

Machine learning with Math

Useop Gim

January 2, 2023

Chapter 1

An introduction 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. *Field*

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 space

$$T : V \rightarrow V'$$

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 $R_i \leftrightarrow R_j$
3. Addition $R_i(x) = R_i + xR_j$

Theorem 1. Row Echelon form elimination and Gaussian Jordan elimination

Both elimination augments 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. $\text{rank}(A) = n$
7. $\text{Null}(A) = 0$
8. $\det(A) \neq 0$
9. $0 \notin \text{eigen}(A)$
10. A is non singular

Definition 13. *Transpose*

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

Definition 14. *determinant*

Let Mutually Exclusive Collectively Exhaustive by permutation

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

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

$$\text{adj}(A) = C^T$$

Definition 16. *Orthogonal*

Inner product of two different vector gives zero

Definition 17. *Funder mental subspace*

Subspace : 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 $\dim(S) = |\text{basis}(S)|$

Row(S) : basis for row space of S thus $\text{row}(S) = \text{col}(S^T)$

Col(S) : basis for col space of S thus $\text{col}(S) = \text{row}(S^T)$

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) : $\dim(\text{null}(A))$

rank(S) : The largest dimension of basis of row space $\text{rank}(S) = \dim(\text{row}(S))$ usually $\text{rank}(S)$ from row space

Ker(A) : $x : G(x) = e_y$

In matrix linear system $\text{Ker}(A) \leftrightarrow Ax = 0$

Therefore $\text{Ker}(A) = \text{Null}(A)$

Img(A) : $y : G(x) = y$ Therefore $\text{Img}(A) = \text{col}(A)$

Theorem 3. *The relationship between subspace*

$$\text{row}(S) = \text{col}(S^T)$$

$$\text{row}(S) \perp \text{Null}(S^T)$$

$$\text{col}(S) = \text{row}(S^T)$$

$$\text{col}(S) \perp \text{Null}(S^T)$$

Theorem 4. *Rank and Nullity theorem*

We know $\dim(\text{Img}(A)) + \dim(\text{Ker}(A)) = \dim(\text{domain}(A))$

Since 1st isomorphism theorem from $V/\text{Ker}(A)$ to $\text{Img}(A)$ (bij,homo)

$$\text{rank}(A) + \text{Nullity}(A) = n$$

Theorem 5. rank

We know that $\text{rank}(A) + \text{Nullity}(A) = n$

And $\text{rank}(A) = \dim(\text{row}(A))$

Also, $\text{Ker}(A) = \text{Null}(A)$ and $\text{Img}(A) = \text{col}(A)$

$\dim(\text{Img}(A)) + \dim(\text{Ker}(A)) = \dim(\text{domain}(A))$

$\Leftrightarrow \dim(\text{Img}(A)) + \dim(\text{Ker}(A)) = n$

$\Leftrightarrow \dim(\text{Img}(A)) + \dim(\text{Null}(A)) = n$

$\Leftrightarrow \dim(\text{col}(A)) + \text{Nullity}(A) = n$

$\Leftrightarrow \dim(\text{col}(A)) + n - \text{rank}(A) = n$

$\Leftrightarrow \dim(\text{col}(A)) = \text{rank}(A)$

$\Leftrightarrow \dim(\text{col}(A)) = \dim(\text{row}(A))$

it doesn't mean $\text{col}(A) = \text{row}(A)$ 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)

$$\lambda = 0$$

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

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

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

Definition 22. Nilpotent

The nilpotent matrices satisfies $M_{m \times m}^n = 0$ for $\exists n \in \mathbb{N}$

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

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 $\text{rank}(\text{Nilpotent}) \in [0, m)$

Definition 23. Trace

Trace is the sum of main diagonal(only saugre) elements

Theorem 8. $\text{Tr}(A)$ and $\det(A)$

$$\text{Tr}(A) = \sum_i^n \lambda_i$$

Because Characteristic equation is 0

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

$$0 = \sum_i \lambda_i - \text{tr}(A)$$

$$\text{tr}(A) = \sum_i \lambda_i$$

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

$\text{am}(\lambda)$ is the number of eigenvalue

$\text{gm}(\lambda)$ is the number of eigenvector

1.3 Applied Decomposition

Chapter 2

An introduction to ML

Theorem 9. *Topological meaning of the neural network*
Vector space, Hyperplane, optimization, statical

Chapter 3

Introduction of Probability

Definition 25. *Probability Density function*

When $F_x(x) = \int_a^b p(x)dx$

$$p(x) = \frac{dF_x(x)}{dx}$$

it is continuous

Definition 26. *Probability Mass function*

$$p(x) = P(X = x)$$

it is discrete

Definition 27. *Cumulative Distribution function*

$$F_X(x) = P\{X \leq x\}$$

There are two type of cdf, discrete and continuous

Definition 28. *Probability mass function*

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(X \leq b) - F_X(X \leq a) = \int_a^b p_X(x)dx$$

