Assignment 2 - CSC/DSC 265/465 - Spring 2020 - Due April 16

Note that problem **Q5** is reserved for graduate students.

Q1: Suppose, conditional on parameter $\theta \in (0,1)$, a random variable X has distribution $X \sim bin(n, \theta)$. Then suppose we assign a prior distribution $\pi(\theta)$ to θ of the form $\pi(1/4) = \pi(1/2) = \pi(3/4) = 1/3$. If $n = 10$ and we observe $X = 4$, give the posterior distribution $\pi(\theta \mid X)$ of θ.

Q2: Suppose the traffic flow rate on a certain highway is expressed as λ vehicles per hour. In other words, the number of vehicles which pass a given point in one hour is, on average, λ.

Then, suppose at one point on this highway there is a toll booth installation with a maximum capacity of $N = 10$ toll booths. The number of tolls currently open is $M \leq N$. When a car approaches the installation, any one of the open tolls is chosen at random. As an approximation, we will assume that the number of vehicles which pass through a given open toll in a time interval of length T has a Poisson distribution with mean $\lambda T/M$. We may also assume that the number of vehicles passing through distinct toll booths are statistically independent.

We will construct a Bayesian model for the estimation of λ. Both λ and M are unknown, and have (independent) prior distributions $\pi_\lambda(\lambda)$, $\pi_M(m)$. The observation will be a low resolution "snapshot" of the toll booth installation. Within a period of $T = 15$ seconds, X is the number of toll booths through which at least one vehicle has passed.

(a) What is $f(x \mid \lambda, M)$, the distribution of X conditional on (λ, M)?

(b) Give an expression for the posterior distribution of the parameter pair (λ, M).

(c) Create a Hastings-Metropolis MCMC algorithm to simulate a sample from the posterior distribution of (λ, M). Include the following elements:

(i) The prior distribution of λ is uniform.

(ii) The prior distribution of M is given by $\pi_M(8) = 4/7$, $\pi_M(9) = 2/7$, $\pi_M(10) = 1/7$, with $\pi_M(m) = 0$ for any $m < 8$.

(iii) A proposal distribution generates proposed state (λ_{new}, M_{new}) from (λ_{old}, M_{old}) in the following way. Set

$$\lambda_{new} = \lambda_{old} + U,$$

where $U \sim unif(-10, 10)$ (but replace any negative λ_{new} with zero). If $M_{old} = 8$ or $M_{old} = 10$ set $M_{new} = 9$ with probability one. If $M_{old} = 9$ set $M_{new} = 8$ or $M_{new} = 10$ with equal probabilty. Note that a proposed value is generated for both parameters λ and M within each transition.

(iv) Allow the MCMC to run for 5,000,000 transitions. However, store the current values of λ, M only at intervals of 1,000 transitions.

(v) Use $M_{old} = 8$ and $\lambda_{old} = X \times 240$ as initital states, for observation X.

Run the MCMC sampler four times, using observations $X = 2, 4, 6, 8$.

(d) For each of the four samples, give an estimate of the marginal posterior distribution of M. Compare this to the prior distribution π_M. How do the posterior distributions of M vary with observations $X = 2, 4, 6, 8$?

(e) Create side-by-side boxplots of the sampled λ values for observations $X = 2, 4, 6, 8$. How do the posterior distributions of λ vary with observations $X = 2, 4, 6, 8$?

Q3: This problem will make use of the `Cars93` data set from the `MASS` library, titled *Data from 93 Cars on Sale in the USA in 1993*. First, select a subset of variables from `Cars93` with the following code:

```
> carsb = Cars93[,c(4,5,6,7,8,12,13,14,15,17,19:22,25,26)]
> names(carsb)
 [1] "Min.Price"          "Price"
 [3] "Max.Price"          "MPG.city"
 [5] "MPG.highway"        "EngineSize"
 [7] "Horsepower"         "RPM"
 [9] "Rev.per.mile"       "Fuel.tank.capacity"
[11] "Length"             "Wheelbase"
[13] "Width"              "Turn.circle"
[15] "Weight"             "Origin"
>
```

The first 15 columns of `carsb` are continuous quantitative automobile features. Column 16 is the factor `Origin` possessing two levels `USA`, `non-USA`, indicating whether or not the manufacturer is located in the USA. Do a log transformation of the first 15 columns:

```
> carsb[,-16] = log(carsb[,-16])
```

The objective will be to build a classifier of `Origin` based on the remaining 15 quantitative features. The class to be predicted will be either `USA` or `non-USA`.

	true `USA`	true `non-USA`
predicted `USA`	n_{11}	n_{12}
predicted `non-USA`	n_{21}	n_{22}

Any record used for testing the predictor is placed in exactly one of the four cells.

