University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 5

Issued: Friday, May 17, 2013 **Due:** Thursday, May 30, 2013

This is the last homework! It will cover material until May 23. You will have two weeks to work on it. Please include your code with your solution.

Problem 5.1

In this question you will use a Dirichlet process mixture of Gaussians to do density estimation for the eruption duration of the Old Faithful Geyser (in “OldFaithful.csv”). Specifically, in the model below,

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

Here, IG denotes the inverse gamma distribution and “Stick” the stick-breaking process.

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

(a) Show the estimated density at the 1st, 10th, 100th and 1000th MCMC iterations.
Hint: You may find the “DPpackage” in R helpful.

(b) Run the MCMC sampling with 1000 iterations. Generate a trace plot of the number of clusters (the number of clusters at the t-th iteration vs. t). Plot a histogram of the number of clusters over the 1000 iterations.
Hint: You may find the “DPpackage” in R helpful.

(c) **Bonus Part (worth 10 credits):**
Write your own code for a Dirichlet process mixture of Gaussians collapsed Gibbs sampler to do density estimation with the Old Faithful Geyser data.

(i) (5 credits) Reproduce the results in (a) and (b). You can initialize the z_i’s by assigning them to the same cluster, i.e. starting with 1 cluster. [*Note: this is a simple, though not ideal initialization. A more intricate initialization starts by assigning the first observation to cluster 1. Then assign the next observation as if it were the last, conditioning on the assignment of the first observation. That is, ignore all other data points. Continue this process until all observations have been assigned.*]

(ii) (5 credits) Evaluate the marginal likelihood at every iteration t:

$$P(y_{1:N}|z_{1:N}) = \prod_{k=1}^{K(t)} \int \left(\prod_{i:z_i=k} p(y_i|z_i = k, \mu_k, \sigma_k^2) \right) p(\mu_k, \sigma_k^2) d\mu_k d\sigma_k^2 \tag{5}$$

Generate a trace plot of the log-likelihood at iteration t vs. t.

Problem 5.2

Suppose that we specify linear splines with two truncated lines for each of the two predictors X_1 and X_2, and therefor four basis in each dimension:

$$1, x_1, (x_1 - \xi_{11})_+, (x_1 - \xi_{12})_+ \tag{6}$$

$$1, x_2, (x_2 - \xi_{21})_+, (x_2 - \xi_{22})_+ \tag{7}$$

(a) How many basis functions does the tensor product splines model contain? Write down the form of all basis in this model.

(b) Suppose X_1 and X_2 are from 0 to 1. Let $\xi_{11} = \xi_{21} = 1/3$ and $\xi_{12} = \xi_{22} = 2/3$. Graphically display in 3-D all basis functions for a tensor product splines model. Please include the name of basis with each plot.
Hint: You may find the function "persp()" helpful.

Problem 5.3

This problem considers daily air quality measurements in New York, from May to September in 1973 (in "airQuality.csv"). There are 153 observations on 6 variables: Ozone (ppb), Solar.R (solar radiation, measured in Langleys), Wind (mph), Temperature (degrees F), Month ($5 - 9$), Day of month ($1 - 31$). The response variable we are interested in is the cube root of ozone. (Note: you need to transform the response variable yourself.)

(a) Consider all 5 predictors. Fit a GAM model with a cubic regression spline to Solar.R, a cubic regression spline to Wind, a two-dimensional thin-plate regression spline to Time and Temperature and an effect for Month (modeled as a factor or a categorical variable). Explain the method by which you choose the smoothing parameter. Please briefly describe the model fit, such as the number of degrees of freedom and if any predictor does or doesn't seem to have a strong association with the response variable.
Hint: You may find the "gam()" function in "mgcv" package helpful.

(b) Plot the resulting 1-D smoothing curve on Solar.R and comment on its relationship with the response variable. Repeat the same exercise to Wind.

(c) Plot the fitted surface of the two-dimensional thin-plate regression spline of Time and Temperature. Describe their relationship with the response variable.

(d) Examine the adequacy of the GAM model you fit, using residual plots of residuals vs. each predictor and the fitted value respectively.

Problem 5.4

In this problem we will do spam classification. Consider the email spam dataset (in "spam.data"). This consists of 4601 email messages, from which 57 features (or covariates) have been extracted. These are as follows:

- 48 features, in $[0, 100]$, giving the percentage of words in a given message which match a given word on the list. The list contains words such as "business", "free", "george", etc.
- 6 features, in $[0, 100]$, giving the percentage of characters in the email that match a given character on the list. The characters are ; ([! $ #
- Feature 55: The average length of an uninterrupted sequence of capital letters
- Feature 56: The length of the longest uninterrupted sequence of capital letters
- Feature 57: The sum of the lengths of uninterrupted sequence of capital letters

More detail about the data can be found in the file "spam.info".

Load the data from "spam.data", in which Columns 1-57 are the features and Column 58 is our response variable, the indicator of spam emails (1 is spam and 0 is non-spam). Divide the dataset into a training set (of size 3065) and a test set (of size 1536) by the indicator in the file "spam.traintest" (1 is test set and 0 is training set). Before analyzing the data, we need to standardize the covariates so that they have mean 0 and unit variance.

(a) **Write your own code** to implement the gradient descent algorithm to fit a logistic regression model with L_2 regularization to all parameters including intercept. Use the test set to choose the penalization parameter λ. For this exercise, you can set the step size in gradient descent algorithm to 0.001 and tuning λ from 0 to 5 by a 0.1 difference. Make binary predictions from the model with the chosen λ and report the error rate on the training and test sets.

(b) (i) Fit a classification tree to the training set. Plot the fitted un-pruned tree. Make binary predictions and report the error rate on the training and test sets.

(ii) Prune the tree and select the optimal tuning parameter λ by minimizing the 10-fold cross validation error with the one standard error rule (1-SE Rule). The 1-SE Rule means that we should choose a simpler model if its penalized RSS is less than 1 SE worse than the next complex model. Plot the fitted pruned tree. Make binary predictions and report the error rate on the training and test sets.

Hint: You may find the "rpart" package in R helpful. The examples in the introduction below are also helpful.
http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

(c) **Bonus part (worth 5 credits):** Binarize the features using $I(x_{ij} > 0)$ for all covariates for all data points. Fit a naive Bayes model with the binarized covariates. Report the error rate on the training and test sets.
Hint: You may find the "klaR" package in R helpful