University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 4

Issued: Friday, April 26, 2013 **Due:** Thursday, May 9, 2013

---

1. This homework will cover materials till May 2. You will have two weeks to work on it. Please include your code with your solution.
2. Just a reminder that the 3-page project midway report is due at 9am on Thursday May 16. You should use the NIPS format provided on the course website, under "Project".
3. You won't have homework due on May 16. The next homework will be released on May 16 and due on May 30.

Problem 4.1

In this question we will analyze a dataset ("CMB.csv") that concerns Cosmic Microwave Background (CMB). The first column is the wavenumber (the x variable), while the second column is the spectrum (the y variable).

In the following you should carefully explain how you fitted the models, for example, in (b) and (c) what kernels and smoothing parameters did you use?

(a) Fit a penalized cubic regression spline with 30 evenly spaced knots using the `mgcv` package.

(b) Fit a Nadaraya-Watson locally constant model.

(i) Set the relative bandwidth to 0.01. Plot the fitted curve.
Hint: The function `locfit(..., deg=0, alpha= specified bandwidth)` *from the* `locfit` *package in* `R` *might be helpful.*

(ii) Set the relative bandwidth to 0.99. Plot the fitted curve.

(iii) Set the relative bandwidth from 0.01 to 0.99 by a 0.01 difference. What is the optimal bandwidth that minimizes GCV score? Plot the fitted curve with the optimal bandwidth.

Hint: The functions `gcvplot(..., deg=0, alpha= a vector of the specified bandwidths )` *from the* `locfit` *package in* `R` *might be helpful.*

(c) Fit a locally linear polynomial model. Repeat the steps (i)-(iii) in part (b).
Hint: The function `locfit(..., deg=1)` *from the* `locfit` *package in* `R` *might be helpful.*

(d) Make a scatter plot of the data points, superimposed with fitted curves by (a)-(c). To see the patterns better, truncate y axis limits to (-1500, 8000). Which of the three models appears to give the best fit just by visual inspection?

Problem 4.2

In this question, we will use a toy example ("toy.csv") with 7 data points. The column "x" includes the covariate values, and the column "y" the observed outcomes. We want to estimate the outcome y by an unknown function $f(x)$.
A Gaussian process (GP) provides a distribution over functions. In this problem, we consider a GP defined as $f \sim GP(0, \kappa)$, where $\kappa(x, x') = \exp\left(-\frac{1}{2}(x - x')^2\right)$.

(a) Draw 100 random samples of f from its prior and plot them.

(b) Draw 100 random samples of f from its posterior and plot them.

(c) Construct the predictive distribution for $x = -2.8$ and plot the resulting 1D predictive distribution.

(d) Construct the predictive distribution for $x = 1$ and plot the resulting 1D predictive distribution. Compare the two predictive distribution by (c) and (d) and explain the difference.

Problem 4.3

In this question you will use a finite Dirichlet mixture of Gaussians to do density estimation for eruption duration of the Old Faithful Geyser (in "OldFaithful.csv"). Specifically, we consider the following model specification:

$$\pi \sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \tag{1}$$

$$\mu_k \mid \sigma_k^2 \sim N(0, \gamma\sigma_k^2) \quad \sigma_k^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 S_0}{2}\right) \quad k = 1, \ldots, K \tag{2}$$

$$z_i \mid \pi \sim \pi \quad i = 1, \ldots, n \tag{3}$$

$$y_i \mid z_i, \{\mu_k, \sigma_k^2\} \sim N(\mu_{z_i}, \sigma_{z_i}^2) \quad i = 1, \ldots, n. \tag{4}$$

Here, IG denotes the inverse gamma distribution and Dir the finite Dirichlet with K components in this case.

Set $\alpha = 1$, $\gamma = 1$, $\nu_0 = 1$ and $S_0 = 1$.

(a) Using $K = 2$, show the estimated density after the 5000th MCMC iterations. Run this procedure 10 times with 10 different sets of initial values for $\pi, \{\mu_k, \sigma_k^2\}$ and plot all resulting densities on one figure.
Hint: You may find the "bayesmix" in R helpful.

(b) Using $K = 10$ and the default initial values in "bayesmix" (if you use other computing languages, just choose initial values by drawing parameters from their priors), show the estimated density after the 5000th MCMC iterations. Comment on any similarities to or differences from the plot of part (a), and justify by examining the outputted posterior samples of π from the sampler.