Machine learning with Math

Useop Gim

January 1, 2023

Chapter 1

An introducion to Linear algebra

1.1 Basic

Definition 1. *Group*
Closure, associativity, identity, inverse
If commutativity, abelian

Definition 2. *Ring*
operation(+,$\times$)
Closure,associativity, identity(+), inverse(+), distributivity
If commutativity, commutativity ring

Definition 3. *Filed*
operation + and times
Closure, commutativity, associativity, identity, inverse, distributivity

Definition 4. *Vector Space*
operation + and scalar times
Closure, commutativity, associativity, identity, inverse, distributivity

Definition 5. *inner product*
Linearity
Conjugate symmetry
Positive-definiteness

Definition 6. *Norm*
Triangle inequality
Absolute homogeneity
Positive-definiteness
All norm is seminorm but not all seminorm(nonnegativity) is Norm

Definition 7. *Transformation*
Let V and V' are vector sapce

$$T : V \to V'$$

Linear Transformation is specific transformation which transforms to linear equation

Definition 8. *Matrix*
Matrix is the system for linear polynomial equation in vector space

1.2 Basic for Vector

Definition 9. *Elementary row operation matrix*
1. Bilinear and Scalar multiplication $R_i(x) = xR_i$
2. Swap $R_ij \Leftrightarrow R_i \leftrightarrow R_j$
3. Addition $R_ij(x) = R_i + xR_j$

Theorem 1. *Row Echelon form elimination and Gaussian Jordan elimination*
Both elimination augmentes the matrix
For RREF, the first non-zero row value is "leading variable"
The zero value for the row or without introduced variable is "free variable"

Theorem 2. *linear equation system of vector space*
1. No solution
2. One solution
3. Infinity many solution

Definition 10. *Homogeneous equation*
It is an equation to zero, whose each term contains the function or 'one of its derivatives'.

Definition 11. *Linearly independent*
If the solution of homogeneous equation has only zero for all terms, it is LD
And we call the solution is trivial solution

Definition 12. *Invertible in Matrix*
1. Linear transformation of A is bijective
2. Since LT of A is bijective, RREF is Identity
3. Since RREF no free variable, Elementary row operation matrix times A is identity
4. Since no free variable, $Ax = 0$ has only trivial solution$\Leftrightarrow$ Linearly independent
5. Since no free variable, $Ax = b$ has only one solution
6. $rank(A) = n$
7. $Null(A) = 0$
8. $def(A) \neq 0$
9. $0 \notin eigen(A)$
10. A is non singular

Definition 13. *Transpose*

$$[A^T]_{ij} = A_{ji}$$

Definition 14. *determinant*
Let Mutually Exclusive Collectively Exhaustive by permutation

$$\sum_{i=1} \left(sign(\epsilon_i) \prod_{j=1} [A]_{j\epsilon_i} \right)$$

Definition 15. *Cofactor expansion(Laplace expansion) and Adjugate of matrix*

$$adj(A) = C^T$$

Definition 16. *Orthogonal*
Inner product of two different vector gives zero

Definition 17. *Funder mental subspace*
Subspce : subset for vector space with linearity
Span : the set of vectors which constructs subspace(S) from linear equation system
Basis : the set of vectors in linearly independent span(S)
Dim : the maximal order of linear independent subset $dim(S) = |basis(S)|$
Row(S) : basis for row space of S thus $row(S) = col(S^T)$
Col(S) : basis for col space of S thus $col(S) = row(S^T)$
Null(S) : null space is basis of linearly independent equation
In RREF : row(S) is the leading variables and null is the free variables
Null(S^T) *: WLG same as above*
Nullity(S) : dim(null(A))
rank(S) : The largest dimension of basis of row space $rank(S) = dim(row(S))$
usually rank(S) from row space
Ker(A) : $x : G(x) = e_y$
In matrix linear system $Ker(A) \leftrightarrow Ax = 0$
Therefore $Ker(A) = Null(A)$
Img(A) : $y : G(x) = y$ *Therefore* $Img(A) = col(A)$

