introduction to machine learning

DR.Amiri

electrical engineering department

Ahmadreza Majlesara 400101861

assignment 1

July 11, 2024

introduction to machine learning

assignment 1

Ahmadreza Majlesara 400101861

All You Need is Combinatorics

In this problem, you are given an *Artificial Machine(!)* and you are asked to teach it in a way that in the end, it would Learn *Intelligence(?)*. So basically, by solving this problem completely, you will know everything needed in this course.

A simple Random Walk

First, suppose that our Machine shows an integer at each time and it starts with 0 at the first round. On each round after that, the Machine randomly and with equal probability, either increases its number by 1 or decrease it by 1 (it can show negative integers too). Calculate the probability of our Machine showing 0 on its monitor again, if we know that it will work for exactly T rounds. (There is no need for your answer to have a closed and simple form and just the solution is important for this part)

solution

We employ the concept of the Khayyam-Pascal triangle to address this problem. this triangle is shown in the figure below.

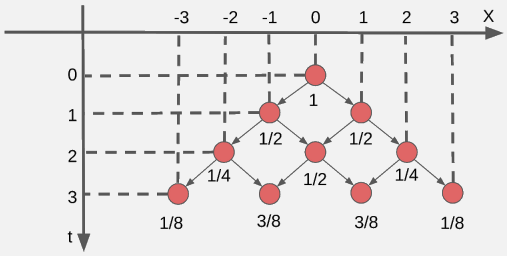

Figure 1. Khayyam-Pascal triangle

For each time step T, if T is odd, the probability of obtaining zero at that time step is 0. However, if T is even and we desire the result to be 0, the counts of 1's and -1's must be equal. Hence, the number of instances where our machine displays 0 on its monitor again is $\binom{T}{\frac{T}{2}}$. This suggests that the number of cases is $\frac{T!}{\frac{T}{2}!\frac{T}{2}!}$. The total number of possible cases is $\Sigma_{i=0}^{T}\binom{T}{i} = 2^T$, so the probability of our machine displaying 0 on its monitor is $\frac{T!}{\frac{T}{2}!\frac{T}{2}!2^T}$.

The key trick here is that if we want to compute the probability of having 0 **until** the time step T, we should not expand nodes of 0 in previous rows of the triangle. This is because in this case we count a stream like $[-1,1,-1,1]$ as one successful case and not two. So, at each time step, we first calculate all successful cases, then we subtract the cases that arise from expanding nodes of 0 in previous rows of the triangle. We also note that for odd T, the number of successful cases is the same as the number of successful cases in $T-1$.
First, we define a function called $N(T)$. This function takes the time step as input and outputs the number of successful cases (subtracting the cases that arise from expanding nodes of 0 in previous rows of the triangle). Then, the probability of having 0 until that time step is $\frac{N(T)}{2^T}$. To find the total number of successful cases, we use the formula mentioned earlier. For subtraction, we utilize the fact that the number of successful cases that arise from expanding nodes of 0 in previous rows (for example, T_1) of the triangle equals $N(T_1)\binom{T-T_1}{\frac{T-T_1}{2}}$. This is because we can choose the number of 1's and -1's in the remaining time steps in $\binom{T-T_1}{\frac{T-T_1}{2}}$ ways and the number of successful cases in T_1 is $N(T_1)$. Finally, we add the probabilities of all even T's until time step T to find the probability.
Now, we can calculate the number of successful cases in T using this formula: *(X is the random variable representing the number of times the machine shows 0 on its monitor until time step T)*

$$N(2)=2 \Rightarrow\ P(X=0|T=2)=\frac{2}{4}=\frac{1}{2}$$

$$N(4)=\binom{4}{2}-N(2)\binom{2}{1}=6-4=2 \Rightarrow\ P(X=0|T=4)=\frac{2}{16}+\frac{1}{2}=\frac{5}{8}$$

$$N(6)=\binom{6}{3}-N(2)\binom{4}{2}-N(4)\binom{2}{1}=20-12-4=4$$

$$\Rightarrow\ P(X=0|T=6)=\frac{4}{64}+\frac{5}{8}=\frac{11}{16}$$

$$N(8)=\binom{8}{4}-N(2)\binom{6}{3}-N(4)\binom{4}{2}-N(6)\binom{2}{1}=70-40-12-8=10$$

$$\Rightarrow\ P(X=0|T=8)=\frac{10}{256}+\frac{11}{16}=\frac{93}{128}$$

now we use another way to get a closed form. **this proof is inspired by proof on catalan numbers**.

if we define a random walk of length N as a sequence $w=\{\varepsilon_n\}_{n=1}^{N}$ elementary steps $\varepsilon \in \{+1,-1\}$ such that $S_n(w)=\sum_{i=1}^{n}\varepsilon_i$ and $S_0=0$ we can introduce the following terminology for certain special types of walks:

- **Balanced walk**: A walk of length $2N$ such that $S_N(w) = 0$. *it contains the same number of up-steps as down-steps*
- **non-negetive walk**: if $S_n(w) \geq 0$ for $1 \leq n \leq N$. *it never goes below the initial position*
- **non-zero walk**: if $S_n(w) \neq 0$ for $1 \leq n \leq N$. *it never returns to the initial position A non-zero walk is either positive or negative depending on whether* $S_n(w) > 0$ *or* $S_n(w) < 0$ *for all* $n > 1$

if we define $A_n = \binom{2n}{n}$ we can proof that this is the number of :

- balanced walks of length $2n$
- non-negative walks of length $2n$
- non-zero walks of length $2n$

we proved that the number of balanced walks of length $2n$ is A_n in previous section. for non-negetive walk of length $2n$ we can proof that the number of this kind of walks with m up-steps and k down-steps where $k \leq m$ is:

$$B(m,k) = \binom{m+k}{k} - \binom{m+k}{k-1}$$

now we can easily count all non-negetive walks of length 2n obtaining:

$$1 + \sum_{k=1}^{n} B(2n-k,k) = 1 + \sum_{k=1}^{n} \binom{2n}{k} - \binom{2n}{k-1} =$$

$$1 + \binom{2n}{1} - \binom{2n}{0} + \binom{2n}{2} - \binom{2n}{1} + \cdots + \binom{2n}{n} - \binom{2n}{n-1} = \binom{2n}{n} = A_n$$

Hence the number of non-negetive walks of length $2n$ is also A_n. Furthermore, the number of positive walks of length $2n$ is the same as the number of non-negetive walks of length $2n-1$ wich is:

$$1 + \sum_{k=1}^{n-1} B(2n-1-k,k) = \binom{2n-1}{n} = \frac{1}{2}A_n$$

by symmetry the number of negative walks of length $2n$ is also $\frac{1}{2}A_n$. so the number of non-zero walks of length $2n$ is A_n. so the probability of not crossing zero in a random walk of length $2n$ is $\frac{A_n}{2^2 n}$. in our problem we can state that the number of non crossing zero walks of length T is

$$P(X \neq 0 | T = T) = \frac{\binom{T}{\lfloor \frac{T}{2} \rfloor}}{2^T}$$

so the probability of crossing zero (wich is the answer to our question) is:

$$P(X = 0 | T = T) = 1 - \frac{\binom{T}{\lfloor \frac{T}{2} \rfloor}}{2^T}$$

Everything is possible!

According to the previous part, show that if T goes to ∞, then the probability of seeing 0 again will converge to 1. Also with a similar reasoning show that we will see every other number at least once too(for this, no calculations is needed).

solution

fo this question we use stirling's approximation. this approximation states that:

stirling's approximation:

for large n:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

in the previous part we showed that the probability of seeing 0 again after T steps is:

$$P(X=0|T=T) = 1 - \frac{\binom{T}{\frac{T}{2}}}{2^T}$$

so we have:

$$P(X=0|T=T) = 1 - \frac{T!}{\frac{T}{2}!\frac{T}{2}!2^T}$$

$$\Rightarrow \lim_{T\to\infty} P(X=0|T=T) = 1 - \frac{\sqrt{2\pi T}\left(\frac{T}{e}\right)^T}{(\sqrt{2\pi\frac{T}{2}}\left(\frac{T}{2e}\right)^{\frac{T}{2}})^2 2^T}$$

$$\Rightarrow \lim_{T\to\infty} P(X=0|T=T) = 1 - \frac{\sqrt{2\pi T}\left(\frac{T}{e}\right)^T}{\pi T\left(\frac{T}{2e}\right)^T 2^T}$$

$$\Rightarrow \lim_{T\to\infty} P(X=0|T=T) = 1 - \frac{\sqrt{2\pi T}\left(\frac{T}{e}\right)^T}{\pi T\left(\frac{T}{e}\right)^T}$$

$$\Rightarrow \lim_{T\to\infty} P(X=0|T=T) = 1 - \frac{\sqrt{2\pi T}}{\pi T}$$

$$\Rightarrow \lim_{T\to\infty} P(X=0|T=T) = 1$$

we can use a similar reasoning to show that the probability of seeing every other number at least once is 1. this is because the number of times we cross a number is $\binom{T}{\frac{T+k}{2}}$ *for each* $(k \leq T)$ and the probability of crossing a number is:

$$P(X = k|T = T) = 1 - \frac{\binom{T}{\frac{T+k}{2}}}{2^T}$$

so like the proof for 0 we can show that the probability of crossing a number converges to 1 as T goes to ∞.

