Stat 274/374: Nonparametric Inference

Assignment 2

Due: Thursday, November 1, 2018

1. *The Mean Shift Algorithm* (20 points)

   Suppose that we wish to estimate the modes of a density p. Let the gradient of p be denoted $g(x) = p'(x)$ (the derivative with respect to x). The gradient defines a family of integral curves π that are solutions of the differential equation

   $$\pi'(t) = g(\pi(t)).$$

   This gives an assignment of points to modes. If p has modes $m_1, \ldots, m_k$ then a point x is assigned to mode m_j if the gradient ascent curve through x (the integral curve through x) leads to m_j.

   Now, suppose we have a sample of (one-dimensional) data $X_1, \ldots, X_n$ from p. Let

   $$\widehat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

   be the kernel density estimator, using a Gaussian kernel. To map a point w to a mode, set $w_0 = w$ and then iterate as follows:

   $$w_{t+1} = \frac{\displaystyle\sum_{i=1}^{n} X_i K\left(\frac{w_t - X_i}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{w_t - X_i}{h}\right)}.$$

   This is, w_{t+1} is the weighted average of the points near w_t. The path $w_0, w_1, \ldots, w_t, \ldots$ converges to a mode m_j. Running this mean shift algorithm on a subset of the points $X_1, \ldots, X_n$ gives an estimate of the modes of p.

   (a) Explain how this algorithm is an approximation to the solution of the differential equation defining the integral curves.

   (b) Two data sets are provided on the Canvas site for this problem. One is a sample of size $n = 500$, the other a sample of size 10,000 from a density p. Use the mean shift algorithm for each data set to give an estimate of the modes. Plot your density estimates, and estimated modes.

2. *The Wizard of Ozone* (20 points)

The data for this problem are daily maximum 8-hour ozone concentration (in parts-per-billion) at 153 sites in the US midwest near Lake Michigan, for 89 days during the summer of 1987. To get the data in R, install and load the package `fields` and use the command `data(ozone2)`. The list `ozone2` contains

- `lon.lat`: longitudes and latitudes of the 153 stations;
- `y`: measurements at each stations on each day;
- `station.id` and `dates`.

(a) Estimate the ozone concentrations of the area on June 18, 1987 using a kernel smoother. In particular,

  i. Specify a grid on the longitude and the latitude using `x=seq(-93,-82,.1)` and `y=seq(40,46,.1)`.

  ii. Write your own code for a 2-d kernel smoother using a Gaussian kernel with bandwidth h and use it to estimate the ozone concentrations at the grid points.

  iii. Perform a cross-validation for choosing the bandwidth h. Plot the cross-validation scores against the bandwidths. What value of the bandwidth will you use? Explain.

  iv. Suppose `z` is the matrix of size `length(x)` by `length(y)` containing your kernel smoothing estimates at the grid points. Visualize your estimate as a heatmap on a regional map by using the following command in R:

```
image.plot(x, y, z, col=rainbow(128,alpha=.5))
US(add=T, lwd=2, col=1)
```

(b) In analogy with the mean shift algorithm in Problem 1, derive an iterative updating rule that takes a starting point on the 2-d plane to a mode of the fitted regression function. Pick a few starting points in the plot you get in part (a), use your updating formula to map them to the modes. Draw the trajectories on top of your heatmap.

(c) Now suppose that you want to estimate the ozone concentrations in the area for several consecutive days. Can you propose a nonparametric estimator, taking into account the smoothness and dependence both spatially and temporally?

3. *How to Win Friends and Influence Functions* (20 points)

Let $X_1, \cdots, X_n \sim F$ where F is strictly increasing with positive density f, and let $\widehat{F}_n$ be the empirical distribution function. Let $\theta = T(F) = F^{-1}(p)$ be the pth quantile.

(a) Show that the influence function of $T(F)$ is

$$L(x) = \begin{cases} \dfrac{p-1}{f(\theta)}, & x \leq \theta \\[2ex] \dfrac{p}{f(\theta)}, & x > \theta. \end{cases}$$

(b) Find the estimate $\widehat{\theta}$ and the estimated standard error of $\widehat{\theta}$.

