Machine Learning Theory

25737-2

Spring Semester 1402-03
Department of Electrical Engineering
Sharif University of Technology

Ali Nikkhah
99102445

Table of content

1 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.

Note that I also done simulation on this question

1.1 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)

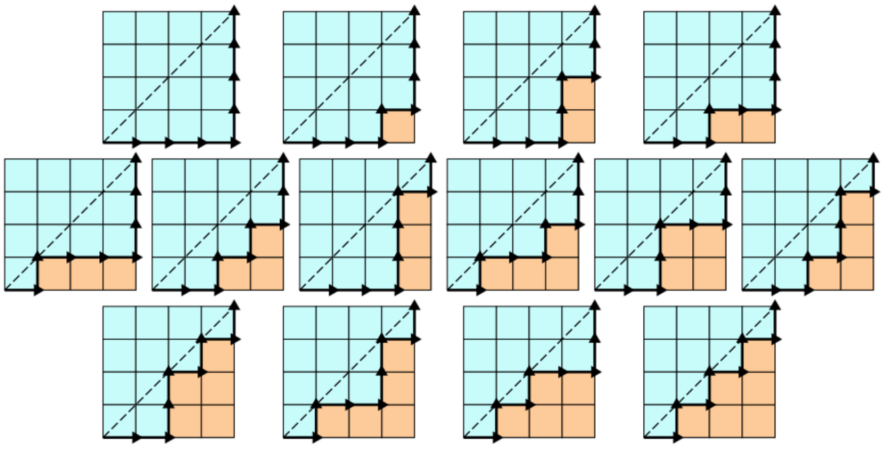
---

the above problem is equivalent to summing up all of the paths that conclude to observing 0 at least at one trial before start of the observation, to the fraction of all of possible occurrences.

$$P(x) = \frac{N(t)}{2^n}$$

Now the problem breaks down to finding value of function N(t)

For this part we use Catalan numbers theorem, stated below.



Catalan numbers statement

For last example, N=4 and by defining a sequence of height of columns we have the set uniquely representing C(4)=14:

[0,0,0,0] [0,0,0,1] [0,0,0,2] [0,0,1,1]

[0,1,1,1] [0,0,1,2] [0,0,0,3] [0,1,1,2] [0,0,2,2] [0,0,1,3]

[0,0,2,3] [0,1,1,3] [0,1,2,2] [0,1,2,3]

The interesting fact is, for our problem Catalan representation on the figure contain all the paths, with at least one intersection with y=x line, and by defining -> as 1 and ^ as -1, intersection with y=x line is equivalent to reaching 0 in the summation screen.

Starting with n=1, N(t)=0

For n=2, the sequence can be {-1,1} and {1,1} (which is indeed Catalan number of 2) thus P(2) = 0.5

For values greater than 2 we have:

n=2k+1: for every odd value of n the function N(t) equals 2 * N(t-1)

Because:

Set of n=2k+1 is [{set of N(t-1)},1] and [{set of N(t-1),-1]

$$P(2k+1) = \frac{2N(2k)}{2^{2k+1}} = P(2k)$$

Now we state our answer for n = 2k,

For every {set of N(t-2)} any combination of adding 1 and -1 to the end and the beginning of the N(t-2) series will be an answer of T=t trials, which means N(t) = N(t-2) * 4

$$P(2k) = P(2k-2) + \frac{C(k)}{2^{2k+1}}$$

Or equivalently,

$$P(n \in 2k) = \sum_{j=0, j=2q}^{n} \frac{C(j)}{2^{2j+1}} = \frac{1}{2} \sum_{j=0, j=2q}^{n} \frac{C(j)}{4^j} \rightsquigarrow \sum_{n=1}^{\lfloor T/2 \rfloor} \frac{1}{2n-1} (\frac{1}{2})^n \binom{2n}{n}$$

By this definition we have:

$$P(2) = \frac{C(0)}{2^1} = \frac{1}{2}$$

$$P(4) = \frac{C(0)}{2^1} + \frac{C(1)}{2^3} = \frac{5}{8}$$

$$P(6) = P(4) + \frac{C(3)}{2^5} = \frac{5}{8} + \frac{2}{2^5} = \frac{22}{32}$$

And we can keep on calculating above formula for every value of n

T	P(T)
1	0.0
2	0.5
3	0.5
4	0.625
5	0.625
6	0.6875
7	0.6875
8	0.7265625
9	0.7265625

Python simulation results of the probability

1.2 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).

By using the formula we calculated earlier we have:

$$\lim_{t\to\infty} P(t) = lim_{t\to\infty} \sum_{j=0,j=2q}^{t} \frac{C(j)}{2^{2j+1}}$$

Now using Integral representation of Catalan number:

$$C_n = \frac{1}{2\pi}\int_0^4 x^n\sqrt{\frac{4-x}{x}}dx = \frac{2}{\pi}4^n\int_{-1}^{1} t^{2n}\sqrt{1-t^2}dt$$

$$\sum_{n=0}^{\infty}\frac{C_n}{4^n} = \sum_{n=0}^{\infty}\frac{\frac{2}{\pi}4^n\int_{-1}^{1} t^{2n}\sqrt{1-t^2}dt}{4^n} = \sum_{n=0}^{\infty}\frac{2}{\pi}\int_{-1}^{1} t^{2n}\sqrt{1-t^2}dt$$

$$= \sum_{n=0}^{\infty}\frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (sin(u))^{2n}cos(u)^2du \text{ using Beta function we conclude}$$

$$\sum_{n=0}^{\infty}\frac{C(n)}{4^n} = 2 \text{ which equivalently means}$$

$$lim_{t\to\infty}\sum_{j=0,j=2q}^{t}\frac{C(j)}{2^{2j+1}} = \sum_{n=0}^{\infty}\frac{C(n)}{2^{2n+1}} = 1$$

So the probability of observing zero at least one time in all trials reaches one when we have ∞ trials in sight.

