Claremont Graduate University, Mathematics M454, Statistical Learning
Lab 1 In-class Exercises
Instructor: Qidi Peng
Feb. 6, 2018

CH 1: Linear Models and CH 2: Classification

All assignments must be done by using R coding. All proofs must be shown by using illustrations or numerical justifications. Use *set.seed(1)* before generating each random variable.

1. **(Taylor series)** Generate the following function

$$f(n) = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots + \frac{1}{n!}.$$

Plot the sequence $(n, f(n))_{n\geq 1}$. Use your figures to show

$$\lim_{n\to\infty} f(n) = e.$$

(Hint: use "for" loop to create the function $f(n)$. Draw a figure $(n, f(n))$ for $n = 1, \ldots, 100$. If possible add the horizontal line $y = e$ in red.)

2. **(The law of large numbers)** By the law of large numbers, one has

$$\frac{\text{number of } A \text{ occurs}}{\text{total number of trials}} \approx \mathbb{P}(A \text{ occurs}).$$

Now one rolls two fair dies of 6 faces. Approximate the probability of obtaining the event "sum of the two outcomes is equal to 6".
(Hint: using R, generate the random variable $X = X_1 + X_2$, where X_i is the outcome for the ith die. Then use "if else" to calculate $\frac{\text{number of } \{X=6\} \text{ in } N \text{ rolls}}{N}$, for $N = 1, 2, 3, \ldots, 1000$. Finally plot this sequence.)

3. **(Compare estimates)** Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d normal random variables $\mathcal{N}(0, 1)$.
Questions.

(a) For $n \geq 1$, generate $Z_1, Z_2, \ldots, Z_n$.

(b) Use figures to show that

$$\overline{Z}_n := \frac{\sum_{i=1}^{n} Z_i}{n} \xrightarrow[n\to\infty]{a.s.} 0.$$

(Hint: Generate $\{Z_1, Z_2, \ldots, Z_n\}$ for $n = 1000$, then draw the figure $(n, \overline{Z}_n)$ for $n = 1, \ldots, 1000$. If possible add the horizontal line $y = 0$ in red.)

(c) Show that the sample median of $Z_1, Z_2, \ldots, Z_n$ is an unbiased estimate of 0:

$$\mathbb{E}[median(Z_1, \ldots, Z_n)] = 0.$$

(Hint: generate 1000 copies of the vector $(Z_1, \ldots, Z_{100})$, return $mean(median(Z_1, \ldots, Z_{100}))$.)

(d) Generate 1000 copies of the vector $(Z_1, \ldots, Z_{100})$, calculate $var(\overline{Z})$ and $var(median(Z))$, compare them. Make comments on the result.

4. **(Simple linear regression)** A Walmart supermarket has 15 cashiers (but they are not all in service). The table below describes the corresponding customer average waiting time VS the number of cashiers in service.

Number of cashiers in service	3	4	5	6	8	10	12
Average waiting time (in min)	16	12	9.6	7.9	6	4.7	4

Questions.

(a) Let the single predictor $X =$"number of cashiers in service" and the response $Y =$"customer's average waiting time". Determine the least squares coefficient estimates of this simple linear regression model; compute the standard errors SE of the estimates.

(b) Determine a 95% chance confidence interval which contains the estimators obtained in **(a)**.

(c) Provide a 95% prediction interval of the "average waiting time" when all the 15 cashiers are in service.

5. **(Discriminant analysis and K-nearest neighbors)** 6 persons step on an elevator. The sample of $(log(weights), gender)$ (weights are in pound) is observed as

$\log(weights)$	5.00	4.70	4.40	5.12	4.30	5.44
Gender	male	female	female	female	male	male

You are told that after a while a seventh person steps on the elevator with $log(weight) = 4.90$, but you don't see her or him.

Questions.

(a) Use quadratic discriminant analysis (QDA) to predict this person's gender.

(b) Use K-nearest neighbors (KNN) to predict this person's gender (taking $K = 3$ in your codes).