# CE956: Statistical Learning
# Department of Computer Engineering
# Sharif University of Technology
# Spring 2019: Room CE204, Sat. & Mon.: 13:30-15:00

## Quiz 01 (30 Points) – (February-23-2019)

## Solution

**Linear Algebra:**

1. What is the Column representation of this linear equation? What do the columns of A represent?

We can write Ax = b as: $A_1*x_1 + A_2*x_2 + \ldots + A_n*x_n = b$ where $A_i$ is the i'th column of A.
Each column of A represents effect of corresponding element of x in the corresponding element in b. If the columns of A are independent, then they correspond to basis vectors.

2. What is a pivot, what is the relation between the pivot, inevitability of A and consistency of solution space?

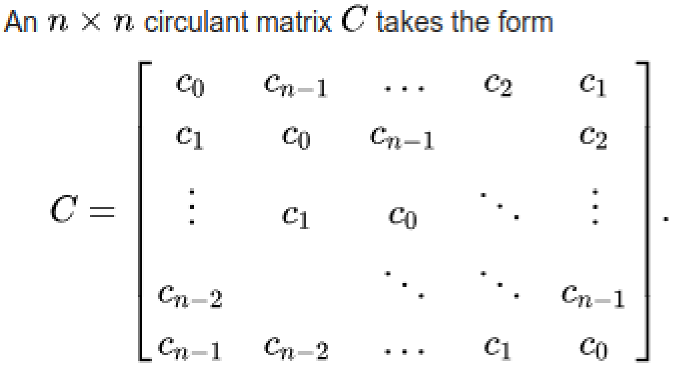
in a reduced row echelon form of a matrix, the leading non-zero element in each row is called a pivot. If A is invertible then it has 'n' pivot positions and has a unique solution.

3. What do Eigenvalues and Eigenvectors represent?

Eigenvetors show points in the problem space which their direction doesn't change under transformation by 'A' and are only multiplied by a scalar which is it's corresponding eigenvalue.

4. What are the properties of a circulant matrix?

A circulant matrix is a special kind of Toeplitz matrix where each row vector is rotated one element to the right relative to the preceding row vector:

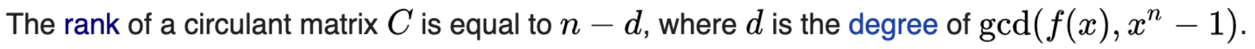


An $n \times n$ circulant matrix $C$ takes the form

$$C = \begin{bmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & c_1 & c_0 \end{bmatrix}.$$

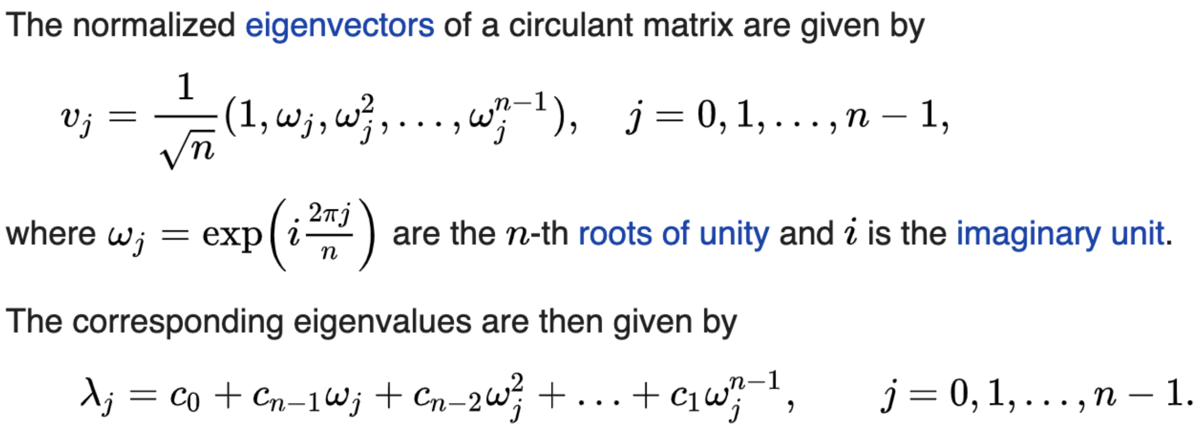
The properties of the circulant matrix:

- Rank:



The rank of a circulant matrix $C$ is equal to $n - d$, where $d$ is the degree of $\gcd(f(x), x^n - 1)$.