1.3 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.

Let X be the random variable representing the number of rounds needed to succeed.
The probability mass function of the geometric distribution is given by:

$P(X = k) = (1 - p)^{k-1}p$

where p is the probability of success on a single trial.

p is probability of correct guess in a n cell array, with iid bernoulli variables:

$$p = (\frac{1}{2})^n$$

If we fail the **remaining** mean number of trials until a success is identical to the original mean.

$$E[X] = p + (1 - p)(1 + E[X]) \rightsquigarrow E[X] = \frac{1}{p} = 2^n$$

Which is equivalent to Brute-force all possible sequences, and algorithm is trivial.

1.4 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.

First let's break down problem with one example:

Hidden Permutation: [2, 4, 1, 3]

- Round 1:
    - Machine guesses the first unrevealed position, starting from the least unrevealed number, which is 1. Incorrect.
    - Moves on to the next least unrevealed number, which is 2. Correct! Moves to the next position.
- Round 2:
    - Machine guesses the second unrevealed position, starting from the least unrevealed number, which is 3. Incorrect.
    - Moves on to the next least unrevealed number, which is 4. Correct! Moves to the next position.
- Round 3:
    - Machine guesses the third unrevealed position, starting from the least unrevealed number, which is 1. Correct! Moves to the next position.
- Round 4:

    - Machine guesses the fourth unrevealed position, starting from the least unrevealed number, which is 3. Correct!

Now let's examine worst case scenario :

Hidden Permutation: [4, 3, 2, 1]

- Position 1:
    - Machine guesses the first unrevealed position, starting from the least unrevealed number, which is 1. Incorrect.
    - Moves on to the next least unrevealed number, which is 2. Incorrect
    - Moves on to the next least unrevealed number, which is 3. Incorrect
    - Moves on to the next least unrevealed number, which is 4, Going to the next position.
- Position 2:
    - Starting from the least unrevealed number, which is 1. Incorrect.
    - Moves on to the next least unrevealed number, which is 2. Incorrect
    - Moves on to the next least unrevealed number, which is 3, Correct and going to the next position
- Round 3:
    - Starting from the least unrevealed number, which is 1. Incorrect.
    - Moves on to the next least unrevealed number, which is 2. Correct, going to last position
- Round 4:
    - Machine guesses the fourth unrevealed position, starting from the least unrevealed number, which is 4. Correct!

Therefore the worst case time for estimation is:

$$\sum_{i=1}^{4} i = \frac{4*5}{2} = 10$$

Now with intuition gained, we can calculate expected time for out Intelligent! Algorithm:

Let Xi be probability of correctly guessing i-th position of our random permutation. Since the permutation of numbers is uniformly distributed (i.e. each number has equal probability to be anywhere), $p_i = \frac{1}{n-i+1}$

Now imagine were in I-th position and our initial guess was wrong then we have:

$$p_i^1 = \frac{1}{n-i+1}$$

$$p_i^2 = \frac{1}{n-i+1-1}$$

$$p_i^3 = \frac{1}{n-i+1-2}$$

…

$$p_i^{n-i+1} = 1$$

Then we define the expected value of I-th position with j-1 wrong guesses:

$$E_i^j = E_i^{n-i+1-n+i} = E_i^1 = \frac{n-i+2}{2}$$

$$\mathbb{E} = \sum_{i=1}^{n} E_i^j = \sum \frac{n+2}{2} - \sum \frac{i}{2} = \frac{n^2+2n}{2} - \frac{n^2+n}{4} = \frac{1}{4}[n^2+3n]$$

Also we know for any algorithm , the worst case there are $\frac{n(n+1)}{2}$ questions asked (wrong guesses) and thus the time it takes should be this value.

If we want to maximize the algorithm performance we should add memory to it (like we just did) and by this trade off we achieve maximal performance.

The key is, no Mather what is our strategy, guessing least number or maximum number or anything in between, in case of wrong guess will lead the same answer, the thing is $Max(p_i^j) = \frac{1}{n-i+1-j+1}$ is always the probability of making a uniform-random guess between {n-i+1} unrevealed numbers ∪ {-j +1} numbers that proven to be not the correct guess for the I-th position.

So we conclude we are solving the problem using the best possible algorithm.

1.5 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).

1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	**8**	9	10	N
1	2	3	4	**5**	6	7	8	9	10	N
1	2	3	4	5	6	7	8	**9**	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	**N**
1	2	**3**	4	5	6	7	8	9	10	N
1	2	3	4	5	6	**7**	8	9	10	N

Random selection of the vector cells

We know that sampling vector indices randomly and then performing operation on them is equivalent to starting from first value of the vector, doing operation till convergence and then going to the next cell. Here is a short proof why this statement is true

Assume we have a series of converging selection of vector cells, each operation is denoted by $S_{indx=i}^{iter=t}$ and list of operations is denoted as $\mathbf{S} = \{S_i^t\}$ and since each index is independent to other indexes of S, we can sort this series by value of $\mathbf{i}$ in increasing order, in other words we can solve the problem for each index independently and since they are identical, we can say:

$$\mathbf{T} = \sum_{i=1}^{N} T_i \rightsquigarrow E[\mathbf{T}] = \sum_{i=1}^{N} E[T_i] = NE[T_i]$$

Or the expected time for running algorithm with a vector with N cells is equivalent of running algorithm for one cell, N times.

