Homework 1, due Tuesday, Sept. 26, 2006
COMP 766-001 – Machine Learning for Bioinformatics

[1] (10 pts) The exponential density has a single parameter λ, and follows the formula $P_\lambda(x) = \lambda e^{-\lambda x}$ for $\lambda > 0$ and $x > 0$. Suppose we are given a set of observations $X = (x_1, x_2, \ldots, x_N)$ to which we want to an exponential density. Following the recipe described in class, derive the maximum likelihood estimate for λ. Show all steps of your derivation.

[2] (10 pts) Find the maximum likelihood estimate for λ for the 522 observations in "Homework_01_data.txt". Plot a histogram of the data, side-by-side with or overlaid on the maximum-likelihood exponential density.

[3] (10 pts) Fit a nonparametric density estimate to the same data. In class we discussed the Parzen windows technique. Recall that this can be written as

$$P(x) = \sum_{i=1}^{N} \frac{1}{N} f_i(x)$$

where

$$f_i(x) = \begin{cases} \frac{1}{h} & \text{if } |x - x_i| \le \frac{h}{2} \\ 0 & \text{otherwise} \end{cases} ,$$

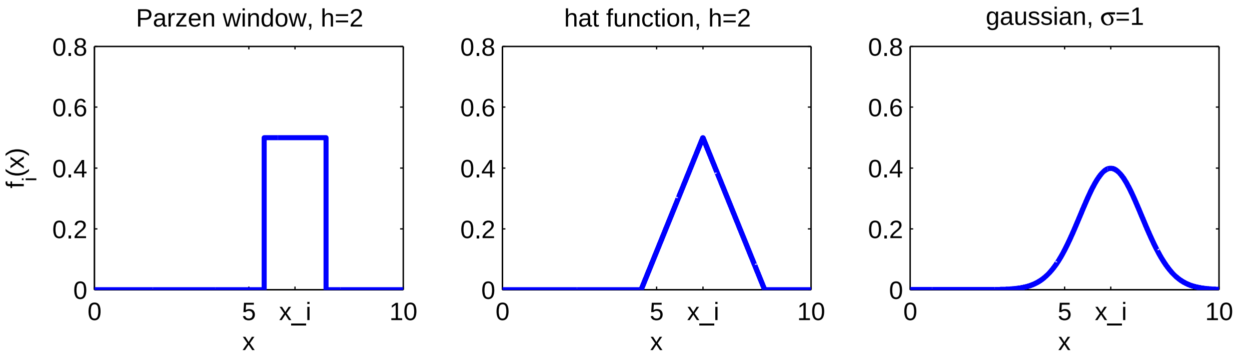
where h is a "width" parameter that can be chosen in various ways. (We'll discuss this more in class tuesday.) For this exercise, show fits based on a different choice for the $f_i(x)$ functions. You may use the hat function (see also picture below),

$$f_i(x) = \begin{cases} \frac{h - x_i + x}{h^2} & \text{if } x_i - h \le x \le x_i \\ \frac{h - x + x_i}{h^2} & \text{if } x_i \le x \le x_i + h \\ 0 & \text{otherwise} \end{cases} .$$

Or, you may use the Gaussian function,

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - x)^2}{2\sigma^2}} .$$

Or, you may use any other function you like, as long as you say what it is.



Like the Parzen windows, the hat function and the Gaussian function have a width parameter, h or σ, that influences the "smoothness" of the resulting density estimate. The hat function is continuous, unlike the Parzen windows, but is not "smooth" in the sense of differentiable. The Gaussian is both continuous and differentiable. After choosing the width parameter (by eye, or by cross-validation), plot the resulting nonparametric density estimate side-by-side with or overlaid on a histogram of the data.

[4] (0 pts) Do you think the data comes from an exponential distribution?