Assignment 1

MATH 7502 - Semsester 2, 2018

Mathematics for Data Science 2

Created by Zhihao Qiao, Maria Kleshnina and Yoni Nazarathy

Question (1)

(a) Let $u = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, v = \begin{bmatrix} -6 \\ 1 \\ -2 \end{bmatrix}, w = \begin{bmatrix} 4 \\ -6 \\ -1 \end{bmatrix},$

prove or disprove that u, v and w lie in the same plane.

(b) Use the dot product to determine the angle between u and v.

(c) Find the length (L2 norm) of u and v. Illustrate that Cauchy-Schwarz inequality and the Triangle Inequality holds for these vectors.

(d) Find the unit vectors associated with u,v and w.

Question 2

(a) Find the unit vectors u_1 and u_2 in the direction of $v = (1, 4)$ and $w = (-2, 1, 2)$ respectively.

(b) Find unit vectors v_1 and v_2 that are parallel to u_1 and u_2 (Are there such vectors that are not u_1 and u_2?).

(c) Find unit vectors w_1 and w_2 that are perpendicular to u_1 and u_2.

(d)Find the dot product of any unit vector $u + w$ and $u - w$.

Question 3

Show the implementation of k-means algorithm (using k=2) for the following table by hand:

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Initialize with individual (1) and (4), in this case the centroids are $m_1 = (1.0, 1.0), m_2 = (5.0, 7.0)$.

(a) Using the initilizations given, verify that two clusters of the next step are $\{1, 2, 3\}$ and $\{4, 5, 6, 7\}$.

(b) Find the new centroids using (a).

(c) Find the most optimized cluster with $k = 2$.

Question 4

Let f be a mapping from $R^2 \to R$,

$$f(x, y) = \left(e^x \sin{(y)}, e^x \cos{(y)}\right)$$

(a) Compute the Jacobian Matrix.

(b) Find its determinant.

(c) State conditions for invertability for every x and y.

Question 5

True or false (give a reason or prove if true or present a counter example if false).

(a) If $u = (1, 1, 1)$ is perpendicular to v and w for any v and w, then v is parallel to w.

(b) If u is perpendicular to v and w, then u is perpendicular to $v + w$.

(c) If u and v are perpendicular unit vectors, then $\|u - v\| = \sqrt{2}$.

Question 6

Suppose each of the vectors $u_1, \dots, u_k$ is a linear combinaton of the vectors $v_1, \dots, v_m$. Assume that w is a linear combination of $u_1, \dots, u_k$.

(a) Show that w is also a linear combianton of $v_1, \dots, v_m$ for the case $m = k = 2$.

(b) Show the above for the genernal m and k. (Note: m does not necessilary equal to k)

Question 7

Determine whether each of the following scalar-valued functions of n-vectors is linear. If it is a linear function, find its inner product representation, i.e Find an n-vector a such that $f(x) = a^T x$. If it is not linear, give specific x, y, α, β such that

$$f(\alpha x + \beta y) \neq \alpha f(x) + \beta f(y)$$

(a) $f(x) = \max_k x_k - \min_k x_k$ (The spread of values of the vector).

(b) $f(x) = x_n - x_1$ (The difference of last element and the first).

(c) $f(x)$ = The median of an n-vector, suppose $n = 2k + 1$ is odd. then the median is the (k+1)th largest number among all entries of x.

(d) Define $x_{n+1} = f(x_n) = x_n + (x_n - x_{n-1})$, for $n \geq 2$ (This is a simple prediction of what x_{n+1} should be based on a straight line drawn through x_n and x_{n-1}.

Question 8

Clustering a collection of vectors into $k = 2$ groups is called 2-way partitioning, since we are partitioning the the vector into 2 groups, with index sets G_1 and G_2. Suppose we run k-means, with $k = 2$, on the n-vectors $x_1, \dots, x_n$ with $x_i \in R^n$.

Show that there is a nonzero vector w and a scalar v that statisfy

$$\begin{aligned} w^T x_i + v \geq 0 \ \text{ for } \ i \in G_1 \\ w^T x_i + v \leq 0 \ \text{ for } \ i \in G_2 \end{aligned}$$

In other words, the affine function $f(x) = w^T x_i + v$ is greater than or equal to zero in the first group, and less than or equal to zero in the second group. This is called the linear separation of the two groups. Hint: Consider the function $\|x - z_1\|^2 - \|x - z_2\|^2$, where z_1 and z_2 are the group representatives.

Question 9

Use the same data set as in Question 3,

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

(a) Write code to find the distance between individuals and the initialization for $k = 2$.

(b) Write code to cluster the data into two set and find the new centorid.

(c) Using (a) and (b), write code to find the most optimized clustering with k=2. (Does your answer agree with Q3?)

(d) Repeat (c) with $k = 3$.

Question 11

Consider the mapping $f : R^2 \to R$

$$f\left(x_1, x_2\right) = \sin\left(x_1\right) \exp\left(x_2\right).$$

(a) Find the Taylor approximation at $z = (1, 1)$.

(b) Plot function f.

(c) Plot the taylor approximation with different numbers of terms, illrustrates the closeness of the approximating to the acutal function, comment on your result.

Question 11

This Question requires some package (using Distributions, Pyplot and etc.) (https://juliastats.github.io/Distributions.jl/latest/fit.html (https://juliastats.github.io/Distributions.jl/latest/fit.html))

(a) Generate four 3-vectors using different distributions from the above document. i.e rand(sampledist(),3).

(b) Find the unit vectors of the 3-vectors you found in (a).

(c) Find the dot product of two 3-vectors generated by the Normal distribution.

(d) Fill the follwing code which generates a vector that stores 20,000 dot products between two 3-vector generated by the Normal distribution

In [ ]:

```
using Distributions,PyPlot

storage=zeros(20000)

for i in 1:20000
    m= ??????????
    a[i]=m
end
```

(e) Using "hist" function from Pyplot, create a histogram plot of the dot products, and comment on your result.

(f) Repeat (d) and (e) with Beta distribution (choosing some parameters), and comment on your result.

Question 12

Plot the solution of the following ODE:

$$\frac{du}{dt} = f(u, p, t).$$

On the time interval $t \in [0, 5]$, where $f(u, p, t) = tu - u^2$ with initial condition $u_0 = 2$.