A naive algorithm

Now, we want to teach the first Learning algorithm to our Machine. For this, suppose that we have n secrets which are either 0 or 1 with equal probability. In other words, our whole secret is a random vector of 0 and 1's in $\{0,1\}^n$, which the Machine isn't aware of. At each round, Machine guesses a random vector of the same form totally random(with same probability for each vector), and it will stop whenever its guess completely match the secret. Calculate the expected time of success for our algorithm and based on that, explain why in practice, using such an algorithm isn't helpful at all.

solution

if we have a random vector of 0 and 1's of length n and our algorithm guesses a random vector of the same form at each round, the probability of success at each round is $\frac{1}{2^n}$.we can model this as a geometric distribution with parameter $p = \frac{1}{2^n}$. the expected number of trials until the first success is $\frac{1}{p} = 2^n$. so the expected time of success for our algorithm is 2^n. as n grows, the expected time of success grows exponentially. so in practice, using such an algorithm isn't helpful at all.

A less naive algorithm

Now, suppose that our hidden secret is a Random Permutation of 1, 2, . . . , n. At each round, Machine guesses the first unrevealed position of this permutation and when it guess becomes true, it will move to the next position. Algorithm will end whenever Machine guesses all of the secret. The algorithm which Machine makes guess at each round is as follows: for each position of permutation, Machine start guessing from the least unrevealed number, and after each wrong guess, guesses the next least unrevealed number. Prove that this algorithm has the best performance, in the means of expected number of time for success. Also, calculate the expected number of times this Machine will guess before success.

solution

we suppose that the random variables $X_1, X_2, \ldots X_n$ represents the number of times the Machine will guess number of an array before success. we also suppose that the random variable S represents the number of times the Machine will guess before success. we can state that:

$$S = \sum_{i=1}^{n} X_i$$

now we need to calculate $E[X_i]$ for each i. we can state that the probability of success at each round is $\frac{1}{n-i+1}$. we define E_{ij} as the expected value of tries for X_i when we had j tries before. we can state that:

$$E_{ij} = \frac{1}{n-i+1} + \frac{n-i+1-j}{n-i+2-j}(E_{i,j+1}+1)$$

by calculating this recursive function can say that $E[X_i] = \frac{n-i}{2} + 1$ so now we have:

$$E[S] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n-i}{2} + 1 = \frac{n(n+3)}{4}$$

Gaussians everywhere

In this part, assume that our secret is a random vector from Normal, n-dimensional distribution. It means that each of it's entries are from Normal distribution(Gaussian with Mean 0 and Variance 1). Now, our algorithm is upgraded so much and it works like this: It starts from an arbitrary vector. At each round, it randomly selects an index. After that, it samples a random number from Normal distribution like x. Then it either adds or removes x from the selected index, and it chooses the one which is closer to the secret in that index(assume that somehow we can know which one is closer). The algorithm will stop whenever at each index, difference between algorithm's number and secret, is less than some fixed t. First show that the best starting point for the algorithm is 0 n vector, then calculate the expected finishing time of algorithm(there is no need to explicit calculation for the expected time and only a sketch of the proof suffices).

solution

first we assume that for each index i the random variable X_i represents the number of times the Machine will guess number of an array before success. if the first value of our algorithm is b we can model this as the gaussian distribution of our samples is shifted by t so it would be $\mathcal{N}(b,1)$. we want our end point to be within the range of t from the secret so we can state that the probability of success at each round is $P(secret[i]-t \leq X_i \leq secret[i]+t) =$

$$\int_{secret[i]-t}^{secret[i]+t} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-b)^2}{2}} dx$$

if we fix all parameters except b and we want to maximize this probability we can differentiate this probability with respect to b and set it to zero. so we have:

$$\frac{d}{db}\int_{secret[i]-t}^{secret[i]+t}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-b)^2}{2}}dx=0$$

$$\Rightarrow\int_{secret[i]-t}^{secret[i]+t}\frac{1}{\sqrt{2\pi}}\frac{2(x-b)}{2\sigma^2}e^{-\frac{(x-b)^2}{2}}dx=0$$

and this is true if $b=0$. so the best starting point for the algorithm is the 0 vector.

Statistics and other friends

LLN and CLT

Briefly explain *Law of Large Numbers* and *Central Limit Theorem*.

solution

Law of Large Numbers (LLN): This law states that as the number of trials increases, the experimental probability of an event approaches the theoretical probability of the event. In other words, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.
Central Limit Theorem (CLT): This theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed ($i.i.d$) random variables approaches a normal distribution, regardless of the shape of the original distribution. This is true even if the original random variables are not normally distributed.

Assumption

Discuss how this two theorems and their implications are related to Statistics. How do you think they are going to be used in the course?

solution

These two theorems are fundamental in statistics. The Law of Large Numbers is used to justify the use of sample means as estimates of population means. It states that as the sample size increases, the sample mean will converge to the population mean. The Central Limit Theorem is used to justify the use of the normal distribution as an approximation to the sampling distribution of the sample mean. It states that the sampling distribution of the sample mean will be approximately normally distributed, regardless of the shape of the population distribution, as long as the sample size is sufficiently large. I think These theorems are used in the course to justify the use of statistical methods and to understand the properties of estimators and test statistics.

Are you sure about that?

Briefly explain Hypothesis Test and Confidence Interval.

solution

Hypothesis Test: A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population. The test is based on the assumption that the null hypothesis is true, and it provides a way to determine whether the null hypothesis should be rejected in favor of an alternative hypothesis.
Confidence Interval: A confidence interval is a range of values that is used to estimate the true value of a population parameter. It is based on the assumption that the sample mean is normally distributed, and it provides a way to estimate the population mean with a certain level of confidence.

Another Assumption

Discuss how this two concepts and their implications are related to Statistics. How do you think they are going to be used in the course?

solution

Hypothesis tests and confidence intervals are used to make inferences about population parameters based on sample data. Hypothesis tests are used to determine whether there is enough evidence to reject the null hypothesis in favor of an alternative hypothesis, while confidence intervals are used to estimate the population parameter with a certain level of confidence. I think these concepts are going to be used in the course to understand the properties of estimators and test statistics, and to make inferences about population parameters based on sample data.

Time to take out your pen!

Consider $X_1, X_2, ..., X_n$ as n independent random variables, having the same distribution as random variable X from [0, 1] interval. Also consider $Y_1, Y_2, ..., Y_n$ as Bernouli random variables independent from each other and also independent from $X_1, X_2, ..., X_n$, each with parameter $X_1, X_2, ..., X_n$ respectively. You are given the values of $Y_1, Y_2, ..., Y_n$, also we don't know the values of $X_1, X_2, ..., X_n$. Based on this, create a 95% confidence interval for $\mu = E[X]$.

solution

we know that for a Bernouli random variable with parameter X , $E[Y] = X$ and $Var[Y] = (1-X)X$. we define a random variable called $Z = \frac{\sum_{i=1}^{n} Y_i}{n}$ for this random variable we can state that $E[Z] = \frac{\sum_{i=1}^{n} E[Y_i]}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$. also we can state that $Var[Z] = \frac{\sum_{i=1}^{n} Var[Y_i]}{n} = \frac{\sum_{i=1}^{n} (1-X_i)X_i}{n}$.

as we dont know the values of X_i we cant have a specific value for $Var[Z]$. but we want to specify an interval for $\mu = E[X]$ we can use this approach:

$$(1-X_i)X_i = X_i - X_i^2 \Rightarrow \frac{d}{dX_i}(X_i - X_i^2) = 1 - 2X_i = 0 \Rightarrow X_{i_{max}} = \frac{1}{2}$$

$$\Rightarrow (1-X_i)X_i \leq \frac{1}{4} \Rightarrow Var[Z] \leq \frac{1}{4n}$$

we can state that if we find an interval with an assumption that each Y_i has the maximum variance this interval is true for all variances. now using CLT we have:

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\mu - \bar{X}}{\sqrt{\frac{1}{4n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

if we solve this inequality for μ and set $\alpha = 5$ we have:

$$P(\bar{X} - z_{\frac{5}{2}}\sqrt{\frac{1}{4n}} \leq \mu \leq \bar{X} + z_{\frac{5}{2}}\sqrt{\frac{1}{4n}}) = 0.95$$

so the interval for the maximum variance is:

$$[\bar{X} - z_{\frac{5}{2}}\sqrt{\frac{1}{4n}}, \bar{X} + z_{\frac{5}{2}}\sqrt{\frac{1}{4n}}]$$

this is an interval with a confidence level of a more than 95% for X.