Definition 29. *Joint Cumulative Distribution function*

$$F_{XY}(x, y) = P(X \leq x) \cap (Y \leq y)$$

Thus for PDF,

$$F_{XY}(x, y) = \int_{-\infty}^y \int_{-\infty}^x p_{XY}(u, v) du dv = \int_{-\infty}^y \int_{-\infty}^x p_{XY}(u, v) dv du$$

Definition 30. *Marginal density function*

From the JCDF, we can obtained the pdf for X and Y from below

$$p_x(x) = \int_{-\infty}^{\infty} p_{XY}(x, y) dy$$

$$p_Y(y) = \int_{-\infty}^{\infty} p_{XY}(x, y) dx$$

Definition 31. *Conditional Probability*

$$P\{A|B\} = \frac{P\{A, B\}}{P\{B\}}$$

Definition 32. *Independence*

$$P(A, B) = P(A)P(B)$$

Definition 33. *Bayesian Probability*

$$P\{A|B\} = \frac{P\{B|A\}P\{A\}}{P\{B|A\}P\{A\} + P\{B|A^c\}P\{A^c\}}$$

In pdf with joint probability

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\int_{-\infty}^{\infty} p_{X|Y}(x|y)p_Y(y)dy}$$

Definition 34. *Independent and identically distributed*

IID means each sample is independent with uniformly picked

Definition 35. *Dirac delta function*

$$\int_{-\infty}^{\infty} \delta(x)f(x)dx = f(0)$$

$$\delta(x) = \begin{cases} \infty & (x = 0) \\ 0 & (x \neq 0) \end{cases}$$

Definition 36. *Expectation*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp_X(x)dx$$

For random variable

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} g(x)p_X(x)dx$$

For joint

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)p_X(x, y)dxdy$$

Definition 37. *Variance*

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Definition 38. *Covariance*

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Definition 39. *Correlation*

$$Cor(X, Y) = \mathbb{E}[XY] = \int_{-\infty}^{\infty} xyp_{XY}(x, y)dxdy$$

Definition 40. *Correlation Coefficient*

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

It is measurement of how likelihood $(X - \mathbb{E}[X])$ and $(Y - \mathbb{E}[Y])$

Chapter 4

An introduction to RL

4.1 Basic

Theorem 10. *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 11. *Markov Process(Markove Chain)*

In the Markov process is a process which following the Markov Property with tuple (S, P)

Theorem 12. *State transition Probability P*

The probability making transform state is called "state transition probability."

Theorem 13. *Reward R*

The reward in Markov Process is obtained from transition.

Theorem 14. *Markov Decision Process*

The Markov Decision Process represent the Markov Process as tuple.

Theorem 15. *Markov Reward Process*

It adds a reward and a discount factor element in process (S, P, R, γ)

Theorem 16. *Markov Decision Process*

It adds an action element in process (S, A, P, R, γ)

Theorem 17. *Policy π*

From the Markov Process, policy is the distribution of all action by current state.

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

Theorem 18. State-value

Let G is total future reward

The state-value is the value of current state

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

Theorem 19. Action-value

The action-value is the value of action for current state

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

Theorem 20. Bellman Expected equation

$$v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v(s_{t+1}) | s_t]$$

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

Theorem 21. Bellman Optimality equation

$$v_*(s) = \max \mathbb{E}[R_{t+1} + \gamma v(s_{t+1}) | s_t]$$

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

Theorem 22. 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 23. Reward by policy

The probability of transition from state s_t to s_{t+1} with policy

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

4.2 Policy Gradient Theorem

Theorem 24. State-value by policy

First convert State-value function as policy form

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

Through the summation property

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

Theorem 25. *Action-value by policy*

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

Theorem 26. *State-value and Action-value relationship*

$$\begin{aligned} v_\pi(s_t) &= \sum_{a_t \in A_t} \pi(a_t | s_t) q_\pi(s_t, a_t) \\ q_\pi(s_t, a) &= r(s_t, a) + v_\pi(s_t) \end{aligned}$$

Theorem 27. *Policy Gradient Theorem*

For the stochastic policy π

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

Therefore

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

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

$$\begin{aligned}
\nabla v_\pi(s) &= \phi(s) + \sum_a \left(\pi(a|s) \sum_{s'} P_{s's}^a \cdot \nabla v_\pi(s') \right) \\
&= \phi(s) + \sum_a \sum_{s'} \pi(a|s) P_{s's}^a \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_a \pi(a|s) P_{s's}^a \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s \rightarrow s', k) \cdot \nabla v_\pi(s') \\
&= \phi(s) + \sum_{s'} \sum_k p_\pi(s \rightarrow s', k) \cdot (\phi(s') + \dots) \\
&= \phi(s) + \sum_{\mathbf{x}} \sum_{k=0} p_\pi(s \rightarrow \mathbf{x}, k) \cdot \phi(\mathbf{x})
\end{aligned}$$

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

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

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

Chapter 5

An introduction to Data

5.1 Basic

Definition 41. *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 42. *Exploitation vs Exploration*

Exploitation is deciding the best action through the using the given samples

Exploration is collecting samples

Definition 43. *TP, TN, FP, FN, SE, SP, FPR, ROC, AUC*

TP is true positive

TN is true negative

FP is false positive (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