(c) Find an expression for an approximate $1 - \alpha$ confidence interval for θ.

(d) Detail the bootstrap algorithm for estimating the standard error of $\widehat{\theta}$.

4. *Pulling Yourself Up by the Bootstrap* (20 points)

For each of the statistical models described below, find 95% confidence intervals for θ based on $B = 1{,}000$ bootstrap replicates. Repeat the experiments 500 times and report the coverage of your confidence interval on the true value of θ.

(a) Let $Y = g(X) + \epsilon$ where $X, Y \in \mathbb{R}$ and $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ we can estimate $\beta = (\beta_0, \beta_1, \beta_2)$ with the least squares estimator $\widehat{\beta}$. Suppose that $g(x)$ is concave and we are interested in the location at which $g(x)$ is maximized. It is easy to see that the maximum occurs at $x = \theta$ where $\theta = -\beta_1/(2\beta_2)$. A point estimate of θ is $\widehat{\theta} = -\widehat{\beta}_1/(2\widehat{\beta}_2)$. Take $n = 100$, $\beta_0 = -1$, $\beta_1 = 2$, $\beta_2 = -1$, and generate $X \sim \text{Unif}(0, 2)$ and $\epsilon \sim N(0, 0.2^2)$.

(b) Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n) \overset{\text{i.i.d.}}{\sim} P$. The partial correlation of X and Y given Z is

$$\theta = -\frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}}$$

where $\Omega = \Sigma^{-1}$ and Σ is the covariance matrix of $(X, Y, Z)^T$. The partial correlation measures the linear dependence between X and Y after removing the effect of Z. A point estimator of θ is $\widehat{\theta} = -\frac{\widehat{\Omega}_{12}}{\sqrt{\widehat{\Omega}_{11}\widehat{\Omega}_{22}}}$. Take $n = 100$, and generate $Z \sim N(0, 1)$, $X = 10Z + \epsilon$, and $Y = 10Z + \delta$ where $\epsilon, \delta \sim N(0, 1)$.

(c) Let $X_1, \ldots, X_n \sim N(0, \Sigma)$ where Σ is a $p \times p$ positive definite matrix. Let θ be the smallest eigenvalue of Σ. A point estimator for θ is the smallest eigenvalue of the sample covariance matrix. Take $n = 100$ and Σ to be the 10×10 identity matrix.

5. *Enjoying the (Dirichlet) Process* (20 points)

(a) Let $w_1, w_2, \ldots$ be the weights generated from the stick-breaking process. Show that $\sum_{j=1}^{\infty} w_j = 1$ with probability 1.

(b) Let $F \sim \text{DP}(\alpha, F_0)$. Show that $\mathbb{E}(F) = F_0$. Show that the prior becomes more concentrated around F_0 as $\alpha \to \infty$.

(c) Find a bound on

$$\mathbb{P}\left(\sup_x |\overline{F}_n(x) - F(x)| > \epsilon\right) \tag{1}$$

where $\overline{F}_n$ is defined by

$$\overline{F}_n = \frac{n}{n+\alpha} F_n + \frac{\alpha}{n+\alpha} F_0.$$

(d) Now set $F_0 = N(0,1)$. Draw 10 random distributions from the Dirichlet process and plot them. Try several different values of α.

(e) For each of $n = 10,\ 25$, and 100, do the following:

   i. Draw $X_1, \ldots, X_n \sim F$ where $F = N(5,3)$. Compute and plot the empirical distribution function and plot a 95% confidence band, using DKW.

   ii. Now compute the Bayesian posterior using a $\mathrm{DP}(\alpha, F_0)$ prior with $F_0 = N(0,1)$. Plot the Bayes estimator $\overline{F}_n$. Compute a 95% Bayesian credible band. Draw several random distributions from the posterior and plot them. (Try a few different values of α.)

   iii. Repeat the entire process above many times (without the plotting), and report the fraction of times the Bayesian confidence bands actually contain F. What does this say about the frequentist properties of the Bayesian confidence bands?