Practice Python

1. Generate a random 20 × 20 grid of two digit natural numbers. Find the largest product of four diagonally adjacent numbers.

2. The Social Security administration has this neat data by year of what names are most popular by gender for babies born that year in the USA. The popular name list for the year 1990 is given in the website `http://mgmt.iisc.ac.in/~parthar`. Write a Python program to identify the names that appear in the list of popular names of both the gender.

Session 1 - Linear Algebra and Linear Programming

3. Write a Python routine to calculate the inverse of a matrix from first principles. Compare the out generated by your routine with the generated by the matrix inversion method in NumPy.

4. An airline operates a set of flights daily with some fixed capacity. Each of these flights are called a flight leg. A passenger might use one or more of these flight legs to travel from him origin to reach the final destination, called the itinerary. The demand and cost of each itinerary is known. The problem faced by the airline is to determine the portion of demand of each itinerary to accept so as to maximise the total revenue. This problem can be formulated as a network flow optimisation problem.

   Let the set of flight legs be **L** and the itineraries be **I**. Let the capacity of a flight leg $l \in \mathbf{L}$ be c_l. Also, let the fare paid by a passenger for an itinerary $i \in \mathbf{I}$ be f_i and the demand for the itinerary is d_i. The decision problem from the airline perspective is to determine the portion of demand x_i that should be carried for itinerary i without violating the capacity constraints and also maximise the total revenue generated. This linear programming problem is as given below in which the function $\mathcal{I}(l)$ maps a flight leg $l \in \mathbf{L}$ to all the itineraries using it.

$$\begin{array}{lrcll} \text{Maximise} & \sum_{i\in I} f_i x_i & & & \\ \text{Subject to} & x_i & \leq & d_i & \forall i \in \mathbf{I} \\ & \sum_{i:i\in\mathcal{I}(l)} x_i & \leq & c_l & \forall l \in \mathbf{L} \end{array}$$

   Write a Python program using the `linprog` function in the SciPy library to solve this problem for the data set provided in the website `http://mgmt.iisc.ac.in/~parthar`. Identify the flight leg that is capacity constrained.

   In the flight leg capacity file you will find entries like `LX01305200209200SLZRHC 11`. The last six characters of the first field indicates the origin and destination of the flight leg. For example, the origin of the above mentioned flight leg is `SLZ` and destinations in `RHC`. Suppose that the destinations city `RHC` is hosting an important event and it results in an increase in demand of those itineraries flying into `RHC` by flat 15% and fares of those itineraries by flat 25%. Now solve the above optimization problem again and identify the effect of this event on the overall revenue and the top 10 most affected itinerary in terms of the optimal seat allocation.

Session 2 - Probability and Statistics

5. Monte Carlo simulation can be used to numerically integrate a funcion $f(x)$ from a to b. Let,

$$F = \int_a^b f(x)dx.$$

x_1 $f_1x_1 + f_2x_2$

This integral can be numerically approximated by averaging samples of the function $f(x)$ at uniform random points in the interval $a \leq x \leq b$. Given N such samples x_i in the required rage the Monte Carlo estimate of this integration is given by

$$\hat{F} = \frac{b-a}{N} \sum_{i=1}^{N} f(x_i).$$

Write a Python program implementing this algorithm. Call this function 100 times and calculate the difference between the estimate and the theoretical value for the following problems.

(a) $f(x) = \int_0^1 e^{-x^2} dx$

(b) $f(x) = \int_1^2 \frac{1}{1+x^2} dx$

(c) $f(x) = \int_0^1 \sqrt{x^4+1} dx$

Construct a confidence interval for the difference and analyse the impact of the choice of N. Perform a hypothesis test with the null hypothesis being the difference between the estimate and the actual value is not zero.

6. Consider n Bernoulli trials with probability of succsess being p. Write a Python method that implements the Bernoulli trial returing success or failure of a trial. Call this method n times and identify the number of successes in these trails. Repeat this experiment N times to estimate the probability mass function of the resultant Binomial distribution. Compare this with the theoretical Binomial distribution with parameters n and p. If n is sufficiently large then the Binomial probabilities can be approximated with Normal distribution. Calculate the probability mass function approximated with the Normal distribution. Plot the theoretical probability mass function, empirical probability mass function and the Normal approximated probability mass function. Study the effect of n on the Normal approximation.

Session 3 - Linear Models

7. Consider the data set on purchasing power parity and national income accounts in international prices for 152 countries over the period 1950 - 1992 (1985 as base year) given in the website `http://mgmt.iisc.ac.in/~parthar`. The objective of this execise is to build multiple linear regression model relating the exchange rate of the local currency to US dollar. There are 28 predictor variables such as population, real per-capita GDP, real investment share of GDP, price level of consumption, rpice level of investment, etc. in the data set. Be sure to consider data transformations and interaction terms in your model. Compare the results of the linear model with that of the K-nearest neighbor regression method. Check the validity of the model with respect to the assumptions made. Build the models for India and three other countries and interpret the models developed.

8. The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit. The data set can be accesses in the website. The data set containts 20 covariates and one binary output variables that indicates if a client subscribed to a term deposit or not. The covariates include demographic variabes, marketing campaign related variables and certain socio-economic attributes. Compare the performance of logictic regression and linear discriminant analysis. Use K-nearest neighbor classification method as a banchmark.