Homework 3

Shao-Ting Chiu (UIN:433002162)

10/18/22

Table of contents

Homework Description . . . 1
Computational Enviromnent Setup . . . 2
Third-party libraries . . . 2
Version . . . 2
Problem 5.1 . . . 3
Problem 5.2 . . . 3
(a) . . . 3
(b) . . . 3
Problem 5.6 . . . 3
(a) . . . 3
(b) . . . 3
Problem 5.10 (Python Assignment) . . . 4
(a) . . . 4
(b) . . . 4
(c) . . . 4

Homework Description

- Course: ECEN649, Fall2022

Problems from the book:

5.1

5.2

5.6 (a,b)

5.10 (a,b,c)

Challenge (not graded):

5.4

5.6 (c,d)

- Deadline: `Oct. 26th, 11:59 pm`

Computational Enviromnent Setup

Third-party libraries

```
%matplotlib inline
import sys # system information
import matplotlib # plotting
import scipy.stats as st # scientific computing
import pandas as pd # data managing
import numpy as np # numerical comuptation
import numba
import sklearn as sk
from numpy import linalg as LA
import scipy as sp
import scipy.optimize as opt
import sympy as sp
import matplotlib.pyplot as plt
from numpy.linalg import inv, det
from numpy.random import multivariate_normal as mvn
from numpy.random import binomial as binom
# Matplotlib setting
plt.rcParams['text.usetex'] = True
matplotlib.rcParams['figure.dpi']= 300
np.random.seed(20221011)
```

Version

```
print(sys.version)
print(matplotlib.__version__)
print(sp.__version__)
print(np.__version__)
print(pd.__version__)
print(sk.__version__)
```

```
3.8.14 (default, Sep  6 2022, 23:26:50)
[Clang 13.1.6 (clang-1316.0.21.2.5)]
3.3.1
1.6.2
1.19.1
1.1.1
1.1.2
```

Problem 5.1

Consider that an experimenter wants to use A 2-D cubic histogram classification rule, with square cells with side length h_n, and achieve consistency as the sample size n increases, for any possible distribution of the data. If the experimenter lets h_n decrease as $h_n = \frac{1}{\sqrt{n}}$, would they be guaranteed to achieve consistency and why? If not, how would they need to modify the rate of decrease of h_n to achieve consistency?

Problem 5.2

Consider that an experimenter wants to use the kNN classification rule and achieve consistency as the sample size n increases. In each of the following alternatives, answer whether the experimenter is successful and why.

(a)

The experimenter does not know the distribution of (X, Y) and lets k increase as $k = \sqrt{n}$.

(b)

The experimenter does not know the distribution but knows that $\epsilon^* = 0$ and keeps k fixed, $k = 3$.

Problem 5.6

Assume that the feature X in a classification problem is a real number in the interval $[0, 1]$. Assume that the classes are equally likely, with $p(x|Y = 0) = 2xI_{\{0 \leq x \leq 1\}}$ and $p(x|Y = 1) = 2(1 - x)I_{\{0 \leq x \leq 1\}}$.

(a)

Find the Bayes error ϵ^*.

(b)

Find the asymptotic error rate ϵ_{NN} for the NN classification rule.

Problem 5.10 (Python Assignment)

(a)

Modify the code in `c05_kernel.py` to obtain plots for $h = 1, 3, 5, 7, 9, 11$[1] and $n = 50, 100, 250, 500$ per class. Plot the classifiers over the range $[-3, 9] $\times$[-3, 9] in order to visualize the entire data and reduce the marker size from 12 to 8 to facilitate visualization. Which classifiers are closest to the optimal classifier? How do you explain this in terms of underfitting/overfitting? See the coding hint in part (a) of Problem 5.8.

(b)

Compute test set errors for each classifier in part (a), using the same procedure as in part (b) of Problem 5.8. Generate a table containing each classifier plot in part (a) with its test set error rate. Which combinations of sample size and kernel bandwidth produce the top 5 smallest error rates?

(c)

Compute expected error rates for the Gaussian kernel classification rule in part (a), using the same procedure as in part (c) of Problem 5.8. Since error computation is faster here, a larger value $R = 200$ can be used, for better estimation of the expected error rates. Which kernel bandwidth should be used for each sample size?

[1] In Problem 5.10, please replace "k=1,3,5,7,9,11" by "h=0.1,0.3,0.5,1,2,5" — Ulisses on Slack