1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N
1	2	3	4	5	6	7	8	9	10	N

Start from the first index till convergence, then next index

After proving this equation for the vector, let's focus on the problem using only one cell of the vector and then using the above property.

For the proof of the first part, let's assume we have a single cell, with starting value ζ then we want to add/subtract this cell with gaussian sample and check convergence and end/iterate. We can chop the process at any iteration and assume current value is the new starting point, denoted by $\overline{\zeta}$ and then solve the problem identically.

$$\zeta \rightarrow \nu \sim \mathcal{N}(0,1) \rightarrow \overline{\zeta} = \zeta \pm \nu \rightarrow \ldots$$

Now lets prove first part:

The distance after each iteration is denoted $D = |\zeta - X|$

Then we have:

$$\mathbb{E}[(\zeta + \Delta - X)^2] = \mathbb{E}[(\zeta - X + \Delta)^2] = \mathbb{E}[(\zeta - X)^2] + 2\mathbb{E}[(\zeta - X)\Delta] + \mathbb{E}[\Delta^2]$$

Then we have

$$\mathbb{E}[(\zeta + \Delta - X)^2] = \mathbb{E}[(\zeta - X)^2] + \mathbb{E}[\Delta^2] = \mathbb{E}[(\zeta - X)^2] + 1$$

If we assume ζ is a constant number then we have:

$$1 + \mathbb{E}[\zeta^2 - 2\zeta X + X^2] = 1 + \zeta^2 - 2\zeta\mathbb{E}[X] + \mathbb{E}[X^2]$$

$$\rightarrow min(2 + \zeta^2) \rightsquigarrow \zeta = 0$$

Also from the first moment we have:

$$min|\mathbb{E}[X - \zeta]| = min|\mathbb{E}[X] - \mathbb{E}[\zeta]| \rightsquigarrow \zeta = 0$$

Another reason for this property would be our secret vector symmetric nature to 0.

We denote expected number of iterations as $D(\zeta)$ then

$$D(\zeta) = \begin{cases} 0 & \text{if } \zeta = X \\ 1 + D(\zeta + \Delta) & \text{if } \zeta + \Delta \text{ gets closer to } X \\ 1 + D(\zeta - \Delta) & \text{if } \zeta - \Delta \text{ gets closer to } X \end{cases}$$

By symmetry of normal pdf probability of the first and second part of iteration cases are the same,

$$D(\zeta) = 1 + \frac{1}{2}D(\zeta + \Delta) + \frac{1}{2}D(\zeta - \Delta)$$

Then we can state that the algorithm is most efficient when we start from 0, due to the symmetry.

2 Statistics and other friends

2.1 LLN & CLT

Briefly explain Law of Large Numbers and Central Limit Theorem.

---

weak law of large numbers:

Suppose sequence of i.i.d random variables $X_1, X_2, X_3, \ldots, X_n$ with parameters μ, σ^2 and let $\overline{X_n}$ denote sample mean of random variables.

Using Chebyshev's inequality we have:

$$P(|Y - \mu_Y| \geq \epsilon) \leq \frac{\sigma_Y^2}{\epsilon^2}$$

$$P(|\overline{X_n} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

$$\delta = \frac{\sigma^2}{n\epsilon^2}$$

As n approaches infinity, δ approaches zero as much as we neeed or any $\epsilon > 0$

Which results in

$$P(|\overline{X_n} - \mu| \geq \epsilon) \leq \delta$$

Which is statement of WLLN.

$$\text{Let } A_n = \{\, |\overline{X_n} - \mu| \geq \epsilon \} \rightsquigarrow \sum_{n=1}^{\infty} P(A_n) \leq \infty$$

Using chebyshev inequality above we have:

$$P(A_n) = P(|\overline{X_n} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

$$\sum_{n=1}^{\infty} P(A_n) \leq \sum_{n=1}^{\infty} \frac{\sigma^2}{n\epsilon^2}$$

This series converges (it's a harmonic series multiplied by a constant), so by the Borel-Cantelli lemma, we have:

$$P(\limsup_{n\to\infty} A_n) = 0$$

Which states for any value of ϵ , $|\overline{X_n} - \mu| \leq \epsilon$ when $n \to \infty$

Central Limit Theorem can be stated as follows:

The Central Limit Theorem (CLT) states that the sum (or average) of a large number of independent and identically distributed (i.i.d.) random variables, regardless of the original distribution, will tend towards a normal distribution.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$$

The characteristic function of a random variable X is defined as
$\phi_X(t) = E[e^{itX}]$

$$\phi_{\bar{X}_n}(t) = E[e^{it\bar{X}_n}] = \prod_{i=1}^{n} E[e^{itX_i/n}] = \left(E[e^{itX}]\right)^n = \left(\phi_X\left(\frac{t}{n}\right)\right)^n$$

$$\phi_X\left(\frac{t}{n}\right) = \phi_X(0) + \frac{t}{n}\phi_X'(0) + \frac{t^2}{2n^2}\phi_X''(0) + \ldots$$

$$\phi_{\bar{X}_n}(t) \approx \left(1 + \frac{it}{n}E[X] - \frac{t^2}{2n^2}E[X^2]\right)^n$$

$$\lim_{n\to\infty} \phi_{\bar{X}_n}(t) = e^{itE[X] - \frac{t^2 E[X^2]}{2}}$$

The resulting characteristic function is that of a normal distribution with mean $E[X]$ and variance $\frac{E[X^2]}{2}$

2.2 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?

