Stat 274/374: Nonparametric Inference

Assignment 1

Due: Thursday, October 13, 2016

1. *Computing and plotting with R* (15 points)

   If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then $\mathbb{E}(\mu_n - \mu)^2 \asymp \sigma^2/n$ and $\sqrt{n}(\mu_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, where $\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. In this problem you are asked to confirm these facts by simulating data in R.

   (a) For each $n = 1, 2, \ldots, 200$, simulate data $X_1, \ldots, X_n \sim N(1, \sigma^2)$, for some σ^2 that you may choose. You can use the R function `rnorm` for this. Compute the sample mean μ_n.

   For each n, repeat this B times where B is chosen at your discretion. Estimate the mean $\mathbb{E}(\mu_n - \mu)^2$ by averaging over the B trials. Plot this empirical mean squared error as a function of n, and compare against σ^2/n. Also plot this on a log-log scale.

   (b) Repeat the above, but now plot (for selected n) a nonparametric estimate of the density of $Z = \sqrt{n}(\mu_n - \mu)$ using `plot(density(Z))` where `Z` are the values of the statistic computed on the B samples. Compare this estimated density to the true normal density in your plot.

2. *Leave-one-out cross-validation* (20 points)

   (a) Let $\widehat{r}_n$ be a linear smoother. Show that the leave-one-out cross-validation score $\widehat{R}(h)$ can be written as

   $$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^n \left(\frac{Y_i - \widehat{r}_n(x_i)}{1 - L_{ii}}\right)^2$$

   where $L_{ii} = \ell_i(x_i)$ is the ith diagonal element of the smoothing matrix L.

   (b) Let $r(x)$ be the Doppler function defined as

   $$r(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x + 0.05}\right).$$

   Generate 1000 observations from the model $Y_i = r(x_i) + \sigma\epsilon_i$ where $x_i = i/n$, $\sigma = 0.1$ and $\epsilon_i \sim N(0, 1)$. Fit a local linear regression on the data with your own choice of a kernel function. Use the formula derived in part (a) to make a plot of cross-validation scores versus bandwidths. Plot the data and your local linear estimates using the optimal bandwidth. Plot the confidence interval $I_n(x) = \widehat{r}_n(x) \pm z_{\alpha/2}\widehat{\sigma}(x)\|l(x)\|$, where $z_{\alpha/2} \approx 1.96$. Is $I_n(x)$ the 95 percent pointwise confidence interval for $r(x)$?

3. *Kernel density estimates for Old Faithful* (15 points)

   The dataset `faithful` in R contains the waiting time between eruptions (`waiting`) and duration of the eruptions (`eruptions`) for the Old Faithful Geyser (see `en.wikipedia.org/wiki/Old_Faithful`). You can load the data using `data(faithful)`.

   For both the waiting time and the duration of the eruptions, estimate the one-dimensional density densities using kernel density estimates, plot the cross-validation score versus the bandwidths (see equation (6.5) in Chapter 6 in AoNS), and plot the estimated density using the optimal bandwidth.

4. *Risk of a two-dimensional Kernel density estimate* (20 points)

   Suppose that we observe n pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ such that (X_i, Y_i) are i.i.d. with a density $p(\cdot, \cdot)$ in $\mathbb{R}^2$. Consider the kernel density estimator defined by

$$\widehat{p}_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right)$$

   where $K : \mathbb{R} \to \mathbb{R}$ is a kernel defined as usual and $h > 0$ is a bandwidth. Now assume that the density $p(\cdot, \cdot)$ belongs to the class of all probability densities on $\mathbb{R}^2$ satisfying

$$|p(x, y) - p(x', y')| \leq L\left(|x - x'|^\beta + |y - y'|^\beta\right), \quad \forall (x, y), (x', y') \in \mathbb{R}^2,$$

   with given constant $0 < \beta \leq 1$ and $L > 0$. Let (x_0, y_0) be a fixed point in $\mathbb{R}^2$. Derive upper bounds for the bias and the variance of $\widehat{p}_n(x_0, y_0)$ and an upper bound on the mean squared error at (x_0, y_0). Find the minimizer $h = h_n^*$ of the upper bound on the risk and the corresponding rate of convergence.

5. *Capital Bikeshare* (30 points)

   In a bike sharing system the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. In this problem, you will try to combine historical usage patterns with weather data to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

   You are provided hourly rental data collected from the Capital Bikeshare system spanning two years. The file `train.txt`, as the training set, contains data for the first 19 days of each month, while `test.txt`, as the test set, contains data from the 20th to the end of the month. The dataset includes the following information:

   `daylabel` - day number ranging from 1 to 731
   `year`, `month`, `day`, `hour` - hourly date
   `season` - 1 = spring, 2 = summer, 3 = fall, 4 = winter
   `holiday` - whether the day is considered a holiday
   `workingday` - whether the day is neither a weekend nor a holiday

`weather` - 1 = clear, few clouds, partly cloudy
2 = mist + cloudy, mist + broken clouds, mist + few clouds, mist
3 = light snow, light rain + thunderstorm + scattered clouds, light rain + scattered clouds
4 = heavy rain + ice pallets + thunderstorm + mist, snow + fog
`temp` - temperature in Celsius
`atemp` - "feels like" temperature in Celsius
`humidity` - relative humidity
`windspeed` - wind speed
`count` - number of total rentals

Predictions are evaluated using the root mean squared logarithmic error (RMSLE), calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(m_i+1)-\log(\widehat{m}_i+1)\right)^2}$$

where m_i is the true count, $\widehat{m}_i$ is the estimate, and n is the number of entries to be evaluated.

(a) For the purpose of evaluating your models, divide the training set into two parts: first 15 days of each month as your new training set and 16th to 19th as your validate set. Using this new training set, fit a linear model on `count` numbers against (a subset of) the time and weather variables. You will first need to transform the count numbers to $\log(\texttt{count}+1)$. Pick those variables which you think are relevant. Be careful about whether to include them numerically or as factors. You might also want to include any interaction terms that you think are necessary. Report the model that you fit, and report the RMSLE score evaluated on your own validate set.

(b) Now keep working on the new training set and take a step further than the linear model. For the *transformed* `count` numbers, compute the mean hourly log counts for each day and make scatterplots of the means versus `daylabel`. To account for this main trend in terms of time, fit local linear regressions to the means against `daylabel`. After fitting the nonparametric curve, use the hourly residuals as your new responses, and fit the same model as in part (a). Report the score evaluated on the validate set.

(c) You can be "more nonparametric" by fitting an additive model. Include `daylabel` in your model and treat it, along with other numerical variables such as temperature, nonparametrically. Again, report your model and the score obtained on the validate set. You can fit the additive model by using the gam package in R.

(d) Now, based on the results you obtained for the previous problems, fit a model on the original training set, and predict the total rental counts for each entry in the test set. Record your predicted counts in a file `assn1-<your_cnet_id>.txt` and send it to `nonparametric16fall@gmail.com`. Your file should contain only one column vector with 6493 entries. We will compute the RMSLE of the predictions, and the points you receive for this part will depend on your relative ranking in the class.