Its all about Tails

Not a very hard inequality

Consider random variable X, with $\mathbb{E}[X] = 0$ and $Var[X] = \sigma^2$. show that for each a > 0, we have:

$$P[X \geq a] \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

solution

we can use markov's inequality to solve this question. this inequality states that:

markov's inequality:
If X is a non-negative random variable and $a > 0$, then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

using this inequality we have:

$$\mathbb{P}(X \geq a) = \mathbb{P}(X + \frac{\sigma^2}{a} \geq a + \frac{\sigma^2}{a}) = \mathbb{P}((X + \frac{\sigma^2}{a})^2 \geq (a + \frac{\sigma^2}{a})^2) \leq \frac{\mathbb{E}[(X + \frac{\sigma^2}{a})^2]}{(a + \frac{\sigma^2}{a})^2}$$

we also know that:

$$\mathbb{E}[(X + \frac{\sigma^2}{a})^2] = Var(X + \frac{\sigma^2}{a}) + \mathbb{E}[X + \frac{\sigma^2}{a}]^2 = \sigma^2 + \frac{\sigma^4}{a^2}$$

so we have:

$$\mathbb{P}(X \geq a) \leq \frac{\sigma^2 + \frac{\sigma^4}{a^2}}{(a + \frac{\sigma^2}{a})^2} = \frac{\frac{\sigma^2(a^2+\sigma^2)}{a^2}}{\frac{(a^2+\sigma^2)^2}{a^2}} = \frac{\sigma^2}{a^2 + \sigma^2}$$

Gaussian

Show that for $X \sim \mathcal{N}(0, \sigma 2)$ and for each $s \in \mathbb{R}$, we have:

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}$$

solution

$$\mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_x(X) dx = \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{sx - \frac{x^2}{2\sigma^2}} dx$$

we can complete the square in the exponent:

$$sx - \frac{x^2}{2\sigma^2} = -(\frac{x^2}{2\sigma^2} - sx + (\frac{\sqrt{2}\sigma s}{2})^2) + (\frac{\sigma s}{\sqrt{2}})^2$$

so we have:

$$\mathbb{E}[e^{sX}] = \frac{e^{\frac{\sigma^2 s^2}{2}}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(\frac{x}{\sqrt{2}\sigma} - \frac{\sigma s}{\sqrt{2}})^2} dx$$

we define $u = \frac{x}{\sqrt{2}\sigma} - \frac{\sigma s}{\sqrt{2}}$ so $du = \frac{dx}{\sqrt{2}\sigma}$ and we have:

$$\mathbb{E}[e^{sX}] = \frac{e^{\frac{\sigma^2 s^2}{2}}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-u^2} \sqrt{2}\sigma du = e^{\frac{\sigma^2 s^2}{2}}$$

this way we can proof that $\mathbb{E}[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}$

Under the Gaussian!

Show that, if X is a random variable which has the property of last part with parameter σ^2 (note that X is not necessarily Gaussian), then for each $t > 0$ we have:

$$\mathbb{P}[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

solution

we can state that $\mathbb{P}[|X| \geq t] = \mathbb{P}[X \geq t] + \mathbb{P}[X \leq -t]$. now we can use chernoff's inequality to solve this question.

chernoff's inequality:

$$\mathbb{P}[X \geq a] \leq e^{-as}\mathbb{E}[e^{sX}], \quad s > 0$$

$$\mathbb{P}[X \leq a] \leq e^{-as}\mathbb{E}[e^{sX}], \quad s < 0$$

now we can say that:

$$\mathbb{P}[X \geq t] \leq e^{-st}\mathbb{E}[e^{sX}]$$

as we proved in the last part, we have:

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}}$$

so we have:

$$\mathbb{P}[X \geq t] \leq e^{-st} e^{\frac{s^2\sigma^2}{2}} = e^{\frac{s^2\sigma^2}{2} - st}$$

now we can solve this equation:

$$\frac{s^2\sigma^2}{2} - st = -\frac{t^2}{2\sigma^2}$$

if we solve this equation for s then $s = \frac{t}{\sigma^2}$. obviously $t > 0$ and $s > 0$ so we proved that:

$$\mathbb{P}[X \geq t] \leq e^{-\frac{t}{2\sigma^2}}$$

with a similar solution we can prove that:

$$\mathbb{P}[X \leq -t] \leq e^{-\frac{t}{2\sigma^2}}$$

so :

$$\mathbb{P}[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Expectable

Show that for any random variable X, we have:

$$\mathbb{E}[max(X,0)] = \int_0^\infty \mathbb{P}(X \geq x)dx$$

solution

we can use the definition of the expectation to solve this question. the definition of the expectation is:

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

so we have:

$$\mathbb{E}[max(X,0)] = \int_{-\infty}^{\infty} max(x,0)f_x(x)dx = \int_{-\infty}^{0} 0f_x(x)dx + \int_0^\infty xf_x(x)dx$$

$$= \int_0^\infty xf_x(x)dx = \int_0^\infty \mathbb{P}(X \geq x)dx$$

we can prove the last equation as below:

$$\int_0^\infty \mathbb{P}(X \geq x)dx = \int_0^\infty \int_x^\infty f_x(t)dtdx = \int_0^\infty \int_0^t f_x(t)dxdt = \int_0^\infty tf_x(t)dt$$

substituting $t = x$ we have:

$$\int_0^\infty tf_x(t)dt = \int_0^\infty xf_x(x)dx$$

so:

$$\int_0^\infty \mathbb{P}(X \geq x)dx = \int_0^\infty x f_x(x)dx = \mathbb{E}[max(X,0)]$$

multivariate Gaussian

suppose that y is a *Gaussian vector* , in other words we have:

$$y \sim \mathcal{N}(\mu, \Sigma)$$

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

show the following statement:

$$p(y_2) = \mathcal{N}(\mu_2, \Sigma_{22})$$

$$p(y_1|y_2) = \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

solution

from defenition of the multivariate Gaussian distribution we know that if we have two jointly Gaussian random variables y_1 and y_2 each with mean μ_1 and μ_2 and covariance Σ_{11} and Σ_{22}, then the joint distribution of y_1 and y_2 is:

$$p(y_1, y_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)}$$

wich the mean vector μ and the covariance matrix Σ are:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

μ_1 and μ_2 are the means of y_1 and y_2 and Σ_{11} and Σ_{22} are the variances of y_1 and y_2 and Σ_{12} and Σ_{21} are the covariances of y_1 and y_2. so we can write y_1 and y_2 as:

$$y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}), y_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

so the first statement is proved. now we can prove the second statement. we know that the conditional distribution of y_1 given y_2 is:

$$p(y_1|y_2) = \frac{p(y_1, y_2)}{p(y_2)}$$

we can write $p(y_1, y_2)$ as:

$$p(y_1, y_2) \propto exp\left\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}$$

$$\Rightarrow p(y_1, y_2) \propto exp\left\{-\frac{1}{2}\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}\right\}$$

shcurr complement states that if we have a block matrix like $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ and A is invertible, then the inverse of the matrix is:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A-BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}$$

using schurr's complement we can write Σ^{-1} as:

$$\Sigma^{-1} = \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \begin{pmatrix} (\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}$$

now we calculate the exponent of the Gaussian distribution:

$$-\frac{1}{2}\begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}$$

we define A , B and C as:

$$A = \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix}$$

$$B = \begin{pmatrix} (\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}$$

$$C = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}$$

now we calculate A and C:

$$A = \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} = \begin{pmatrix} y_1-\mu_1-\Sigma_{22}^{-1}\Sigma_{21}(y_2-\mu_2) \\ y_2-\mu_2 \end{pmatrix}^T$$

$$C = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} = \begin{pmatrix} y_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2) \\ y_2-\mu_2 \end{pmatrix}$$

so we have:

$$-\frac{1}{2}ABC =$$

$$= -\frac{1}{2}(y_1-\mu_1-\Sigma_{22}^{-1}\Sigma_{21}(y_2-\mu_2))(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(y_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2)) \quad (*)$$

$$-\frac{1}{2}(y_2-\mu_2)^T\Sigma_{22}^{-1}(y_2-\mu_2) \quad (**)$$

as we proved in previous part $(**)$ is aqtually $p(y_2) = \frac{1}{\sqrt{\Sigma_{22}}} e^{\left((y_2-\mu_2)^T\Sigma_{22}^{-1}(y_2-\mu_2)\right)}$ so as we mentioned before$(p(y_1,y_2) = p(y_2)p(y_1|y_2))$ we can prove that the $(*)$ is the exponent of the conditional distribution of y_1 given y_2 so we have:

$$p(y_1|y_2) \propto exp\{*\}$$

so by defenition we can say that the conditional distribution of y_1 given y_2 is:

$$p(y_1|y_2) = \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

and this way the second statement is proved.

