STAT 391
Homework 5
Out Tuesday May 5, 2020
Due Tuesday May 12, 2020

©Marina Meilă
mmp@stat.washington.edu

Problem 1 - Model families

You have a set $\mathcal{D}$ of n samples from a distribution f^{true}, and you use them to to estimate f in a model class $\mathcal{F}$. Answer the following questions without giving proofs.

a. The table below lists several possible f^{true} distributions, and several model classes. Fill in each entry of the table with the model class of the estimated density f. E.g., ifyou believe that if f^{true} is normal and the model class is $\mathcal{F}_1$ the class of normal distributions, enter $\mathcal{F}_1$ in the top-left square of the table. Enter "other" whenever neither $\mathcal{F}_{1,2,3,4}$ is the answer.

f^{true}	$\text{Normal}(0, 5)$	$\text{Exponential}(\lambda_0 = 5)$	$\text{Uniform}_{[-1,1]}$
$\mathcal{F}_1 = \{\text{Normal}(\mu, \sigma^2)\}$	c		
$\mathcal{F}_2 = \{\text{Exponential}(\lambda)\}$			
$\mathcal{F}_3 = \{\text{Uniform}_{[\alpha,\beta]}\}$			b
$\mathcal{F}_4 = \text{KDE}_h$	d		

The kernel in $\mathcal{F}_4$ is the Gaussian kernel. Questions **b,c,d** refer to the table cells marked with these letters; $\alpha, \mu, \ldots$ refer to the ML estimates of the parameters from the respective families. In other words, you are asked to compare the estimated parameters with the true parameters.

b. If you answered $\mathcal{F}_3$, then circle or mark one of the following statements, otherwise skip this question.

$\alpha < -1$ $\quad$ $\alpha = -1$ $\quad$ $\alpha > -1$ $\quad$ I didn't answer $\mathcal{F}_3$.

c. If you answered $\mathcal{F}_1$, then circle or mark one of the following statements, otherwise skip this question.

$\mu < 0$ $\quad$ $\mu = 0$ $\quad$ $\mu > 0$ $\quad$ I didn't answer $\mathcal{F}_1$.

d. If you answered $\mathcal{F}_1$, then circle or mark one of the following statements, otherwise skip this question.

$\mu < 0$ $\quad$ $\mu = 0$ $\quad$ $\mu > 0$ $\quad$ I didn't answer $\mathcal{F}_1$.

Problem 2 - Estimating h by cross-validation

For this problem, submit your code.

In this problem you will compute and plot a kernel density estimate of the corresponding densities f and g given below (you have calculated these densities in homework 4).

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$g(x) = \begin{cases} 4x & 0 \leq x \leq 0.5 \\ 4(1-x), & 0.5 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

a. Read in the training set D consisting of $n = 1000$ samples from f and validation set D_v of $m = 300$ samples from files `hw5-f-train.dat, hw5-f-valid.dat`. Use the Gaussian kernel and find the optimal kernel width h by cross-validation. For this, construct $f_h(x)$ the density estimated from D with kernel width h. Then compute the likelihood $L_v(h)$ of the data in D_v under f_h. Also compute $L(h)$, the likelihood of the training set D under f_h. Repeat this for several values of h and plot $L_v(h)$ and $L(h)$ as a function of h on the same graph. (Suggested range of h: 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5).

Let h^* be the h that maximizes $L_v(h)$. Make a plot of $f_{h^*}(x)$ (by, for instance, computing the $f_{h^*}(x)$ values on a grid $x = -0.5, -0.49, -0.48, \ldots 1.49, 1.5$). Plot the true $f(x)$ on the same graph.

The homework you hand in should contain: the formula(s) you used for f_h, the formula(s) you used to compute $L_v(h)$ and $L(h)$ and the required graphs. It is OK to replace likelihoods with log-likelihoods in the plots and equations.

b. Do the same for G and g, reading data from the files `hw5-g-train.dat, hw5-g-valid.dat`.

c. Compare the optimal h's and the quality of the plots in **a, b**. Which of the densities looks easier to approximate? Which of the optimal kernels widths is larger, the one used for f or the one used for g? Can you suggest an explanation why?

[**d.–Extra credit**] Repeat **a,b,c** with the Epanechnikov kernel and the same range of h values. Compare the values of h obtained. What is one similarity and one difference between the graph of f_h here and in **a.**, respectively of g_h here and in **b**?