Theorem 3. *The relationship between subspace*
$row(S) = col(S^T)$
$row(S) \perp Null(S^T)$
$col(S) = row(S^T)$
$col(S) \perp Null(S^T)$

Theorem 4. *Rank and Nullity theorem*
We know $dim(Img(A)) + dim(Ker(A)) = dim(domain(A))$
Since 1st isomorphism theorem from $V/Ker(A)$ *to* $Img(A)$ *(bij,homo)*

$$rank(A) + Nullity(A) = n$$

Theorem 5. *rank*
We know that $rank(A) + Nullity(A) = n$
And $rank(A) = dim(row(A))$
Also, $Ker(A) = Null(A)$ *and* $Img(A) = col(A)$
$dim(Img(A)) + dim(Ker(A)) = dim(domain(A))$
$\Leftrightarrow dim(Img(A)) + dim(Ker(A)) = n$
$\Leftrightarrow dim(Img(A)) + dim(Null(A)) = n$
$\Leftrightarrow dim(col(A)) + Nullity(A) = n$
$\Leftrightarrow dim(col(A)) + n - rank(A) = n$
$\Leftrightarrow dim(col(A)) = rank(A)$
$\Leftrightarrow dim(col(A)) = dim(row(A))$
it doesn't mean col(A) = row(A) it just indicates the rank (dim)

Definition 18. *Eigen value*
eigen value is obtained from the solution $Ax = \lambda x \rightarrow \lambda I - A = 0, x \neq 0$

Definition 19. *Characteristic equation*

$$\lambda I - A = 0$$

Definition 20. *Characteristic polynomial equation*

$$P_A(\lambda) = det(\lambda I - A)$$

Definition 21. *Eigen vector*
The basis of $(\lambda I - A)x = 0$ *with* $x \neq 0$

Theorem 6. *Eigen space is Singular*
If eigenvector is existed, then

$$det(\lambda I - A) = 0$$

If $P_A(\lambda) = 0$ *then* $(\lambda I - A)x = 0 \Rightarrow x = (\lambda I - A)0$
However, $x \neq 0$ *and so it is contradiction*

Theorem 7. *Zero Eigen value*
If eigen value is zero then $(\lambda I - A)x = 0 \Rightarrow Ax = 0$
Thus it is linearly independent ($x \neq 0$*)*
Hence x be the null of(A)

$$\lambda = 0$$

$$\Rightarrow Ax = 0 \Rightarrow x \in Null(A) \wedge x \neq 0$$

$$\Rightarrow Null(A) \text{ has no trivial solution}$$

$$\Rightarrow rank(A) + Nullity(A) = n \rightarrow rank(A) \in [0, n)$$

Definition 22. *Nilpotent*
The nilpotent matrices satisfies $M^n_{m \times m} = 0$ *for* $\exists n \in \mathbb{N}$

$$Ax = \lambda x \Rightarrow A^2x = A(Ax) = A(\lambda x) = \lambda^2 x$$

Through induction

$$A^n x = \lambda^n x \Rightarrow (A^n - \lambda)x = 0$$

We know that $A^n = 0$
Thus $\lambda = 0$ *Therefore all eigenvalues(*λ*) of it are 0*
Hence $rank(Nilpotent) \in [0, m)$

Definition 23. *Trace*
Trace is the sum of main diagonal(only sauqre) elements

Theorem 8. *Tr(A) and det(A)*
Tr(A) $= \sum_i^n \lambda_i$
Because Characteristic equation is 0

$$tr(\lambda I - A) = \sum_i (\lambda_i - a_{ii})$$

$$0 = \sum_i \lambda_i - tr(A)$$

$$tr(A) = \sum_i \lambda_i$$

det(A) $= \prod_i^n \lambda_i$ *Because Characteristic polynomial equation is zero (non-invertible)*

$$P_A(\lambda) = det(\lambda I - A)$$

Since λ *is root of polynomial* P

$$P_A(\lambda) = \prod_i (\lambda - \lambda_i)$$

Through above two equation

$$det(\lambda I - A) = \prod_i (\lambda - \lambda_i)$$

Then let $\lambda = 0$

$$det(A) = \prod_i (\lambda_i)$$

Definition 24. *Algebraic multiplicity and Geometric multiplicity*
$am(\lambda)$ *is the number of eigenvalue*
$gm(\lambda)$ *is the number of eigenvector*