- Eigenvectors and Eigenvalues:



The normalized eigenvectors of a circulant matrix are given by

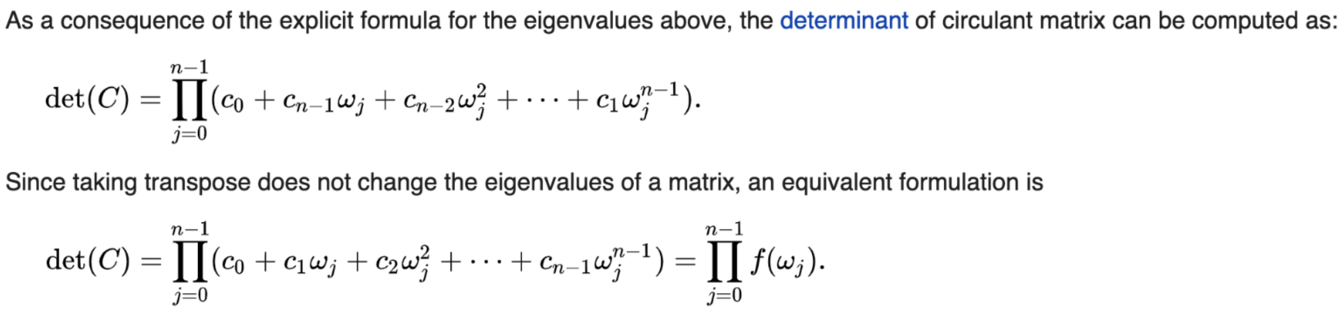
$$v_j = \frac{1}{\sqrt{n}}(1, \omega_j, \omega_j^2, \ldots, \omega_j^{n-1}), \quad j = 0, 1, \ldots, n-1,$$

where $\omega_j = \exp\left(i\frac{2\pi j}{n}\right)$ are the $n$-th roots of unity and $i$ is the imaginary unit.

The corresponding eigenvalues are then given by

$$\lambda_j = c_0 + c_{n-1}\omega_j + c_{n-2}\omega_j^2 + \ldots + c_1\omega_j^{n-1}, \qquad j = 0, 1, \ldots, n-1.$$

- Determinant:



As a consequence of the explicit formula for the eigenvalues above, the determinant of circulant matrix can be computed as:

$$\det(C) = \prod_{j=0}^{n-1}(c_0 + c_{n-1}\omega_j + c_{n-2}\omega_j^2 + \cdots + c_1\omega_j^{n-1}).$$

Since taking transpose does not change the eigenvalues of a matrix, an equivalent formulation is

$$\det(C) = \prod_{j=0}^{n-1}(c_0 + c_1\omega_j + c_2\omega_j^2 + \cdots + c_{n-1}\omega_j^{n-1}) = \prod_{j=0}^{n-1} f(\omega_j).$$

5. What is a Markov Matrix?

Markov matrix is a stochastic matrix showing transition probabilities of a Markov chain. For a right markov matrix sum of each row equals 1 and for a left Markov matrix column summation equals 1.

**Stochastic Processes:**

1. Please explicitly define what a random process and random field means?

Random process: A random process is a time-varying function that is assigned to the outcome of a random experiment.

Random field: Given a probability space, an *X*-valued random field is a collection of *X*-valued random variables indexed by elements in a topological space.

2. Please specify what are the weak and strong stationarity?

Strict, or strong, stationarity means that in a Stochastic Process the probability distribution of the random variable (RV) tossed in each time instant is exactly the same along time, and that the joint probability distribution of RVs in different time instants is invariant to time shifting (this joint probability is usually evaluated with correlation or covariance).

In a weak, or wide-sense, stationary Stochastic Process only the mean and the correlation and covariance of the RV are invariant to time shift (e.g. the variance of the RV eventually changes with time).