Conditional multivariate Gaussian

Take:

$$z \in \mathscr{R}^D, y \in \mathscr{R}^K, W \in \mathscr{R}^{K\times D}, b \in \mathscr{R}^K$$

if we have:

$$p(z) = \mathcal{N}(\mu_z, \Sigma_z)$$
$$p(y|z) = \mathcal{N}(Wz + b, \Sigma_{y|z})$$

show the following statement:

$$p(z,y) = \mathcal{N}(\mu, \Sigma)$$
$$\mu = \begin{pmatrix} \mu_z \\ W\mu_z + b \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & W\Sigma_z W^T + \Sigma_{y|z} \end{pmatrix}$$
$$p(z|y) = \mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$$
$$\mu_{z|y} = \Sigma_{z|y}(W^T\Sigma_{y|z}^{-1}(y - b) + \Sigma_z^{-1}\mu_z)$$
$$\Sigma_{z|y}^{-1} = \Sigma_z^{-1} + W^T\Sigma_{y|z}^{-1}W$$

solution

to prove the first statement we use the previous question result. we calculate $p(y,z)$ first.as we proved in the previous question:

$$p(y_1,y_2) = p(y_2)p(y_1|y_2) \propto exp\left\{-\frac{1}{2}\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{pmatrix}\right\}$$

$$= e^{\left\{-\frac{1}{2}(y_2-\mu_2)^T\Sigma_{22}^{-1}(y_2-\mu_2)\right\}} \times$$
$$e^{\left\{-\frac{1}{2}(y_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2))^T(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(y_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2))\right\}}$$

so we have:

$$y_1 = y \quad , y_2 = z \quad , \mu_1 = W\mu_z + b \quad , \mu_2 = \mu_z$$
$$\Sigma_{11} = \Sigma_{y|z} + W\Sigma_z W^T \quad , \Sigma_{22} = \Sigma_z \quad , \Sigma_{12} = W\Sigma_z \quad , \Sigma_{21} = \Sigma_z W^T$$

$$exp\left\{-\frac{1}{2}\begin{pmatrix} y-W\mu_z-b \\ z-\mu_z \end{pmatrix}^T \begin{pmatrix} \Sigma_{y|z}+W\Sigma_z W^T & W\Sigma_z \\ \Sigma_z W^T & \Sigma_z \end{pmatrix}^{-1} \begin{pmatrix} y-W\mu_z-b \\ z-\mu_z \end{pmatrix}\right\}$$

$$= e^{\left\{-\frac{1}{2}(z-\mu_z)^T\Sigma_z^{-1}(z-\mu_z)\right\}} \times$$

$$e^{\left\{-\frac{1}{2}(y-W\mu_z-b-W\Sigma_z\Sigma_z^{-1}(z-\mu_z))^T(\Sigma_{y|z}+W\Sigma_z W^T-W\Sigma_z\Sigma_z^{-1}\Sigma_z W^T)^{-1}(y-W\mu_z-b-W\Sigma_z\Sigma_z^{-1}(z-\mu_z))\right\}}$$

$$\Rightarrow p(y,z) =\propto exp\left\{-\frac{1}{2}(z-\mu_z)^T\Sigma_z^{-1}(z-\mu_z)\right\} \times$$

$$exp\left\{-\frac{1}{2}(y-Wz-b)^T(\Sigma_{y|z})^{-1}(y-Wz-b)\right\} = p(z)p(y|z)$$

$$\Rightarrow p(y,z) = \mathcal{N}(\mu,\Sigma), \mu = \begin{pmatrix} W\mu_z+b \\ \mu_z \end{pmatrix}, \Sigma = \begin{pmatrix} W\Sigma_z W^T+\Sigma_{y|z} & \Sigma_z W^T \\ W\Sigma_z & \Sigma_z \end{pmatrix}$$

$$\Rightarrow p(z,y) = \mathcal{N}(\mu,\Sigma), \mu = \begin{pmatrix} \mu_z \\ W\mu_z+b \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & W\Sigma_z W^T+\Sigma_{y|z} \end{pmatrix}$$

for the next part we know that $p(z,y) = p(z|y)p(y)$ so we can write:

$$p(z,y) \propto e^{\left\{-\frac{1}{2}\begin{pmatrix} z-\mu_z \\ y-Wz-b \end{pmatrix}^T \begin{pmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & W\Sigma_z W^T+\Sigma_{y|z} \end{pmatrix}^{-1} \begin{pmatrix} z-\mu_z \\ y-Wz-b \end{pmatrix}\right\}}$$

as we proved in the previous question if we have a block matrix like $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ and A is invertible, then the inverse of the matrix is:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -D^{-1}C & I \end{pmatrix} \begin{pmatrix} (A-BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ 0 & I \end{pmatrix}$$

so we have:

$$\Sigma^{-1} = ABC$$

where

$$A = \begin{pmatrix} I & 0 \\ -W\Sigma_z W^T-\Sigma_{y|z} & I \end{pmatrix}$$

$$B = \begin{pmatrix} (\Sigma_z-\Sigma_z W^T(W\Sigma_z W^T+\Sigma_{y|z})^{-1}W\Sigma_z)^{-1} & 0 \\ 0 & (W\Sigma_z W^T+\Sigma_{y|z})^{-1} \end{pmatrix}$$

$$C = \begin{pmatrix} I & -\Sigma_z W^T(W\Sigma_z W^T+\Sigma_{y|z})^{-1} \\ 0 & I \end{pmatrix}$$

by matrix mulipication we can say that:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_z^{-1}+W^T\Sigma_{y|z}^{-1}W & -W^T\Sigma_{y|z}^{-1} \\ -\Sigma_{y|z}^{-1}W & \Sigma_{y|z}^{-1} \end{pmatrix}$$

and its diagonalization is:

$$\Sigma^{-1} = \begin{pmatrix} I & 0 \\ -\Sigma_{y|z}^{-1} W \Lambda_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \begin{pmatrix} I & -\Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1} \\ 0 & I \end{pmatrix}$$

where $\Lambda_{11} = \Sigma_z^{-1} + W^T \Sigma_{y|z}^{-1} W$ and $\Lambda_{22} = \Sigma_{y|z}^{-1} - \Sigma_{y|z}^{-1} W \Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1}$. now we have:

$$p(z,y) \propto$$

$$e^{\left\{ -\frac{1}{2} \begin{pmatrix} z-\mu_z \\ y-Wz-b \end{pmatrix}^T \begin{pmatrix} I & 0 \\ -\Sigma_{y|z}^{-1} W \Lambda_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix} \begin{pmatrix} I & -\Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} z-\mu_z \\ y-Wz-b \end{pmatrix} \right\}}$$

so by matrix multipixation we have:

$$p(z,y) \propto e^{\left\{ -\frac{1}{2}(y-W\mu_z-b)^T \Lambda_{22} (y-W\mu_z-b) \right\}} \times$$

$$e^{\left\{ -\frac{1}{2}(z-\mu_z-\Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1}(y-W\mu_z-b))^T \Lambda_{11} (z-\mu_z-\Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1}(y-W\mu_z-b)) \right\}}$$

as we can see the first part of the exponent is the exponent of the Gaussian distribution of y so as we know $p(z,y) = p(z|y)p(y)$ the second exponent is $p(z|y)$ so we have:

$$\mu_{z|y} = \Lambda_{11}^{-1} W^T \Sigma_{y|z}^{-1} (y - W\mu_z - b) + \mu_z$$

and

$$\Sigma_{z|y}^{-1} = \Lambda_{11} = \Sigma_z^{-1} + W^T \Sigma_{y|z}^{-1} W$$

so if we rewrite the $\mu_{z|y}$ we have:

$$\mu_{z|y} = \Sigma_{z|y} (W^T \Sigma_{y|z}^{-1} (y-b) + (\Sigma_{z|y})^{-1} - W^T \Sigma_{y|z}^{-1} W) \mu_z$$

$$\Rightarrow \mu_{z|y} = \Sigma_{z|y} (W^T \Sigma_{y|z}^{-1} (y-b) + \Sigma_z^{-1} \mu_z)$$

Gaussian Mixture models

Now assume that in the previous part, the prior distribution is a mixture of K gaussian distributions (GMM), that is , $p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$, for which we clearly have $\sum_{k=1}^{K} \pi_k = 1$.Prove the posterior distribution is another GMM, and calculate its parameters.
Hint to avoid a common mistake: The posterior coefficients π_k' are not equal to π_k.

solution

we use a same approach as two previous questions:

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

$$p(y|z) = \mathcal{N}(Wz+b, \Sigma_{y|z})$$

$$p(y,z) = p(y|z)p(y)$$

$$p(y,z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)\mathcal{N}(Wz+b, \Sigma_{y|z})$$

$$= \sum_{k=1}^{K} \pi_k e^{\left\{-\frac{1}{2}(z-\mu_k)^T\Sigma_k^{-1}(z-\mu_k)\right\}} e^{\left\{-\frac{1}{2}(y-Wz-b)^T\Sigma_{y|z}^{-1}(y-Wz-b)\right\}}$$

so we can write $p(z,y)$ as below:

$$p(z,y) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu'_k, \Sigma'_k)$$

which from the previous part we know:

$$\mu'_k = \begin{pmatrix} \mu_k \\ W\mu_k + b \end{pmatrix}$$

$$\Sigma'_k = \begin{pmatrix} \Sigma_k & \Sigma_k W^T \\ W\Sigma_k & W\Sigma_k W^T + \Sigma_{y|z} \end{pmatrix}$$

as we know we can write $p(y) = \int_{-\infty}^{\infty} p(z,y)dz$ so we have:

$$p(y) = \sum_{k=1}^{K} \pi_k \int_{-\infty}^{\infty} \mathcal{N}(\mu'_k, \Sigma'_k)dz$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(W\mu_k + b, W\Sigma_k W^T + \Sigma_{y|z})$$

so we can write the posterior distribution as:

$$p(z|y) = \frac{p(z,y)}{p(y)} = \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mu'_k, \Sigma'_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(W\mu_k + b, W\Sigma_k W^T + \Sigma_{y|z})}$$

$$\Rightarrow p(z|y) = \sum_{k=1}^{K} \pi'_k \mathcal{N}(W\mu'_k + b, W\Sigma'_k W^T + \Sigma_{y|z})$$

and we can calculate the π'_k as:

$$\pi'_k = \frac{\pi_k \mathcal{N}(W\mu_k + b, \Sigma_{y|z}^{-1}(I - W(\Sigma_k + W^T\Sigma_{y|z}^{-1}W)^{-1}W^T)\Sigma_{y|z}^{-1})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(W\mu_k + b, \Sigma_{y|z}^{-1}(I - W(\Sigma_k + W^T\Sigma_{y|z}^{-1}W)^{-1}W^T)\Sigma_{y|z}^{-1})}$$

Estimators are the Key!