1.3 Applied Decomposition

Chapter 2

An introduction to RL

2.1 Basic

Theorem 9. *Markov Property*
It is the probability of event which is related with the past events.

$$P(x_t|x_0, x_1, ..x_{t-1}) = P(x_t)$$

The Markov model is based on the above property.

Theorem 10. *Markov Process(Markove Chain)*
In the Markov process is a process which following the Markov Property with tuple (S, P)

Theorem 11. *State transition Probability* P
The probability making transform state is called "state transition probability."

Theorem 12. *Reward* R
The reward in Markov Process is obtained from transition.

Theorem 13. *Markov Decision Process*
The Markov Decision Process represent the Markov Process as tuple.

Theorem 14. *Markov Reward Process*
It adds a reward and a discount factor element in process (S, P, R, γ)

Theorem 15. *Markov Decision Process*
It adds an action element in process (S, A, P, R, γ)

Theorem 16. *Policy* π
From the Markov Process, policy is the distribution of all action by current state.

$$\pi(a|s) = P\left[A_t = a|S_t\right]$$

Theorem 17. *State-value*
Let G is total future reward
The state-value is the value of current state

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

Theorem 18. *Action-value*
The action-value is the value of action for current state

$$q(s,a) = \mathbb{E}\left[G_t | S_t = s, A_t = a\right]$$

Theorem 19. *Bellman Expected equation*

$$v_\pi(s) = \mathbb{E}\left[R_{t+1} + \gamma v(s_{t+1}) | s_t\right]$$
$$q_\pi(s,a) = \mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) | s_t, a_t\right]$$

Theorem 20. *Bellman Optimality equation*

$$v_*(s) = max\mathbb{E}\left[R_{t+1} + \gamma v(s_{t+1}) | s_t\right]$$
$$q_*(s,a) = max\mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) | s_t, a_t\right]$$

Theorem 21. *Transition by policy*
The probability of transition from state s_t to s_{t+1} with policy

$$P^\pi(s_{t+1}|s_t) = \sum_{a \in A_t} \pi(a|s_t) p(s_{t+1}|s_t, a) r(s,a)$$

Theorem 22. *Reward by policy*
The probability of transition from state s_t to s_{t+1} with policy

$$R^\pi(s_t, a) = \sum_{a \in A_t} \pi(a|s_t) r(s_t, a)$$

2.2 Policy Gradient Theorem

Theorem 23. *State-value by policy*
First convert State-value function as policy form

$$\begin{aligned} v_\pi(s_t) &= \mathbb{E}_\pi\left[R_{t+1} + \gamma v(s_{t+1})\right] \\ &= R^\pi_{s_t} + \gamma \sum_{s_t} v(s_{t+1}) \\ &= \sum_{a \in A} \pi(a|s_t) r(s_t, a) + \gamma \sum_{s_{t+1} \in S} \sum_{a \in A_t} \pi(a|s_t) p(s_{t+1}|s_t, a) v_\pi(s_{t+1}) \end{aligned}$$

Through the summation property

$$v_\pi(s_t) = \sum_{a \in A_t} \pi(a|s_t) r(s_t, a) + \gamma \sum_{a \in A_t} \pi(a|s_t) \sum p(s_{t+1}|s_t, a) v_\pi(s_{t+1})$$

Theorem 24. *Action-value by policy*

$$
\begin{aligned}
q_\pi(s_t, a) &= \mathbb{E}\left[R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) | s_t, a_t\right] \\
&= r(s, a) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1} | s, a) \sum_{a' \in A} \pi(a' | s_{t+1}) q_\pi(s_{t+1}, a')
\end{aligned}
$$

