Assignment 3: Numerical Problems

September 27, 2019

1. Methods of Classification

Generate 100 (respectively, 250) observations from each of the two classes, and call it the TRAINING set (respectively, the TEST set).

(a) Calculate the Bayes' risk from the test set assuming the underlying probability distributions to be known.

Forget the probability distribution, and consider the data sets ONLY!

(b) Implement LDA and QDA, and compute their misclassification rates on both the training and test sets.

(c) Consider the (i) Gaussian and (ii) uniform kernels with the bandwidth matrix to be $h^d I_d$ for both classes. Construct the kernel density estimate for both classes with an *adaptive* choice of h (see David Scott's book). Compute the misclassification rates for kernel discriminant analysis for this choice of h on both the training and test sets.

Plot the misclassification rate for kernel discriminant analysis for varying values of h on both the training and test sets.

(d) Plot the misclassification rates for the kNN classifier with the l_2 norm for varying values of k on the training, test sets and leave-one-out cross-validation [LOOCV]. Report the *adaptive* choice of k using LOOCV.

(e) Least squares regression with appropriate response variables (Problem 1 in Quiz 1) on this data, and report the misclassification rates on both the training and test sets.

(f) Perform logistic discriminant analysis on this data set, and report the misclassification rates on both the training and test sets.

(g) Use SVM with the *linear, quadratic* and *radial basis function* kernels on this data set, and report the misclassification rates on both the training and test sets.

Consider $\pi_1 = \pi_2 = 0.5$ and $\pi_1 = 0.1, \pi_2 = 0.9$, for $d = 2, 10, 25$ and 50. The probability models for the two classes (say, f_1 and f_2) are as follows:

1. $N_d(0_d, I_d)$ and $N_d(\mu_d, I_d)$, where $\mu_d = (d, 0_{d-1})^T$.
2. $N_d(0_d, I_d)$ and $SN_d(0_d, I_d)$ (Problem 2.(c) in Mid-semester exam).
3. $C_d(0_d, I_d)$ and $C_d(\mu_d, 2I_d)$, where $\mu_d = (0_{d-1}, d)^T$.
4. $[N_d(0_d, I_d) + N_d(2_d, I_d)]/2$ and $[N_d(1_d, I_d) + N_d(3_d, I_d)]/2$.
5. $N_d(0_d, I_d)$ and $N_d(\mu_d, I_d)$, where $\mu_d = (1, \ldots, d)^T$.

Write a comprehensive report (based on these examples and the methods you analyze) on what you infer out of this comprehensive numerical study.

2. Curse of Dimensionality

(a) Draw a sample of size 20 from $N_d(0_d, I_d)$, and compute the length of each of the vectors. Estimate the distribution using a kernel density estimate with an adaptive choice of h, and plot it for $d = 2, 5, 10, 20, 50, 100, 200$ and 500. Is the mean 0_d a representative summary of the distribution in high dimensions?

(b) Consider a sample of size 3 from $N_d(0_d, I_d)$, say $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$. Take $d = 2, 5, 10, 20, 50, 100, 200$ and 500. Compute the angle $\angle(\mathbf{X}_i, \mathbf{X}_j)$ for $1 \leq i < j \leq 3$. Make a table of each of these values.

(c) Consider a training sample of size 20 by generating 10 observations from f_1, and 10 observations from f_2. Take $d = 2, 5, 10, 20, 50, 100, 200$ and 500. Compute the misclassification rate for 1NN, kernel discriminant analysis, SVM with *linear* and *radial basis function* kernels and the Bayes' classifier, and plot them w.r.t. d for the following probability distributions:

1. $N_d(0_d, I_d)$ and $N_d((1, 0_{d-1})^T, 1/4I_d)$.
2. $N_d(0_d, I_d)$ and $N_d(1_d, 1/4I_d)$.
3. $N_d(0_d, I_d)$ and $N_d(0_d, 1/4I_d)$.

Write a report on what you infer out of this numerical study.

---

3. Cross Validation

Consider the regression model $y = 2x^3 + x^2 - 2x + 5 + e$.

Let $x \sim U(-2, 2)$ and $e \sim N(0, 1)$. Generate random samples of size $n = 100$ from both distributions. For each pair of (x, e), compute a y using the model stated above. Plot (y, x) for these 100 observations.

Now, forget this model, and consider ONLY the data!

Fix a $q \in \mathbb{N}$, and consider the data (y, x). Fit a regression model using the response y on the variables $1, x, \ldots, x^q$. Plot the MSE values for varying values of $q \in \{2, 5, 10, 20\}$.

Choose the 'best' value of q using AIC, BIC and method of cross-validation [leave one out, 2-fold, 5-fold]. **Explain what you observe out of this numerical exercise.**

A useful link: http://robjhyndman.com/hyndsight/crossvalidation/.

Total points: 20.

The use of statistical software R is encouraged for writing codes. You can use R packages.

Discussions are encouraged, but NO COPYING. If copying is established (IITK is now using https://www.urkund.com/ to deal with plagiarism), then both (or more) students will get ZERO.

Submit this assignment (in .pdf only), and relevent R codes (.in txt only) to the gmail id: `assignment.stat.iitk@gmail.com` **only. In the subject of the email, please use the format 'roll number - name' (i.e, 12345 - Subhajit).**

Deadline : 09.11.2019 at 11:59pm (IST).
I will go by the time recorded on Gmail.