Suppose we have a random vector $X \in \mathbb{R}^d$. All elements are assumed to be *i.i.d* random variables. Assume that we have an observation x. We want to fit a probability distribution to this data and we are going to use the Maximum Likelihood Estimator for that.

MLE 1

Assume that each X_i is a Bernoulli random variable, *i.e.*, $p_{x_i} = \theta^{x_i}(1-\theta)^{1-x_i}$. Also assume that we have observed m ones and k zeros. Find the distribution parameter θ.

solution

from slides we know that we can define *negetive log likelihood* as:

$$NLL(\theta) = -\sum_{n=1}^{N} \log p(y_n|x_n,\theta)$$

so we have an optimization problem:

$$\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)$$

now from defenition we have:

$$NLL(\theta) = -\log \prod_{n=1}^{N} p(y_n|\theta) = -\log \prod_{n=1}^{N} \theta^{\mathbb{I}(y_n=1)}(1-\theta)^{\mathbb{I}(y_n=0)}$$

$$= -\sum_{n=1}^{N} \mathbb{I}(y_n=1)\log\theta + \mathbb{I}(y_n=0)\log(1-\theta)$$

$$= -\log\theta \sum_{n=1}^{N} \mathbb{I}(y_n=1) - \log(1-\theta)\sum_{n=1}^{N} \mathbb{I}(y_n=0)$$

as it is mentioned in the question, we have m ones and k zeros, so we have:

$$\hat{\theta}_{MLE} = \arg\min_{\theta}(-m\log\theta - k\log(1-\theta))$$

so:

$$\frac{d}{d\theta}(-m\log\theta - k\log(1-\theta)) = -[\frac{m}{\theta} - \frac{k}{1-\theta}] = 0$$

$$\Rightarrow m - \theta(m+k) = 0 \Rightarrow$$

$$\hat{\theta}_{MLE} = \frac{m}{m+k}$$

MLE 2

Assume that each X_i is a Normal random variable, *.i.e* $p_{x_i} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$. Find the mean and variance of the distribution.

solution

in this question we suppose:

- $Y \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta = (\mu, \sigma^2)$
- $D = \{y_1, y_2, ..., y_N\}$
- $\hat{\theta}_{MLE} = \{\hat{\mu}_{MLE}, \hat{\sigma^2}_{MLE}\}$

now we use the same approach as the last question. we can define *negetive log likelihood* as:

$$NLL(\theta) = -\sum_{n=1}^{N} \log p(y_n|\mu, \sigma^2)$$

so we have an optimization problem:

$$\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)$$

now from defenition we have:

$$NLL(\theta) = -\log \prod_{n=1}^{N} p(y_n|\mu, \sigma^2) = -\log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n-\mu)^2}{2\sigma^2}}$$

$$= -\sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_n-\mu)^2}{2\sigma^2}$$

$$= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mu)^2$$

now we have:

$$\hat{\mu}_{MLE} = \frac{\partial}{\partial\mu} NLL(\mu, \sigma^2) = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n-\mu) = 0$$

$$\Rightarrow \sum_{n=1}^{N}(y_n-\mu) = 0 \Rightarrow \sum_{n=1}^{N} y_n = N\mu \Rightarrow$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{n=1}^{N} y_n$$

we also have:

$$\hat{\sigma^2}_{MLE} = \frac{\partial}{\partial\sigma^2} NLL(\mu, \sigma^2) = \frac{N}{4\pi\sigma^2} - \frac{1}{2\sigma^4}\sum_{n=1}^{N}(y_n-\mu)^2 = 0$$

$$\Rightarrow \frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(y_n-\mu)^2 \Rightarrow$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n-\mu)^2$$

$$\hat{\sigma^2}_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(y_n-\mu)^2$$

Bias-Variance

Show that for any estimator $\hat{\theta}$ of the parameter θ, we have the following:

$$\mathbb{E}[(\hat{\theta}-\theta)^2] = (\mathbb{E}[\hat{\theta}]-\theta)^2 + Var(\hat{\theta})$$

solution

$$\mathbb{E}[(\hat{\theta}-\theta)^2] = \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2]$$
$$= \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2$$
$$= \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 + \mathbb{E}[\hat{\theta}]^2 - \mathbb{E}[\hat{\theta}]^2$$

as we know:

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

and also:

$$\mathbb{E}[\hat{\theta}]^2 - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2 = (\mathbb{E}[\hat{\theta}]-\theta)^2$$

so it can be proved that:

$$\mathbb{E}[(\hat{\theta}-\theta)^2] = (\mathbb{E}[\hat{\theta}]-\theta)^2 + Var(\hat{\theta})$$

Linear Regression

Consider the following *Linear Regression model*.

$$Y_i = ax_i + b + Z_i$$

Z_i's are *i.i.d* and of $\mathcal{N}(0,\sigma^2)$ distribution. We know the value of σ and we are given n data like $(x_1,Y_1),(x_2,Y_2),\ldots,(x_n,Y_n)$. Using MLE, say how can we estimate $\hat{a}$,$\hat{b}$.(No calculations is needed for this part)

solution

we know that Z_i's are *i.i.d* and of $\mathcal{N}(0,\sigma^2)$ distribution. so we can say that:

$$p_{Z_i} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z_i^2}{2\sigma^2}}$$

we also can say that:

$$Z_i = Y_i - ax_i - b$$

so we can rewrite the distribution of Z_i as:

$$p_{Z_i} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - ax_i - b)^2}{2\sigma^2}}$$

now we suppose that:

- $\theta = (a,b)$
- $D = \{(x_1,Y_1),(x_2,Y_2),\ldots,(x_n,Y_n)\}$
- $\hat{\theta}_{MLE} = \{\hat{a}_{MLE}, \hat{b}_{MLE}\}$

so we can define *negetive log likelihood* as:

$$NLL(\theta) = -\sum_{n=1}^{N} \log p(z_n|a,b)$$

so we have an optimization problem:

$$\hat{\theta}_{MLE} = \arg\min_{\theta} NLL(\theta)$$

now from defenition we have:

$$NLL(a,b) = -\log \prod_{n=1}^{N} p(z_n|a,b) = -\log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_n - ax_n - b)^2}{2\sigma^2}}$$

$$= -\sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(Y_n - ax_n - b)^2}{2\sigma^2}$$

$$= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(Y_n - ax_n - b)^2$$

so we can say that:

$$\hat{a}_{MLE} = \frac{\partial}{\partial a} NLL(a,b) = -\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n(Y_n - ax_n - b) = 0$$

and also:

$$\hat{b}_{MLE} = \frac{\partial}{\partial b} NLL(a,b) = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(Y_n - ax_n - b) = 0$$

by solving these two equations we can find $\hat{a}_{MLE}$ and $\hat{b}_{MLE}$.

Blind estimation

We are given $X_1, X_2, \ldots, X_n$ independent samples from X distribution with mean μ and $Var(X) = \sigma^2$. We want to do an ε-accurate estimation of μ. Which means that we want our estimation to be in the $(\mu - \varepsilon, \mu + \varepsilon)$ range. Show that for an ε-accurate estimation, if we have $n = \mathcal{O}(\frac{\sigma^2}{\varepsilon^2})$, then with probability at least $\frac{3}{4}$ we will reach our goal.

solution

we use chebyshev's inequality to solve this question. chebyshev's inequality is defined as:

chebyshev's inequality:

$$\mathbb{P}(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

if we define sample mean as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,the variance of $\bar{X}$ is $\frac{\sigma^2}{n}$. so by using this inequality we can say that:

$$\mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow 1-\mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \geq 1-\frac{\sigma^2}{n\varepsilon^2} > \frac{3}{4}$$

$$\Rightarrow \mathbb{P}(|\bar{X}-\mu| < \varepsilon) \geq 1-\frac{\sigma^2}{n\varepsilon^2} > \frac{3}{4}$$

now we solve the right side of the inequality:

$$1-\frac{\sigma^2}{n\varepsilon^2} > \frac{3}{4} \Rightarrow \frac{\sigma^2}{n\varepsilon^2} < \frac{1}{4}$$

solving this inequality for n we have:

$$n > \frac{4\sigma^2}{\varepsilon^2}$$

so we can say that if we have $n = \mathcal{O}(\frac{\sigma^2}{\varepsilon^2})$, then with probability at least $\frac{3}{4}$ we will reach our goal.