Theorem 25. *State-value and Action-value relationship*

$$
v_\pi(s_t) = \sum_{a_t \in A_t} \pi(a_t | s_t) q_\pi(s_t, a_t)
$$

$$
q_\pi(s_t, a) = r(s_t, a) + v_\pi(s_t)
$$

Theorem 26. *Policy Gradient Theorem*
For the stochastic policy π

$$
\begin{aligned}
\nabla v_\pi &= \nabla \left(\sum_a \pi q \right) \\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \sum \pi(a|s) \nabla q_\pi \right) \\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s) \sum_{s'} \nabla \left(\sum_{s'} P^a_{s's} \cdot q_\pi(s', a') \right) \right) \\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s) \sum_{s'} \nabla \left(\sum_{s'} P^a_{s's} \cdot (r + v_\pi(s')) \right) \right) \\
&= \sum_a \left(\nabla \pi(a|s) q_\pi + \pi(a|s) \sum_{s'} P^a_{s's} \cdot \nabla v_\pi(s') \right)
\end{aligned}
$$

Therefore

$$
v_\pi(s) = \sum_a \left(\nabla \pi(a|s) q_\pi(s, a) + \pi(a|s) \sum_{s'} P^a_{s's} \cdot \nabla v_\pi(s') \right)
$$

Then the $v_\pi(s,a)$ *is repeated Thus let* $\phi(s) = \sum_a \nabla\pi(a|s) q_\pi(s,a)$

$$
\begin{aligned}
\nabla v_\pi(s) &= \phi(s) + \sum_a \left(\pi(a|s) \sum_{s'} P^a_{s's} \cdot \nabla v_\pi(s') \right) \\
&= \phi(s) + \sum_a \sum_{s'} \pi(a|s) P^a_{s's} \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_a \pi(a|s) P^a_{s's} \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s \to s', k) \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s \to s', k) \cdot (\phi(s') + ...) \\
&= \phi(s) + \sum_x \sum_{k=0} p_\pi(s \to x, k) \cdot \phi(x)
\end{aligned}
$$

Andso let $\sum_{k=0} p_\pi(s \to x, k)$ *as* $\eta(s)$

$$
\begin{aligned}
\nabla v_\pi &= \sum_s \eta(s)\phi(s) \\
&= \left(\sum_s \eta(s) \right) \sum \frac{\eta(s)}{\sum \eta(s)} \phi(s) \\
&= \sum_s d_\pi(s)\phi(s) && \textit{Since } \textstyle\sum \eta \textit{ is constant} \quad d \textit{ is the stationary distribution} \\
&= \sum_s d_\pi(s) \sum_a \nabla\pi(a|s) q_\pi(s,a)
\end{aligned}
$$

Hence the derivative of the expected value function is obateind from the gradient of policy without taking derivative for the reward function

Chapter 3

An introducion to Data

3.1 Basic

Definition 25. *DP, MC, TD*
DP is dynamic programming, it uses the model base
MC is Monte Carlo, it uses bunch of samples then estimates the probability (ex calculate circle of pi)
TD is Temporal Difference, it uses the difference of the transitional value from one step behind state

Definition 26. *Exploitation vs Exploration*
Exploitation is deciding the best action through the using the given samples
Exploration is collecting samples

Definition 27. *TP, TN, FP, FN, SE, SP, FPR, ROC, AUC*
TP is true positive
TN is true negative
FP is false postive (type 1 error)
FN is false negative (type 2 error)
SE is Sensitivity $\frac{TP}{TP+FN}$ *which is the rate of correct positive*
SP is Specificity $\frac{TN}{TN+FP}$ *which is the rate of correct negative*
FPR is False Positive Rate which is the rate of wrong positive
ROC is the the curve by vertical SE horizontal FPR or (1-SP)
AUC is the area of under ROC