{aligned}$$

   which diverges (and is certainly not 1). This result goes through identically for $N > 1$ and/or $k > 1$.

   (c) Consider the *generalized kernel density estimator* $\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{d_k(x)}(d(x, x_i))$ with univariate Gaussian kernels of width σ and a single observed point x_1. Again, $N=k=1$ and the estimator simplifies to

   $$\hat{P}(x) = \frac{1}{\sqrt{2\pi(x-x_1)^2}} e^{-\frac{(x-x_1)^2}{2(x-x_1)^2}} = \frac{1}{\sqrt{2\pi(x-x_1)^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi e}|x-x_1|}$$

   whose integral diverges just as for the k-NN estimator.

   (d) The *variable kernel density estimator*, given by $\hat{P}(x) = \frac{1}{N}\sum_{i=1}^{N} K_{d_{ik}}(d(x, x_i))$, is indeed a proper density function. This has the same form as the kernel density estimator in part (a), because the kernel widths are independent of the query point x, and hence this is a proper density estimator.
   Consider a query point x a long way from a small cluster of data points. The kernel width for the $k-NN$ density estimator would be large, but for the variable kernel density estimator they would be small, because the data points are close to each other. Thus, the variable kernel density estimator is not strictly sensitive to the amount of data near the query point.

2. *Ex. 20.10 in Russell & Norvig*

   (a) Consider the ideal case in which the bags were infinitely large so there is no statistical fluctuation in the sample. With two attributes (say, *Flavor* and *Wrapper*), we have five unknowns: θ gives the relative sizes of the bags, θ_{F1} and θ_{F2} give the proportion of cherry candies in each bag, and θ_{W1} and θ_{W2} give the proportion of red wrappers in each bag. In the data, we observe just the flavor and wrapper for each candy; there are four combinations, so three independent numbers

can be obtained. We can write equations for these numbers in terms of the unknown parameters. For example, the number of cherry candies with red wrappers is

$$r_c = N\theta\theta_{F1}\theta_{W1} + N(1-\theta)\theta_{F2}\theta_{W2} \ .$$

With three such equations, we cannot recover five unknowns. This failure is a consequence not of the specific details of the EM algorithm but of the nature of the optimization problem. With three attributes, on the other hand, there are eight combinations and seven numbers can be obtained, enough to recover the seven parameters.

(b) We repeat in detail the computation described on AIMA, page 823: In the fully observable case we would estimate $\theta_{F_1}^{(1)}$ directly from the observed counts of cherry and lime candies from bag 1 and their flavors. Because the bag is a hidden variable we calculate the expected counts instead. The expected count $\hat{N}(Bag = 1)$ is the sum, over all candies, of the probability that the candy came from bag 1 (for brevity, we substitute the variables' names with their initials, for example we substitute $flavor$ by f):

$$\hat{N}(B=1) = \sum_{j=1}^{N} P(B_j = 1 | f_j, w_j, h_j) =$$

$$\sum_{j=1}^{N} \frac{P(f_j|B_j=1)P(w_j|B_j=1)P(h_j|B_j=1)P(B_j=1)}{\sum_i P(f_j|B_j=i)P(w_j|B_j=i)P(h_j|B_j=i)P(B_j=i)} =$$

$$\sum_f \sum_w \sum_h N(f,w,h) \frac{P(f|B=1)P(w|B=1)P(h|B=1)P(B=1)}{\sum_i P(f|B_j=i)P(w|B=i)P(h|B=i)P(B=i)} =$$

$$N(f,w,h) \frac{\theta_{F_1}^{(0)}\theta_{W_1}^{(0)}\theta_{H_1}^{(0)}\theta^{(0)}}{\theta_{F_1}^{(0)}\theta_{W_1}^{(0)}\theta_{H_1}^{(0)}\theta^{(0)} + \theta_{F_2}^{(0)}\theta_{W_2}^{(0)}\theta_{H_2}^{(0)}(1-\theta^{(0)})} +$$

$$N(f,w,\neg h) \frac{\theta_{F_1}^{(0)}\theta_{W_1}^{(0)}(1-\theta_{H_1}^{(0)})\theta^{(0)}}{\theta_{F_1}^{(0)}\theta_{W_1}^{(0)}(1-\theta_{H_1}^{(0)})\theta^{(0)} + \theta_{F_2}^{(0)}\theta_{W_2}^{(0)}(1-\theta_{H_2}^{(0)})(1-\theta^{(0)})} + \ldots =$$

$$273 \cdot \frac{0.6 \cdot 0.6 \cdot 0.6 \cdot 0.6}{0.6 \cdot 0.6 \cdot 0.6 \cdot 0.6 + 0.4 \cdot 0.4 \cdot 0.4 \cdot (1-0.6)} +$$

$$93 \cdot \frac{0.6 \cdot 0.6 \cdot (1-0.6) \cdot 0.6}{0.6 \cdot 0.6 \cdot (1-0.6) \cdot 0.6 + 0.4 \cdot 0.4 \cdot (1-0.4) \cdot (1-0.6)} + \ldots =$$

$$273 \cdot 0.8351 + (93+79+104) \cdot 0.6922 + (100+90+94) \cdot 0.5 + 167 \cdot 0.3077 = 612.4$$

Similarly, the *expected* count of cherry candies from bag 1 is given by:

$$\hat{N}(B=1, f=cherry) = \sum_{j:f_j=cherry} P(B_j = 1 | f_j = cherry, w_j, h_j) =$$

$$\sum_{j:f_j=cherry} \frac{P(f_j=cherry|B_j=1)P(w_j|B_j=1)P(h_j|B_j=1)P(B_j=1)}{\sum_i P(f_j=cherry|B_j=i)P(w_j|B_j=i)P(h_j|B_j=i)P(B_j=i)} =$$

$$\sum_w \sum_h N(f=cherry,w,h) \frac{P(f=cherry|B=1)P(w|B=1)P(h|B=1)P(B=1)}{\sum_i P(f=cherry|B_j=i)P(w|B=i)P(h|B=i)P(B=i)} =$$

$$273 \cdot 0.8351 + (93+104) \cdot 0.6922 + 90 \cdot 0.5 = 409.3$$

Our estimate for $\theta_{F_1}^{(1)}$ is therefore $409.3/612.4 \approx 0.6684$.