Eigenvalues

Assume $\mathbf{A}$ is a 2×2 matrix with λ_1 and λ_2 being its eigenvalues. If $\lambda_1 \neq \lambda_2$, prove:

$$e^A = \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I + \frac{e^{\lambda_1} - e^{\lambda_2}}{\lambda_1 - \lambda_2} A$$

solution

first of all we prove the fact that if $\lambda_1 \neq \lambda_2$, then A is diagonalizable. We know that A is diagonalizable if and only if it has n linearly independent eigenvectors. Then we have to prove that if $\lambda_1 \neq \lambda_2$, then v_1 and v_2 are linearly independent. we assume that:

$$\alpha v_1 + \beta v_2 = 0$$

then we have:

$$\alpha\lambda_1 v_1 + \beta\lambda_1 v_2 = 0 \quad (1)$$

we can also state that:

$$\alpha A v_1 + \beta A v_2 = 0 = \alpha\lambda_1 v_1 + \beta\lambda_2 v_2 \quad (2)$$

by subtracting (2) from (1) we have:

$$\beta(\lambda_2 - \lambda_1)v_2 = 0$$

by the fact that $\lambda_1 \neq \lambda_2$, then $\beta = 0$. *The same can be done for* α.so we can conclude that v_1 and v_2 are linearly independent. so A is diagonalizable. we can write A as:

$$A = V\Lambda V^{-1} \quad (3)$$

where V is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. by taylor expansion we have:

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!} = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots \quad (4)$$

by (3) and some matrix multiplication we have:

$$A^n = V\Lambda^n V^{-1}$$

so we can rewrite (4) as:

$$e^A = I + V\Lambda V^{-1} + \frac{V\Lambda^2 V^{-1}}{2!} + \cdots = I + \sum_{i=1}^{\infty} \frac{V\Lambda^i V^{-1}}{i!} = V\sum_{i=0}^{\infty} \frac{\Lambda^i}{i!} V^{-1}$$

now we can write $V\frac{\Lambda^i}{i!}V^{-1}$ as:

$$V \begin{bmatrix} \frac{\lambda_1^i}{i!} & 0 \\ 0 & \frac{\lambda_2^i}{i!} \end{bmatrix} V^{-1}$$

we know that:

$$e^{\lambda_1} = \sum_{i=0}^{\infty} \frac{\lambda_1^i}{i!}$$

$$e^{\lambda_2} = \sum_{i=0}^{\infty} \frac{\lambda_2^i}{i!}$$

so we can write $\sum_{i=0}^{\infty} \frac{\Lambda^i}{i!}$ as:

$$\begin{bmatrix} e^{\lambda_1} & 0 \\ 0 & e^{\lambda_2} \end{bmatrix}$$

so we can write e^A as:

$$e^A = V \begin{bmatrix} e^{\lambda_1} & 0 \\ 0 & e^{\lambda_2} \end{bmatrix} V^{-1} \quad (*)$$

now we need to prove that the right hand side of the equation is equal to the left hand side. we write the right hand side as:

$$\frac{V\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I\, V^{-1} + \frac{e^{\lambda_1} - e^{\lambda_2}}{\lambda_1 - \lambda_2} V\Lambda V^{-1}$$

so we have:

$$V \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I\, V^{-1} + V \begin{bmatrix} \frac{\lambda_1 e^{\lambda_1} - \lambda_1 e^{\lambda_2}}{\lambda_1 - \lambda_2} & 0 \\ 0 & \frac{\lambda_2 e^{\lambda_1} - \lambda_2 e^{\lambda_2}}{\lambda_1 - \lambda_2} \end{bmatrix} V^{-1}$$

$$= V \begin{bmatrix} \frac{\lambda_1 e^{\lambda_1} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} & 0 \\ 0 & \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_2}}{\lambda_1 - \lambda_2} \end{bmatrix} V^{-1} = V \begin{bmatrix} e^{\lambda_1} & 0 \\ 0 & e^{\lambda_2} \end{bmatrix} V^{-1} \quad (**)$$

by $(*)$ and $(**)$ we can conclude that the right hand side of the equation is equal to the left hand side. so we proved the equation.

SVD Decomposition

If the SVD decomposition of matrix A is defined as $A = U\Sigma V^T$, then the pseudo-inverse matrix is defined as $A^\dagger = V\Sigma^{-1}U^T$

pseudo-inverse

Show that if A has full row rank, then we have: $A^\dagger = A^T(A^TA)^{-1}$ and if it has full column rank, then we have: $A^\dagger = (A^TA)^{-1}A^T$. The invertibility of square matrices does not need to be proven.

solution

if A has full row rank we know that $U^{-1} = U^T$ and $V^{-1} = V^T$. so we have:

$$A^T(A^TA)^{-1} = A^TA^{-1}A^{T^{-1}} = V\Sigma U^TV\Sigma^{-1}U^TU\Sigma V^T$$

$$= V\Sigma U^TV\Sigma^{-1}\Sigma V^T = V\Sigma U^TVV^T = V\Sigma U^T = A^\dagger$$

similarly if A has full column rank we have:

$$(A^TA)^{-1}A^T = A^{-1}A^{T^{-1}}A^T = A^{-1} = V\Sigma^{-1}U^T = A^\dagger$$

SVD

Find the SVD decomposition of the matrix $A = \begin{bmatrix} 1 & 3 & 1 \\ 2 & -1 & 2 \end{bmatrix}$

solution

$$A^TA = \begin{bmatrix} 1 & 2 \\ 3 & -1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 3 & 1 \\ 2 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 1 & 5 \\ 1 & 10 & 1 \\ 5 & 1 & 5 \end{bmatrix}$$

$$\Rightarrow \det(A^TA - \lambda I) = \det\begin{bmatrix} 5-\lambda & 1 & 5 \\ 1 & 10-\lambda & 1 \\ 5 & 1 & 5-\lambda \end{bmatrix} = 0$$

$$\Rightarrow \lambda_1 = 10+\sqrt{2}, \lambda_2 = 10-\sqrt{2}, \lambda_3 = 0$$

$$\Rightarrow \sigma_1 = \sqrt{10+\sqrt{2}}, \sigma_2 = \sqrt{10-\sqrt{2}}, \sigma_3 = 0$$

$$\Rightarrow \hat{v}_1 = \frac{1}{2}(1,\sqrt{2},1), \hat{v}_2 = \frac{1}{2}(1,-\sqrt{2},1), \hat{v}_3 = \frac{1}{\sqrt{2}}(-1,0,1)$$

$$\Rightarrow u_1 = \frac{Av_1}{\sigma_1} = (\frac{\sqrt{2+\sqrt{2}}}{2}, \frac{\sqrt{2-\sqrt{2}}}{2}), u_2 = \frac{Av_2}{\sigma_2} = (-\frac{\sqrt{2-\sqrt{2}}}{2}, \frac{\sqrt{2+\sqrt{2}}}{2})$$

$$\Rightarrow U = \begin{bmatrix} \frac{\sqrt{2+\sqrt{2}}}{2} & -\frac{\sqrt{2-\sqrt{2}}}{2} \\ \frac{\sqrt{2-\sqrt{2}}}{2} & \frac{\sqrt{2+\sqrt{2}}}{2} \end{bmatrix}, \Sigma = \begin{bmatrix} \sqrt{10+\sqrt{2}} & 0 & 0 \\ 0 & \sqrt{10-\sqrt{2}} & 0 \end{bmatrix}$$

$$, V = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\Rightarrow A = U\Sigma V^T$$

Vector differentiation

Prove the following vector differentiation formulas.

statement 1

$$\nabla_x(a^T x) = \nabla_x(x^T a) = a$$

solution

we can write $a^T x$ as:

$$a^T x = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

Then we have:

$$\nabla_x(a^T x) = \begin{bmatrix} \frac{\partial(a^T x)}{\partial x_1} \\ \frac{\partial(a^T x)}{\partial x_2} \\ \vdots \\ \frac{\partial(a^T x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

we can also write $x^T a$ as:

$$x^T a = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

Then we have:

$$\nabla_x(x^T a) = \begin{bmatrix} \frac{\partial(x^T a)}{\partial x_1} \\ \frac{\partial(x^T a)}{\partial x_2} \\ \vdots \\ \frac{\partial(x^T a)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a$$

this concludes the proof.

statement 2

$$\nabla_x(Tr\{xx^T A\}) = \nabla_x(x^T Ax) = (A + A^T)x$$

first we denote $B = x^T Ax$ so:

solution

$$B = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix}$$

$$= x_1(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n) +$$

$$+ \cdots + x_n(a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n)$$

Then we have:

$$\nabla_x(x^T Ax) = \begin{bmatrix} \frac{\partial(x^T Ax)}{\partial x_1} \\ \frac{\partial(x^T Ax)}{\partial x_2} \\ \vdots \\ \frac{\partial(x^T Ax)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + (a_{12} + a_{21})x_2 + \cdots + (a_{1n} + a_{n1})x_n \\ a_{21}x_1 + 2a_{22}x_2 + \cdots + (a_{2n} + a_{n2})x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + 2a_{nn}x_n \end{bmatrix}$$

so:

$$\nabla_x(x^T Ax) = (A + A^T)x$$

we also can say that $Tr\{xx^T A\} = Tr\{x^T Ax\}$ so:

$$\nabla_x(Tr\{xx^T A\}) = \nabla_x(Tr\{x^T Ax\})$$

as we proved that:

$$B = x_1(a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n) +$$

$$+ \cdots + x_n(a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n)$$

so $Tr\{x^T Ax\} = B$ so:

$$\nabla_x(Tr\{xx^T A\}) = \nabla_x(Tr\{x^T Ax\}) = \nabla_x(x^T Ax) = (A + A^T)x$$

Gradient without explicit differentiation!