---

1. Relevance to Statistics:

- Law of Large Numbers (LLN):
  - Implication: As the sample size increases, the sample mean converges to the true population mean. This provides a foundation for the reliability of sample statistics as estimators of population parameters.
  - Application in Statistics: LLN is the basis for understanding the precision and accuracy of statistical estimates, such as sample means and proportions. It assures statisticians that larger sample sizes lead to more reliable estimates.
- Central Limit Theorem (CLT):
  - Implication: The distribution of the sample mean (or sum) becomes approximately normal, regardless of the original distribution of individual observations. This is particularly powerful because the normal distribution is well-understood.
  - Application in Statistics: CLT is widely used in hypothesis testing and constructing confidence intervals. It allows statisticians to make inferences about population parameters using the normal distribution, even if the population distribution is unknown or not normal.

2. Machine Learning Applications:

- Data Preprocessing:
  - LLN: In machine learning, when dealing with large datasets, LLN assures us that statistical properties computed from the data, such as means and variances, are likely to be close to the true population values.
  - CLT: Large datasets in machine learning often involve the aggregation of many random variables. The CLT supports the assumption of normality in many machine learning algorithms, making them more robust and reliable.
- Modeling and Inference:
  - LLN: In the context of model training, LLN reinforces the idea that larger datasets lead to more accurate model parameter estimates.
  - CLT: Machine learning algorithms often rely on assumptions of normality, and the CLT facilitates the use of statistical tests and confidence intervals in assessing model performance and uncertainty.

- Statistical Learning Theory:
  - LLN: The LLN is foundational in statistical learning theory, assuring that empirical averages converge to expected values as the sample size grows.
  - CLT: The CLT is crucial in establishing the distributional properties of sample statistics, providing insights into the behavior of learning algorithms and their convergence.

2.3 Are you sure about that?

Briefly explain Hypothesis Test and Confidence Interval.

Hypothesis Testing is a statistical method used to make inferences about a population parameter based on a sample of data. The process involves formulation a null hypothesis noted by H_0 and alternative hypothesis noted by H_a and in the process we conclude that given collected data whether we can reject null hypothesis in favor of the alternative, or not.

In the process we compare p-value with the significance level, denoted by α and reject or fail-to-reject the null hypothesis.

Confidence interval is a range of values constructed from sample data, within we except a population parameter to lie with a certain level of confidence. It provides a measure of the precision or uncertainty associated with a point estimate.

A 95% confidence interval, for example, means that if we were to take many samples and construct a confidence interval from each, we would expect about 95% of those intervals to contain the true population parameter.

2.4 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?

Relationship to Statistics:

Hypothesis Test:

- Parameter Estimation:
    - Hypothesis tests are often used to make decisions about population parameters. For example, testing whether a population mean is equal to a certain value or comparing means from two different populations.
- Scientific Inquiry:
    - Hypothesis testing is fundamental in scientific research. It helps researchers assess whether observed effects or differences are statistically significant, providing evidence for or against a particular hypothesis.
- Decision Making:
    - Businesses and policymakers use hypothesis tests to inform decisions. For instance, testing the effectiveness of a new product or determining whether a policy change has a significant impact.

Confidence Interval:

- Parameter Estimation with Uncertainty:
    - Confidence intervals provide a range of plausible values for a population parameter, acknowledging the uncertainty associated with the estimate. This is crucial in expressing the precision of an estimate.
- Comparisons and Inference:
    - Confidence intervals can be used to compare population parameters. If the intervals of two groups do not overlap, it may suggest a significant difference between them.
- Visual Representation:

    - Confidence intervals are useful in visually representing the precision of an estimate. This aids in the interpretation of statistical results for a broader audience.

Application in Machine Learning:

Hypothesis Test:

- Model Evaluation:
    - In machine learning, hypothesis tests can be used to assess the significance of differences in model performance metrics. For example, testing whether the performance improvement of a new model is statistically significant.
- Feature Importance:
    - Hypothesis testing can be employed to evaluate the significance of individual features in a model, helping identify which features contribute significantly to predictive performance.
- A/B Testing:
    - In online experiments, A/B testing often involves hypothesis testing to determine whether changes to a website or application result in statistically significant improvements.

Confidence Interval:

- Model Confidence:
    - Confidence intervals for model parameters help express the uncertainty associated with model coefficients. This is crucial for understanding the stability of the model.
- Prediction Intervals:
    - In regression tasks, prediction intervals, a type of confidence interval for individual predictions, provide a range within which future observations are likely to fall.
- Comparing Models:
    - Confidence intervals for model performance metrics (e.g., accuracy or F1 score) can be used to compare different models. If the intervals do not overlap, it may suggest a significant difference in performance.

