ECE 411 Homework 6 (Fall 2024)
Instructor: Dr. Chau-Wai Wong
Material Covered: Classification, Maximum Likelihood, Generalized Linear Models

Problem 1 (20 points) [Maximum Likelihood Estimator (MLE)]

a) Calculate the MLE for variance θ (or σ^2, a more common notation if you prefer) for a random sample $X_1, \ldots, X_n$ drawn from the normal distribution with PDF shown as follows:

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-(x-\mu)^2/2\theta}, \quad -\infty < x < \infty. \tag{1}$$

Once you are done with partial differentiating against θ, try to do partial derivative directly against σ. Leave your intermediate steps there and comment on where you have encountered difficulty.

b) Calculate the MLE for parameter b for a random sample $X_1, \ldots, X_n$ drawn from an exponential distribution with PDF of the following form:

$$f(x;b) = \frac{1}{b}\, e^{-x/b}, \quad x \geq 0. \tag{2}$$

c) The exponential distribution is more often parameterized using the rate parameter λ with PDF of the following form:

$$f(x;\lambda) = \lambda\, e^{-\lambda x}, \quad x \geq 0. \tag{3}$$

Use the invariance principle of MLE, show that $\hat{\lambda}_{\text{MLE}} = 1/\bar{X}$.

Problem 2 (20 points) [Generalized Linear Model (GLM)] Response $Y_i \sim \text{B}(n, p_i)$ is a binomial random variable in which n is known. The (conditional) PDF is shown as follows:

$$\mathbb{P}[Y_i = k | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = \binom{n}{k} p_i^k (1-p_i)^{n-k}, \quad k \in \{0, 1, \ldots, n\}. \tag{4}$$

a) Explain why the linear regression may not the best fit to find the relation between Y_i and a set of predictors $X_{i,1}, \ldots, X_{i,q}$.

b) One proposes to link the conditional mean μ_i and the predictors $\underset{\sim}{x}_i$ using a generalized linear model shown as follows:

$$g(\mu_i) = \underset{\sim}{\beta}^T \underset{\sim}{x}_i \tag{5}$$

where $g(u) = \log(\frac{u}{n-u})$ and $\mu_i = \mathbb{E}[Y_i | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = np_i$. From the variable transformation viewpoint, show that $g(\cdot)$ matches the ranges for the two sides of Eq. (5).

c) Rewrite the PDF into an exponential family form shown as follows:

$$f_Y(y;\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{6}$$

where θ is the natural parameter. Show that $g(\cdot)$ in (b) is the canonical link function when taking μ_i as the input.

Problem 3 (20 points) [Various Classification Methods on the Stock Market and the Weekly Data]

a) Complete *ISL**P**-4.7.1–7*. Be super concise when reporting the results.

b) (10' bonus) Complete *ISL**P**-4.8.13*. Be selective when reporting the results.

Problem 4 (Bonus, 20 points) [k-Nearest Neighbors]

a) (10') Complete *ISLP or ISLR-2.4.7*. Repeat (a)–(c) for $(X_1, X_2, X_3) \in \{(1,2,3), (1,-1,1)\}$.

b) (10') Using a programming language of your choice, refactor your code into a function named `MyKnn` with the following input and output variables. We have shown below examples in Python and Matlab, but you may also use R.

Python:

```
def MyKnn(x1, x2, x3, k):
    ...
    return Y
```

Matlab:

```
function Y = MyKnn(x1, x2, x3, k)
...
end
```

The file containing the function should be named `MyKnn` with extension `.py`, `.m`, or `.r` and appended to the homework submission in plaintext. The performance of `MyKnn` will be manually assessed, and bonus points will be given solely on the percentage of correct classifications using test data. You can assume that when the function is evaluated, the input variables `x1`, `x2`, `x3` will be any value in $\mathbb{R}$, `k` will be less than 6, and the return value being checked against will be either `"Red"` or `"Green"`.

(You are given 3 required problems. The rest of time should be devoted to the term project released on 10/18.)