University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 4

Issued: Friday, April 24, 2013 **Due:** Thursday, May 8, 2013

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 fit the models, for example, in (b) and (c) what smoothing parameters did you use (for a gaussian kernel)?

(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? Also, compare the two GCV-optimal curves from the local linear and local constant fits. Name two key differences between the curves.

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 $\kappa(x, x') = \exp\left(-\frac{1}{\sigma^2}(x - x')^2\right)$ and $\sigma^2 = 2, 10$.

(a) For $\sigma^2 = 2$, draw 100 random samples of f from its prior and plot them.

(b) For $\sigma^2 = 2$, draw 100 random samples of f from its posterior and plot them.

(c) For $\sigma^2 = 10$, draw 100 random samples of f from its prior and plot them.

(d) For $\sigma^2 = 10$, draw 100 random samples of f from its posterior and plot them.

(e) Interpret your plots from part (a)-(d). Why are there regions of low variability and regions of high variability (in the posterior plots)? What does changing σ^2 do to the prior/posterior plots? What does this imply about how the choice of σ^2 can affect your posterior samples of your function f?

(f) For $\sigma^2 = 2$, construct the predictive distributions for $x = 1$ and $x = -2.8$ and plot the resulting 1D predictive distributions. What are the fully-specified forms of these distributions? Why are the variances so different (use the posterior plot to help you)?

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 "Old-Faithful.csv").
Specifically, we consider the following model specification:

$$\begin{aligned}
\pi &\sim \text{Dir}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right) && (1)\\
\mu_k \mid \sigma_k^2 &\sim N(0, \gamma\sigma_k^2) \quad \sigma_k^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 S_0}{2}\right) \quad k = 1, \dots, K && (2)\\
z_i \mid \pi &\sim \pi \quad i = 1, \dots, n && (3)\\
y_i \mid z_i, \{\mu_k, \sigma_k^2\} &\sim N(\mu_{z_i}, \sigma_{z_i}^2) \quad i = 1, \dots, n. && (4)
\end{aligned}$$

Here, IG denotes the inverse gamma distribution and Dir the finite Dirichlet with K components in this case. Fit the densities using the first 242 observations, leave the last 30 observations for testing.

Set $\alpha = 1$, $\gamma = 5$, $\nu_0 = 0.1$ and $S_0 = 1$.

(a) Using $K = \{2, 10\}$, show the estimated densities from 5000 MCMC iterations by using the averaged density estimates from the last 500 iterations and plot the resulting densities ontop of the histogram of the data.
Hint: You may find the "bayesmix" package in R helpful.

(b) Using $K = \{2, 10\}$, show the estimated densities using the EM algorithm (for Maximum Likelihood) estimates as the final $\{\hat{\pi}, \hat{\mu}_k, \hat{\sigma}_k^2\}$ estimates and plot the resulting densities ontop of the histogram of the data.
Hint: You may find the "mclust" package in R helpful.

(c) What do you think is an appropriate number of clusters, 2 or 10? (Hint: think about how many you'd need to adequately fit the data). Are there any systematic differences between the density estimates from part A and B? If there are, do you think one method is better than the other in this situation?

(d) Calculate the log-likelihood of the test data for all 4 scenarios. In the Bayesian setting, the predictive likelihood is estimated as:

$P(Y^*|Y) = \int P(Y^*|\theta)P(\theta|Y)d\theta \approx \frac{1}{500}\sum_{i=1}^{500} P(Y^*|\theta^{(i)})$

for observed data Y and test data Y^*, where $\theta^{(i)}$ is a draw of the model parameters from the i'th MCMC iteration. Use iterations [4501:5000].

For the EM approach, we just compute the test likelihood using our plug-in estimator:

$P(Y^*|X^*, \hat{\theta}) = \sum_{i=1}^{K} \hat{\pi}_i \times [N(X^* \mid \hat{\mu}_i, \hat{\sigma_i^2})]$

Draw a boxplot of these 500 $\log(P(Y^*|\theta^{(i)}))$ values for $K = 2, 10$, and super-impose lines for the $\log(P(Y^*|Y))$ averaged estimate and the EM test log-likelihood $\log(P(Y^*|X^*, \hat{\theta}))$, where $\hat{\theta}$ is our final EM model parameter estimates.