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

$$Y_1, Y_2, \ldots, Y_n \sim Bernoulli(X_1, \ldots, Xn)$$

$$Z = \sum_{i=1}^{n} \frac{1}{n} Y_i = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$E[Z] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i] \rightarrow Z = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$E[Y_i] = X_i$$

$$Var[Y_i] = X_i(1 - X_i)$$

$$Var[Z] = \frac{1}{n^2} \sum_{i=1}^{n} Var[Y_i] = \frac{1}{n^2} \sum_{i=1}^{n} X_i(1 - X_i)$$

$$max(1 - X_i)(X_i) = \frac{1}{4} \text{ for } X_i = 0.5$$

$$Var[Z] \leq \frac{1}{4n}$$

Then we use CLT and we have:

$$p(-z_{\frac{\alpha}{2}} \leq \frac{\mu - \overline{X}}{Var[Z]} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

For a 95% confidence interval we have:

And then we conclude interval for maximum variance is:

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

3 Its all about Tails

3.1 Not a very hard inequality

Consider random variable X, with E[X] = 0 and Var[X] = σ^2. show that for each a > 0, we have:

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

---

we use markov's inequality which states that:

$$P[Y \geq t] \leq \frac{E[Y]}{t}$$

by defining new random variable Y we have:
$Y = (X - E[X])^2 = X^2 + E[X]^2 - 2E[X]X = X^2 + E[X]^2 = X^2$

$Var[X] = \sigma^2 = E[X^2] - E[X]^2 \rightarrow E[X^2] = \sigma^2$

$E[Y] = E[X^2] = \sigma^2$

$P[Y \geq a^2] \leq \frac{E[Y]}{a^2} \rightarrow P[Y \geq a^2] \leq \frac{\sigma^2}{a^2}$

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

$P[\,|X| \geq a] \leq \frac{\sigma^2}{a^2}$ which is the chebyshev's inequality

Now by applying Markov inequality and Chebyshev's inequality we have:

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

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

By substituting $t = \frac{\sigma^2}{a}$ we have:

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

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

3.2 Gaussian

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

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

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$E[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{sx - \frac{x^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{2sx\sigma^2 - x^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-(x - s\sigma^2)^2 + s^2\sigma^4}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-(x - s\sigma^2)^2}{2\sigma^2}} e^{\frac{s^2\sigma^2}{2}} dx$$

$\Phi(\infty) = 1$ so we have

$$= e^{\frac{s^2\sigma^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-(x - s\sigma^2)^2}{2\sigma^2}} dx = e^{\frac{s^2\sigma^2}{2}}$$

In a more general form we have:

$$E[e^{s(X-\mu)}] = e^{\frac{s^2\sigma^2}{2}}$$

3.3 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:

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

By using Chernoff's Method we have for any random variable X:

$$P(X > \epsilon) \leq inf_{t \geq 0} e^{-t\epsilon} E[e^{tX}]$$

We know that this function can satisfy MGF property we stated earlier

$$M_X(t) = E[e^{tX}] \leq e^{\frac{\sigma^2 t^2}{2}}$$

$$P(X > \epsilon) \leq inf_{t \geq 0} e^{-t\epsilon} E[e^{tX}]$$

$$P(X > \epsilon) \leq inf_{t \geq 0} e^{-t\epsilon} e^{\frac{\sigma^2 t^2}{2}}$$

which is minimized by substitution of t with the value below, and yields tail bound

by substitution of $t = \dfrac{\epsilon}{\sigma^2} > 0$ and by using symmetry we have:

$$P(|X| > \epsilon) \leq 2e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Which is the desired proof by setting $\epsilon \to t$

- X sub-Gaussian implies

$$\begin{aligned} P(X - \mu \geq t) &\leq \exp(-t^2/(2\sigma^2)), \\ P(X - \mu \leq -t) &\leq \exp(-t^2/(2\sigma^2)), \\ P(|X - \mu| \geq t) &\leq 2\exp(-t^2/(2\sigma^2)). \end{aligned}$$

Reference : Oxford statistics course

3.4 Expectable

Prove that for any random variable Z we have:

$$E[Max(0,Z)] = \int_0^{\infty} P(Z \geq x)dx$$

---

by rule LOTUS we have: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

Then by defining $g(z) = (z < 0) \rightarrow 0, (z \geq 0) \rightarrow z$

$$E[Max(0,Z)] = \int_{-\infty}^{\infty} f_Z(z)max(0,z)dz$$

$$= \int_0^{\infty} z f_Z(z)dz$$

$$= \int_0^{\infty} \int_0^{z} dx f_Z(z)dz$$

$$= \int_0^{\infty} \int_0^{z} f_Z(z)dxdz$$

$$= \int_0^{\infty} \int_x^{\infty} f_Z(z)dzdx$$

$$= \int_0^{\infty} P(Z \geq x)dx$$

3.5 ∗Multivariate Gaussian

Y is a Gaussian random vector and we have:

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

$$Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

By definition we know that if Y is a gaussian random variable, then y1 and y2 are jointly normal, in other words

$\forall a_1, a_2 : a_1 y_1 + a_2 y_2$ is a normal random variable.