3. What is an ergodic process? What is the condition for a process to be ergodic?

A stochastic process is said to be ergodic if its statistical properties can be deduced from a single, sufficiently long, random sample of the process. For example, the process is said to be mean-ergodic if the time average estimate

$$\hat{\mu}_X = \frac{1}{T}\int_0^T X(t)\,dt$$

converges in squared mean to the ensemble average $\mu_X$ as T -> ∞. Likewise, the process is said to be autocovariance-ergodic if the time average estimate

$$\hat{r}_X(\tau) = \frac{1}{T}\int_0^T [X(t+\tau)-\mu_X][X(t)-\mu_X]\,dt$$

converges in squared mean to the ensemble average. A process which is ergodic in the mean and autocovariance is sometimes called ergodic in the wide sense.

4. What is a doubly stochastic process?

A doubly stochastic model is a type of model that can arise in many contexts, but in particular in modelling time-series and stochastic processes. The basic idea for a doubly stochastic model is that an observed random variable is modelled in two stages. In one stage, the distribution of the observed outcome is represented in a fairly standard way using one or more parameters. At a second stage, some of these parameters (often only one) are treated as being themselves random variables.
For example, the observed values in a point process might be modelled as a Poisson process in which the rate (the relevant underlying parameter) is treated as being the exponential of a Gaussian process.

5. What is power spectrum of a stochastic process?

The power spectrum of a stationary stochastic process is the Fourier Transform of it's autocorrelation function:

$$\textit{autocorrelation function}: \quad R(\tau) = \mathbb{E}[X(t+\tau)X^*(t)]$$

$$\textit{power spectrum}: \quad S(\alpha) = \int_{-\infty}^{\infty} R(\tau)e^{-j\alpha\tau}d\tau$$

6. What is the relation between input and output spectrum an and LTI system with stochastic input?

A system is linear when we can write:

$$L[a_1X_1(t)+a_2X_2(t)] = a_1L[X_1(t)]+a_2L[X_2(t)]$$

A system is time invariant if its response to X(t+c) be Y(t+c).

For a linear time invariant system we have:

$$y(t) = (x * h)(t)$$

7. What is the difference between the frequentist and Bayesian point of view?

The essential difference between Bayesian and Frequentist statisticians is in how probability is used. Frequentists use probability *only* to model certain processes broadly described as "sampling." Bayesians use probability more widely to model both sampling and other kinds of uncertainty. Suppose we are interested in the average height *h* in inches of all adult males in a country.

A Bayesian statistician would begin with a "prior distribution," meaning a probability distribution reflecting the state of knowledge about *h* before collecting any data. We do clearly have some prior information: *h* is certainly between 60 and 84 inches, and more likely near the middle of this range. After collecting some data (e.g. a random sample from that country's adult males), the Bayesian would update the prior distribution in light of the data to get a new probability distribution for *h* called the posterior distribution. The posterior distribution reflects our state of knowledge about *h* after collecting data. Using the posterior distribution, the Bayesian can make a statement such as:

$$P(70 \leq h \leq 74) = 95\%$$

Frequentists do not allow themselves to make such statements. For a Frequentist, *h* is simply an unknown constant which either lies in the range [70, 74] or does not. To the Frequentist, the probability statement above is meaningless. Frequentists only allow probability statements about sampling. An example of a legal probability statement for a Frequentist is:

$$P(70 \leq H \leq 74) = 95\%$$

where *H* is a random draw from the population of adult males in the U.S. Frequentist techniques, such as confidence intervals and hypothesis tests, provide ways to make statements that resemble Bayesian probability statements but which only use probability in the Frequentist way.

In terms of parametric distributions p(x,θ), frequentists assume θ is fixed but unknown whereas Bayesian assume θ is random and unknown. Therefore, they assume a prior density for θ and compute the posterior distribution.

**Machine Learning:**

1. What is the difference between transductive and inductive learning?

Inductive learning is traditional supervised learning. We learn a model from labeled examples, and try to predict the labels of examples we have not seen or know about. Transductive learning is less ambitious. We learn on lots of examples, but we only try to predict on a known (test) set of unlabeled examples.