(a) Recall the odds representation of Baye's Rule, in this application:

$$\begin{aligned} Odds(\text{true } \texttt{USA} \mid \text{predicted } \texttt{USA}) &= LR_+ \times Odds(\text{true } \texttt{USA}) \\ Odds(\text{true } \texttt{USA} \mid \text{predicted } \texttt{non-USA}) &= LR_+ \times Odds(\text{true } \texttt{USA}) \end{aligned}$$

Express LR_+ and LR_- in terms of the elements $(n_{11}, n_{12}, n_{21}, n_{22})$ of the confusion table. Create an `R` function that inputs the confusion table, and outputs a single vector with elements (CE, LR_+, LR_-), where CE is classification error.

(b) Using the function `lda()` fit a classifier using linear discriminant analysis (LDA). Use the function of Part (a) to calculate (CE, LR_+, LR_-). Do not use cross-validation to fit this classifier.

(c) Using the function `qda()` fit a classifier using quadratic discriminant analysis (QDA). Use the function of Part (a) to calculate (CE, LR_+, LR_-). Do not use cross-validation to fit this classifier. Does the QDA classifier appear to improve the LDA classifier?

(d) Repeat Parts (b) and (c) using the `CV = TRUE` option of the `lda()` and `qda()` functions. This yields the predictions resulting from *leave-one-out* cross-validation. How does this change the comparison between the LDA and QDA classifier. Give a brief explanation.

Q4: This problem will make use of the `Pima.tr` data set from the `MASS` library. From the `help` page:

> A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases.

Note that `Pima.tr` is a subset of the complete data set containg $n = 200$ records. First, select a subset of variables from `Pima.tr` with the following code:

```
> pima1 = rbind(Pima.tr)[,c(2,3,4,5,6,8)]
> names(pima1)
[1] "glu"  "bp"   "skin" "bmi"  "ped"  "type"
>
```

The first 5 columns of `pima1` are continuous quantitative automobile features. Column 6 is the factor `type` possessing two levels `Yes` (diabetic according to WHO criteria) and `No`. The objective will be to build a classifier of `type` (ie. to predict diabetes) based on the remaining 5 quantitative features.

(a) Determine the frequencies of the `Yes` and `No` outcomes. If a classifier simply predicted the most frequent outcome for all inputed features, what would the classification error be (assuming the outcome prevalences found in the data hold)?

(b) Create an `R` function which accepts a vector `k.list` of values of K (the neighborhood size of the classifier), a training set of features, and a paired training set of classes. For each K in `k.list` a KNN fit will be evaluated using LOOCV. Use the `knn.cv()` function from the library `class`. The function should output the summary (CE, LR_+, LR_-) for each K in `k.list` (See Question 3).

(c) Apply the function of Part (b) to the data set. Use `k.list = seq(1,125,2)`. What advantage is there to using only odd numbers for K? Plot the estimated CE against K. Superimpose on the plot a horizontile line representive the value of CE estimated for the simple classifier described in Part (a). What is the minimum value of CE, and for what value of K is this attained? Give also the maximum and minimum values of LR_+ and LR_- attained for any K.

(d) Create side-by-side boxplots of the 5 features (make sure you use a common vertical axis). Then normalize each of the 5 features by subtracting the mean, then dividing by the standard deviation. Repeat Part (c). What difference do you notice? Given that the default metric used to define the neighborhood is Euclidean distance, why might normalizing the features improve the classifier?

Q5: [**For Graduate Students**] We will revist Question 2. To simplify the problem we will assume that M is fixed and known to be $M = 10$.

(a) Derive the log-likelihood function of λ for a given observation $X = x$. Plot this function against λ for $X = 6$.

(b) Give a closed form expression for the maximum likelihood estimate (MLE) of λ. Create a table of the MLEs for $x = 0, 1, \ldots, 9$.

(c) What happens when $x = 10$? Is it possible to give an MLE for λ, of any kind, for this case?