$QQ^T = I$

$\Sigma = QDQ^T$

Where D is a diagonal matrix, which is PD

Then we can define

$D^{0.5}(D^{0.5})^T = D$

Then we have

$A = QD^{0.5}Q^T$

$AA^T = A^TA = QD^{0.5}Q^TQD^{0.5}Q^T = QDQ^T = \Sigma$

By this definition we can say $Y = AZ + \mu$ because:

$EY = E[AZ + \mu] = AE[Z] + \mu = \mu$

And $C_Y = AC_ZA^T = AA^T = \Sigma$

By this definition we can write down:

$f_Y(y0 = f_z(H(y))\,|J|$

$Z = A^{-1}(Y - \mu)$

$$J = det(A^{-1}) = \frac{1}{det(A)}$$

$$f_Y(y_1, y_2)_= \frac{1}{|det(A)|} f_z(A^{-1}(y-\mu))$$

$$f_Y(y) = \frac{1}{(2\pi)\sqrt{det(\Sigma)}} e^{-\frac{(A^{-1}(y-\mu))^T(A^{-1}(y-\mu))}{2}}$$

$$= \frac{1}{(2\pi)\sqrt{det(\Sigma)}} e^{-\frac{((y-\mu)^T)(A^{-1})^T A^{-1}((y-\mu))}{2}}$$

$$= \frac{1}{(2\pi)\sqrt{det(\Sigma)}} e^{-\frac{1}{2}((y-\mu)^T \Sigma^{-1}(y-\mu))}$$

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} \Sigma_{22} & -\Sigma_{12} \\ -\Sigma_{21} & \Sigma_{11} \end{bmatrix} \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$-\frac{1}{2|\Sigma|}((y-\mu)^T\Sigma^{-1}(y-\mu)) = -\frac{1}{2|\Sigma|}[y_1-\mu_1 \quad y_2-\mu_2]\begin{bmatrix} \Sigma_{22} & -\Sigma_{12} \\ -\Sigma_{21} & \Sigma_{11} \end{bmatrix}\begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{bmatrix}$$

$$= -\frac{1}{2|\Sigma|}\begin{bmatrix} \Sigma_{22}(y_1-\mu_1) - \Sigma_{21}(y_2-\mu_2) & -\Sigma_{12}(y_1-\mu_1) + \Sigma_{11}(y_2-\mu_2) \end{bmatrix}\begin{bmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{bmatrix}$$

$$= -\frac{1}{2|\Sigma|}[\Sigma_{22}(y_1-\mu_1)^2 + \Sigma_{11}(y_2-\mu_2)^2 + (y_1-\mu_1)(y_2-\mu_2)(\Sigma_{21}+\Sigma_{12})]$$

Which is the multivariate form of Normal gaussian distribution

09397521561

3.5.1 Now we can calculate $p(y_2)$ for special case :

$$p(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1$$

$$= \frac{1}{2\pi |\Sigma|^{0.5}} \int_{-\infty}^{\infty} e^{-\frac{1}{2|\Sigma|}[\Sigma_{22}(y_1-\mu_1)^2+\Sigma_{11}(y_2-\mu_2)^2+(y_1-\mu_1)(y_2-\mu_2)(\Sigma_{21}+\Sigma_{12})]} dy_1$$

$$= \frac{e^{-\frac{\Sigma_{11}(y_2-\mu_2)^2}{2|\Sigma|}}}{2\pi |\Sigma|^{0.5}} \int_{-\infty}^{\infty} e^{-\frac{1}{2|\Sigma|}[\Sigma_{22}(y_1-\mu_1)^2+(y_1-\mu_1)(y_2-\mu_2)(\Sigma_{21}+\Sigma_{12})]} dy_1$$

$$= \frac{e^{-\frac{\sigma_1^2(y_2-\mu_2)^2}{2(1-\rho^2)\sigma_1^2\sigma_2^2}}}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \int_{-\infty}^{\infty} e^{-\frac{1}{2|\Sigma|}[\Sigma_{22}(y_1-\mu_1)^2+(y_1-\mu_1)(y_2-\mu_2)(\Sigma_{21}+\Sigma_{12})]} dy_1$$

$$\frac{e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}}}{2\pi\sigma_2\sqrt{1-\rho^2}} = \mathcal{N}(\mu_2, \sqrt{1-\rho^2}\sigma_2)$$

By setting $\rho = 0$ the equation simplifies to:

$$= \frac{e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_1\sqrt{2\pi}\sigma_2} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma_1^2}(y_1-\mu_1)^2} dy_1$$

$$= \frac{e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_1\sqrt{2\pi}\sigma_2} \sqrt{2\pi}\sigma_1\Phi(\infty)$$

$$\rightsquigarrow p(y_2) = \frac{e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2} = \mathcal{N}(\mu_2, \sigma_2)$$

3.5.2 Alternative Approach:

In this approach we use the theorem that all conditional distributions of a multivariate normal distribution are normal.

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

We know that any subset of random variables X (denoted by X_A here) on another subset X_B can be driven from conditioning the joint distribution on X_B and is, indeed normal.

(Proof : https://statproofbook.github.io/P/mvn-cond.html#mjx-eqn-eq:mvn)

We need to find parameters μ, Σ for resulting distribution:

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

Define: $z = y_1 + Ay_2$ and $A = -\Sigma_{12}\Sigma_{22}^{-1}$

$$cov(z, y_2) = cov(y_1, y_2) + cov(Ay_2, y_2) = \Sigma_{12} + Avar(y_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$$

Thus we conclude z and y_2 are uncorrelated, jointly normal, independent.

$$E(z) = \mu_1 + A\mu_2$$

$$E(y_1|y_2) = E(z - Ay_2|x_2) = E(z|y_2) - E(Ay_2|y_2) = E(z) - Ay_2 = \mu_1 + A(\mu_2 - y_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2)$$

$$E(y_1|y_2) = \mu_1 - A(y_2-\mu_2)$$

$$var(y_1|y_2) = var(z - Ay_2|y_2)$$

$$= var(z|y_2) + var(Ay_2|y_2) - Acov(z, -y_2) - cov(z, -y_2)\hat{A}$$

$$= var(z|y_2) = var(z)$$

$$= var(x_1 + Ax_2) = \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma 21$$

$$= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\mu = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2-\mu_2)$$

$$\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$p(y_1|y_2) = Y \sim \mathcal{N}(\mu, \Sigma)$$