Another method to calculate gradients without explicit differentiation is using an equivalent definition of gradients: For small ΔX, we have:

$$f(x + \Delta x) - f(x) \approx \langle \nabla_x f(x), \Delta x \rangle$$

Using this equation and the fact that in the matrix space, the Frobenius inner product is defined as $\langle A,B\rangle = Tr\{A^TB\}$, prove the following for a symmetric matrix X:

$$\nabla_X(-\log\det\{X\}) = -X^{-1}$$

Hint: You may need an eigenvalue decomposition somewhere in your solution.

solution

$f(X) = \log\det\{X\}$ so:

$$f(X+\Delta X) = \log\det\{X+\Delta X\}$$
$$= \log\det\{X^{\frac{1}{2}}(I+X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}})X^{\frac{1}{2}}\}$$
$$= \log\det\{X\} - \log\det\{I+X^{-\frac{1}{2}}\Delta X X^{-\frac{1}{2}}\}$$
$$= \log\det\{X\} - \sum_{i=1}^{n}\log(1+\lambda_i)$$

fo small λ_i we have:

$$\log(1+\lambda_i) \approx \lambda_i$$

so we can rewrite the equation as:

$$f(X+\Delta X) = \log\det\{X\} + \sum_{i=1}^{n}\lambda_i$$

$$= \log\det\{X\} + Tr\{X^{-1}\Delta X\}$$

we know that $f(X) = \log\det\{X\}$ so based on the given equation we have:

$$f(X+\Delta X) - f(X) \approx \langle \nabla_X f(X), \Delta X\rangle$$

$$\Rightarrow \log\det\{X+\Delta X\} - \log\det\{X\} \approx \langle \nabla_X \log\det\{X\}, \Delta X\rangle$$
$$= Tr\{\nabla_X \log\det\{X\}^T \Delta X\} = TrX^{-1}\Delta X$$

based on the last equation we can say that:

$$\nabla_X \log\det\{X\} = X^{-1}$$

so:

$$\nabla_X(-\log\det\{X\}) = -X^{-1}$$

Gradient without explicit differentiation! Part 2

Using the method of the previous part, prove the following:

$$\nabla_X Tr\{X^{-1}A\} = -X^{-T}A^TX^{-T}$$

Hint: An asymmetric matrix is not always diagonalizable! Use another method for the difference of matrices in your solution.

solution

like the previous part in this part we define $f(x) = Tr(X^{-1}A)$ so:

$$f(X+\Delta X) = Tr((X+\Delta X)^{-1}A)$$

using taylor expansion for $(X+\Delta X)^{-1}$ for small ΔX we have:

$$(X+\Delta X)^{-1} \approx X^{-1} - X^{-1}\Delta X X^{-1}$$

so we have:

$$f(X+\Delta X) = Tr(X^{-1}A) - Tr(X^{-1}\Delta X X^{-1}A)$$
$$\Rightarrow f(X+\Delta X) - f(X) = -Tr(X^{-1}\Delta X X^{-1}A)$$

using the fact that $f(X+\Delta X) - f(X) \approx \langle \nabla_X f(X), \Delta X \rangle$ we have:

$$-Tr(X^{-1}\Delta X X^{-1}A) \approx \langle \nabla_X Tr(X^{-1}A), \Delta X \rangle$$

$$\Rightarrow -Tr(X^{-1}\Delta X X^{-1}A) \approx Tr\{\nabla_X Tr(X^{-1}A)^T \Delta X\}$$
$$\Rightarrow \nabla_x f(x)^T = -X^{-1}AX^{-1}$$

so:

$$\nabla_X Tr\{X^{-1}A\} = -X^{-T}A^T X^{-T}$$

Matrix Frobenius Norm

We define the Frobenius norm of a matrix $A \in \mathbb{R}^{m\times n}$ as:

$$||A||_F = \sqrt{Tr\{A^T A\}}$$

prove the following statements:

Statement 1

prove $||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2}$

solution

we can write A by its column as $A = [a_1, a_2, \cdots, a_n]$, where a_i is the i^{th} column of A. Then we have:

$$A^T A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix} [a_1, a_2, \cdots, a_n] = \begin{bmatrix} a_1^T a_1 & a_1^T a_2 & \cdots & a_1^T a_n \\ a_2^T a_1 & a_2^T a_2 & \cdots & a_2^T a_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n^T a_1 & a_n^T a_2 & \cdots & a_n^T a_n \end{bmatrix}$$

Then we have:

$$Tr\{A^T A\} = \sum_{i=1}^{n} a_i^T a_i = \sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2$$

so we can rewrite Frobenius norm as:

$$||A||_F = \sqrt{Tr\{A^T A\}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2}$$

Statement 2

Given singular values of matrix A : $||A||_F = \sqrt{\sum_{i=1}^{r}\sigma_i(A)^2}$

solution

we know that the singular value decomposition of a matrix A is given by:

$$A = U\Sigma V^T$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with singular values of A on its diagonal. Then we have:

$$A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma V^T$$

Then we have:

$$Tr\{A^TA\} = Tr\{V\Sigma^T\Sigma V^T\} = Tr\{\Sigma^T\Sigma V^TV\} = Tr\{\Sigma^T\Sigma\}$$

Then we have:

$$||A||_F = \sqrt{Tr\{A^TA\}} = \sqrt{Tr\{\Sigma^T\Sigma\}} = \sqrt{\sum_{i=1}^{r}\sigma_i(A)^2}$$

Statement 3

conclude $max\{\sigma_i(A)\} \leq ||A||_F \leq \sqrt{r}max\{\sigma_i(A)\}$

solution

in the previous section we proved that:

$$||A||_F = \sqrt{\sum_{i=1}^{r}\sigma_i(A)^2}$$

Then we have:

$$max\{\sigma_i(A)\}^2 \leq \sum_{i=1}^{r}\sigma_i(A)^2 \leq rmax\{\sigma_i(A)\}^2$$

this comes from the fact that $max\{\sigma_i(A)\} \geq \sigma_i(A)$ for all i . Then we have:

$$max\{\sigma_i(A)\} \leq \sqrt{\sum_{i=1}^{r}\sigma_i(A)^2} \leq \sqrt{r}max\{\sigma_i(A)\}$$

Then we have:

$$max\{\sigma_i(A)\} \leq ||A||_F \leq \sqrt{r}max\{\sigma_i(A)\}$$

Right or wrong!

Determine the correctness or incorrectness of the following items with sufficient reasoning.

statement 1

if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ are matrices with full rank and $AB = 0$ then we have: $p+m \leq n$.

solution

if A is a $m \times n$ matrix and B is a $n \times p$ matrix, we can define them as below:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

if we suppose that $n \leq m$, if A is full rank then we have $rank(A) = n$ and $dim(N(A)) = n-n$. so we have $dim(N(A)) = 0$. in this case if $AB = 0 \Rightarrow B = 0$ and B is not full rank so we have a contradiction. so our assumption is wrong and we have $m \leq n$.
now if we suppose that $m \leq n$, if A is full rank then we have $rank(A) = m$ and $dim(N(A)) = n-m$. by the fact that $AB = 0$ we can say that the columns of B are in the null space of A so we have $dim(C(B)) \leq n-m$ now if we suppose that $n \leq p$ by the fact that B is full rank we have $rank(B) = n$ so $dim(C(B)) = n$ so we have $n \leq n-m$ which is a contradiction so we have $p \leq n$ and $dim(C(B)) = p$ so we have $p+m \leq n$.

statement 2

If for some integer values $k \geq 1$ we have $A^k = 0$ then $A-I$ is a matrix with full rank.

solution

if $k = 1$ then we have $A^1 = 0$ so $A-I = -I$ and this matrix is invertible so it is full rank.
if $k > 1$ and we have $A^k = 0$ then $A^k - I = -I$. we can factorize the left side of equation as below:

$$A^k - I = (A-I)(A^{k-1} + A^{k-2} + \cdots + A + I) = -I$$

if we name $B = A^{k-1} + A^{k-2} + \cdots + A + I$ then we have $(A-I)(-B) = I(*)$. on the other hand we can factorize the left side as below:

$$A^k - I = B(A-I) = -I$$

then we have $(-B)(A-I) = I(*)(*)$.
by the definition of the inverse of a matrix we can say that if $AB = BA = I$ the A has an inverse and its inverse is B.
so we can say that by the equations $(*)$,$(**)$, $A-I$ is invertible and its invert is $-B$. hence $A-I$ is a full rank matrix.

statement 3

for every $A, B \in \mathbb{R}^{n\times n}$, the eignvectors of AB are equal to the eigenvectors of BA.

solution

we can prove that this statement is wrong by a counter example.if we define A and B as below:

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}$$

we can calculate AB and BA as below:

$$AB = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 8 \\ 0 & 4 \end{pmatrix}$$

$$BA = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 0 & 4 \end{pmatrix}$$

the eignvalues of both AB and BA are equal to 3 and 4. if we want to calculate the eignvectors of AB and BA we can calculate the null space of $AB-\lambda I$ and $BA-\lambda I$ for $\lambda = 3,4$. if we calculate the null space of $AB-3I$ and $BA-3I$ we have:

$$AB-3I = \begin{pmatrix} 0 & 8 \\ 0 & 1 \end{pmatrix} \quad BA-3I = \begin{pmatrix} 0 & 6 \\ 0 & 1 \end{pmatrix}$$

so the null space of $AB-3I$ is equal to $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and the null space of $BA-3I$ is equal to $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. now if we calculate the null space of $AB-4I$ and $BA-4I$ we have:

$$AB-4I = \begin{pmatrix} -1 & 8 \\ 0 & 0 \end{pmatrix} \quad BA-4I = \begin{pmatrix} -1 & 6 \\ 0 & 0 \end{pmatrix}$$

so the null space of $AB-4I$ is equal to $\begin{pmatrix} 8 \\ 1 \end{pmatrix}$ and the null space of $BA-4I$ is equal to $\begin{pmatrix} 6 \\ 1 \end{pmatrix}$. so we can say that the eignvectors of AB and BA are not equal.

Calculating normalized eigenvectors from eigenvalues!

We wish to prove the following identity, for diagonalizable matrix A, with diagonalization $A = \sum_{i=1}^{n} \lambda_i(A) v_i(A) v_i(A)^H$, where $\lambda_i(A)$, $1 \leq i \leq n$ are the eigenvalues and $v_i(A)$, $1 \leq i \leq n$ are the corresponding eigenvectors, where all eigenvalues are non-zero and the matrix has a simple spectrum.

$$\left|v_{i,j}(A)\right|^2 \prod_{\substack{k=1 \\ k\neq i}}^{n} (\lambda_i(A) - \lambda_k(A)) = \prod_{k=1}^{n-1} \left(\lambda_i(A) - \lambda_k(M_{ij})\right)$$

Where M_{ij} is the submatrix formed by removing the ith row and jth column from the original matrix, A.

To do this we take the following steps.

Step 1

For any square matrix S , with eigenvalues $\lambda_1(S), \lambda_2(S), \cdots, \lambda_n(S)$ prove that

$$\det\{S\} = \prod_{i=1}^{n} \lambda_i(S)$$

.

solution

from eigenvalue decomposition of S we have $S = V\Lambda V^{-1}$, where Λ is the diagonal matrix of eigenvalues of S and V is the matrix of eigenvectors of S.

$$S = V\Lambda V^{-1} \quad , \quad \Lambda = \begin{bmatrix} \lambda_1(S) & 0 & \cdots & 0 \\ 0 & \lambda_2(S) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n(S) \end{bmatrix}$$

so by the properties of the determinant we have

$$\det\{S\} = \det\{V\Lambda V^{-1}\} = \det\{V\}\det\{\Lambda\}\det\{V^{-1}\}$$

we also know that $\det\{V^{-1}\} = \frac{1}{\det\{V\}}$, so

$$\det\{S\} = \det\{V\}\det\{\Lambda\}\frac{1}{\det\{V\}} = \det\{\Lambda\}$$

as Λ is a diagonal matrix , $\det\{\Lambda\} = \prod_{i=1}^{n} \lambda_i(S)$, so we have

$$\det\{S\} = \prod_{i=1}^{n} \lambda_i(S)$$

Step 2

Prove $A\ \text{adj}(A) = \det(A)I = \text{adj}(A)A$, in wich $\text{adj}(A)$ is the adjugate matrix of A. We define the coefficients of the $\text{adj}(A)$ by $\text{adj}(A)_{ij} = (-1)^{i+j}\det\{M_{ji}\}$

solution

we know that determinant of a matrix can be written as below:

$$\det\{A\} = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}\det\{M_{ij}\}$$

so we have:

$$(A\text{adj}(A))_{kl} = \sum_{i=1}^{n} a_{ki}(-1)^{i+l}\det\{M_{li}\}$$

if we assume $k = l$ we have:

$$(A\text{adj}(A))_{ll} = \sum_{i=1}^{n} a_{li}(-1)^{i+l}\det\{M_{li}\} = \det\{A\}$$

and if we assume $k \neq l$ we have:

$$(A\text{adj}(A))_{kl} = \sum_{i=1}^{n} a_{ki}(-1)^{i+l}\det\{M_{li}\} = 0$$

so if we define expansion of the determinant of the matrix as A^{kl} as:

$$A^{kl} = \begin{cases} a_{ij} & \text{if } i \neq l \\ a_{kj} & \text{if } i = l \end{cases}$$

this expansion has two equal rows so we can say that:

$$A\text{adj}(A) = \det\{A\}I$$

similarly we can prove that $\text{adj}(A)A = \det\{A\}I$

Step 3

show that $\text{adj}(A)$ has a diagonalization $\text{adj}(A) = \sum_{i=1}^{n}\left(\prod_{k=1,k\neq i}^{n}\lambda_k(A)\right)v_i(A)v_i(A)^H$.

solution

we know that by definition $A^n v_i = \lambda_i^n v_i$, so we can say that:

$$A^{-1}v_i = \frac{1}{\lambda_i}v_i \Rightarrow A^{-1} = \sum_{i=1}^{n}\frac{1}{\lambda_i}v_i v_i^H$$

we also know that $A^{-1} = \frac{\text{adj}(A)}{\det\{A\}}$, so we can say that:

$$\text{adj}(A) = \det\{A\}A^{-1} = \det\{A\}\sum_{i=1}^{n}\frac{1}{\lambda_i}v_i v_i^H$$

as we proved in the first part of this question, $\det\{A\} = \prod_{i=1}^{n}\lambda_i(A)$, so we have:

$$\text{adj}(A) = \sum_{i=1}^{n}\left(\prod_{k=1,k\neq i}^{n}\lambda_k(A)\right)v_i(A)v_i(A)^H$$

Step 4

now prove the identity.

solution

in the last part if we replace A with $\lambda I - A$ we have:

$$\text{adj}(\lambda I - A) = \sum_{i=1}^{n}\left(\prod_{k=1,k\neq i}^{n}(\lambda - \lambda_k(A))\right)v_i(A)v_i(A)^H$$

we know that $\text{adj}(\lambda I - A) = \det\{\lambda I - A\}I$, so we have:

$$\det\{\lambda I - A\}I = \sum_{i=1}^{n}\left(\prod_{k=1,k\neq i}^{n}(\lambda - \lambda_k(A))\right)v_i(A)v_i(A)^H$$

we also know that $\det\{\lambda I - A\} = \prod_{i=1}^{n}(\lambda - \lambda_i(A))$, so we have:

$$\prod_{i=1}^{n}(\lambda - \lambda_i(A))I = \sum_{i=1}^{n}\left(\prod_{k=1,k\neq i}^{n}(\lambda - \lambda_k(A))\right)v_i(A)v_i(A)^H$$

by multiplying both sides by $\left(\prod_{k=1,k\neq i}^{n}(\lambda - \lambda_k(A))\right)^{-1}$ we have:

$$\left|v_{i,j}(A)\right|^2\prod_{\substack{k=1\\k\neq i}}^{n}(\lambda_i(A) - \lambda_k(A)) = \prod_{k=1}^{n-1}\left(\lambda_i(A) - \lambda_k(M_{ij})\right)$$

Optimization

In the following lessons, you will become familiar with various classifiers, one of which is Support Vector Machine or SVM for short. In this question, we aim to examine this classifier for inseparable data. As you will see later, to find the best classifier, we will encounter an optimization problem with inequality constraints as follows.

$$\begin{cases} \min\limits_{\mathbf{w},w_0,\eta} & J(\mathbf{w},w_0,\eta)=\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum\limits_{i=1}^{N}\eta_i \\ \text{s.t.} & y_i(\mathbf{w}^T x_i+w_0)\geq 1-\eta_i \quad , \quad \eta_i\geq 0 \quad i=1,2,...,N \end{cases}$$

step 1

Formulate the Lagrangian for the above problem.

solution

we can define the lagrangian as

$$L(\mathbf{w},w_0,\eta,\alpha,\beta)=\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i=1}^{N}\eta_i-\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T x_i+w_0)-1+\eta_i)-\sum_{i=1}^{N}\beta_i\eta_i$$

step 2

Obtain the solution to the problem by applying the Karush-Kuhn-Tucker (KKT) conditions.

solution

We can solve the problem by applying the KKT conditions. The KKT conditions are as follows:

- primal feasibility: $\begin{cases} y_i(\mathbf{w}^T x_i+w_0)\geq 1-\eta_i \\ \eta_i\geq 0 \end{cases}$
- dual feasibility: $\begin{cases} \alpha_i\geq 0 \\ \beta_i\geq 0 \end{cases}$
- stationary: $\begin{cases} \nabla_{\mathbf{w}}L=\mathbf{w}-\sum\limits_{i=1}^{N}\alpha_i y_i x_i=0 \\ \frac{\partial}{\partial w_0}L=-\sum\limits_{i=1}^{N}\alpha_i y_i=0 \\ \frac{\partial}{\partial \eta}L=C-\alpha_i-\beta_i=0 \end{cases}$
- Complementary slackness: $\begin{cases} \alpha_i(y_i(\mathbf{w}^T x_i+w_0)-1+\eta_i)=0 \\ \beta_i\eta_i=0 \end{cases}$