$y_1, y_2 \sim \mathcal{N}(\mu, \Sigma)$

3.5.3 Now we can calculate $p(y_2)$ for general case:

Which is also a gaussian normal variable,

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

$p(y_1 \mid y_2) = Y \sim \mathcal{N}(\mu, \Sigma)$ where:

$$\Sigma = \sigma_1^2 - \rho^2 \frac{\sigma_1^2}{\sigma_2^2}$$

$$\mu = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y_2 - \mu)$$

$$p(y_1 \mid y_2) = \frac{1}{\sqrt{2\pi \mid \Sigma \mid}} \exp\left(-\frac{1}{2}(y_1 - \mu)^T \Sigma^{-1} (y_1 - \mu)\right)$$

$$p(y_1 \mid y_2) = \frac{1}{\sqrt{2\pi\left(\sigma_1^2 - \rho^2 \frac{\sigma_1^2}{\sigma_2^2}\right)}} \exp\left(-\frac{1}{2}\left(y_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right)^T \left(\sigma_1^2 - \rho^2\frac{\sigma_1^2}{\sigma_2^2}\right)^{-1} \left(y_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right)\right)$$

$$p(y_2) = \frac{\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)}{\mathcal{N}(\mu, \Sigma)} = \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(y_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right]\right)}{\frac{1}{\sqrt{2\pi\left(\sigma_1^2 - \rho^2\frac{\sigma_1^2}{\sigma_2^2}\right)}} \exp\left(-\frac{1}{2}\left(y_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right)^T \left(\sigma_1^2 - \rho^2\frac{\sigma_1^2}{\sigma_2^2}\right)^{-1} \left(y_1 - \mu_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right)\right)}$$

$$p(y_2) = \frac{1}{\sqrt{(2\pi) \mid \Sigma_{22} \mid}} \exp\left(-\frac{1}{2}(y_2 - \mu_2)^T \Sigma_{22}^{-1} (y_2 - \mu_2)\right) = \mathcal{N}(\mu_2, \sigma_2)$$

3.6 Conditional multivariate Gaussian

If we have

$p(\mathbf{z}) = \mathcal{N}(\mu_z, \Sigma_z)$

$p(\mathbf{y} \mid \mathbf{z}) = \mathcal{N}(\mathbf{W}z + b, \sigma_y)$

Then show the statements:

1. $p(\mathbf{z}, \mathbf{y}) = \mathcal{N}(\mu, \Sigma)$

Where

$$\mu = \begin{bmatrix} \mu_z \\ W\mu_y + b \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & \Sigma_{y|z} + W\Sigma_z W^T \end{bmatrix}$$

2. $p(\mathbf{z} \mid \mathbf{y}) = \mathcal{N}(\mu_{z|y}, \Sigma_{z|y})$

Where

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

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

$$P(y_1, y_2) = p(y_2)p(y_1 \mid y_2) = Kexp(-\frac{1}{2\mid\Sigma\mid}((y-\mu)^T\Sigma^{-1}(y-\mu))) = Kexp(-\frac{1}{2\mid\Sigma\mid}[y_1-\mu_1 \quad y_2-\mu_2]\begin{bmatrix}\Sigma_{22} & -\Sigma_{12}\\ -\Sigma_{21} & \Sigma_{11}\end{bmatrix}\begin{bmatrix}y_1-\mu_1\\ y_2-\mu_2\end{bmatrix})$$

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

With substituting $y_1 = y$, $y_2 = z$

Let Y and Z be random variables with joint distribution $p(y, z)$ then we write $p(Y \mid Z)$ in this form:

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

Then we conclude:

$$p(y, z) = Ke^{-\frac{1}{2}(z-\mu_z)^T\Sigma_z^{-1}(z-\mu_z)}e^{-\frac{1}{2}(y-Wz-b)^T(\Sigma_{y|z})^{-1}(y-Wz-b)}$$

$$p(y, z) = \mathcal{N}(\mu, \Sigma) \text{ where } \mu = \begin{bmatrix} W\mu_z + b \\ \mu_z \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} W\Sigma_z W^T + W\Sigma_z & \Sigma_z W^T \\ W\Sigma_z W^T & \Sigma_z \end{bmatrix}$$

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

3.6.2 Then we use property $p(z, y) = p(y)p(z \mid y)$

By assuming A invertible, matrix inverse is:

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

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

Then we have:

$$A = \begin{bmatrix} I & 0 \\ -\Sigma_z W^T (\Sigma_{y|z}^{-1} & I \end{bmatrix}$$

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

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

Then we conclude by ABC

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

3.7 Gaussian Mixture models

prior distribution: mixture of k G,GMM

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

Prove posterior distribution:

K gaussian distributions exist, then in order to generate a sample, first we select one of the distributions, we choose gaussian Z with probability α_k then we have:

$$X \sim \mathcal{N}(\mu_z, \Sigma_z)$$

$$\Theta = \{\mu_1, .., \mu_k, \Sigma_1, .., \Sigma_k, \alpha_1, .., \alpha_k\}$$

$$p(x;\Theta) = \sum_{k=1}^{k} P(Z = k;\Theta) p(x \mid Z = k;\Theta) = \sum_{k=1}^{K} \alpha_k \phi(x;\mu_k,\Sigma_k)$$

$$= \sum_{k=1}^{K} \frac{1}{(2\pi)^{n/2} |\Sigma|^{0.5}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

$$\hat{\Theta} = \{\hat{\mu}_1, \ldots, \hat{\Sigma}_1, \ldots, \hat{\alpha}_1, \ldots\}$$

Then we use Bayes's rule:

$$P(Z_i \mid x_i; \hat{\Theta}) = \frac{\phi(x_i; \hat{\mu}_k, \hat{\Sigma}_k)\hat{\alpha}_k}{\sum_{k=1}^{K} \phi(x_i; \hat{\mu}_k; \hat{\Sigma}_k)\hat{\alpha}_k)}$$

$$P(Z) = \sum_{i=1}^{K} \phi(x_i; \hat{\mu}_k; \hat{\Sigma}_k) \frac{\hat{\alpha}_k}{\sum_{k=1}^{K} \phi(x_i; \hat{\mu}_k; \hat{\Sigma}_k)\hat{\alpha}_k}$$

Which is also, a GMM

Drive of Q function:

$$Q_t(\Theta) = \mathbb{E}_{Z|X}[log p(X, Z, \Theta)]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z_i|x_i,\Theta_t}[log o(x_i, z_i; \Theta)]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i; \Theta_t) log p(x_i, z_i; \Theta)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k}}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})} log p(x_i; z_i; \Theta)$$

$$\Theta_{t+1} = arg_{\Theta} max Q_t(\Theta)$$

$$s.t. \sum_{k=1}^{K} \alpha_k = 1$$

Solving using Lagrangian:

$$L(\Theta, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k}}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})} log(\alpha_k \phi(x_i, \mu_k, \Sigma_k) + \lambda(\sum_{k=1}^{K} \alpha_k - 1)$$

$$\frac{dL(\Theta, \lambda}{d\alpha_k} = -\frac{1}{\alpha_k} \sum_{i=1}^{n} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k}}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})} + \lambda$$

$$\zeta_{t,i,k} = \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k})}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})}$$

$$\alpha_k = -\frac{1}{\lambda} \sum_{i=1}^{n} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k})}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})}$$

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k}}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})} = -\lambda$$

$$\lambda = -n$$

$$\alpha_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{t,k} \phi(x_i, \mu_k, \Sigma_{t,k}}{\sum_{h=1}^{K} \alpha_{t,h} \phi(x_i; \mu_{t,h}; \Sigma_{t,h})} = \frac{1}{n} \sum_{i=1}^{n} \zeta_{t,i,k}$$

$$\frac{dL(\Theta, \lambda)}{d\mu_k} = \sum \zeta_{t,i,k} \frac{d}{d\mu_k} log(\alpha_k \phi(x_i; \mu_k; \Sigma_k) = \sum_{i=1}^{n} \zeta_{t,i,k} [-\Sigma_k^{-1}(x_i - \mu_k)]$$

$$0 = \sum_{i=1}^{n} \zeta_{t,i,k} [-2\Sigma_k^{-1}(x_i - \mu_k)]$$

$$\mu_k = \frac{1}{\sum_{i=1}^{n} \zeta_{t,i,k}} \sum_{i=1}^{n} \zeta_{t,i,k} x_i$$

Solving for covariance matrix we have:

$$\frac{dL(\Theta,\lambda)}{d\Sigma_k} = \sum_{i=1}^{n} \zeta_{t,i,k} \frac{d}{d\Sigma_k} log\alpha_k \phi(x_i,\mu_k,\Sigma_k) = \sum_{i=1}^{n} \zeta_{t,i,k} \frac{d}{d\Sigma_k} \alpha_k \phi(x_i,\mu_k,\Sigma_k)$$

$$= \sum_{i=1}^{n} \zeta_{t,i,k} [-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}]$$

Thus we have:

$$\Sigma_{k,t+1} = \frac{1}{\sum_{i=1}^{n} \zeta_{t,i,k}} (\sum_{i=1}^{n} \zeta_{t,i,k}(x_i-\mu_i)(x_i-\mu_i)^T)$$

4 Estimators are the Key!

Suppose we have a random vector $X \in R^d$. 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.

4.1 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.

Given that $X_i \sim Bernoulli(\theta)$ the Likelihood function is:

$$L(\theta) = \theta^m(1-\theta)^k$$

$$log(L(\theta)) = mlog(\theta) + klog(1-\theta)$$

$$\frac{d}{d\theta} logL(\theta) = \frac{m}{\theta} - \frac{k}{1-\theta} = 0$$

$$\theta = \frac{m}{m+k}$$

4.2 MLE 2

Assume that each X_i is a Normal random variable,find mean and variance of the distribution.

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$logL(\mu,\sigma^2)=-\frac{n}{2}log(2\pi)-\frac{n}{2}log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\frac{d}{d\mu}logL(\mu,\sigma^2)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0$$

$$\mu=\frac{1}{n}\sum_{i=1}^{n}X_i$$

$$\frac{d}{d\sigma}logL(\mu,\sigma^2)=\frac{2n}{\sigma}-\frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2=0 \rightsquigarrow \sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}$$

4.3 Bias-Variance

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

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

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

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

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

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

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

4.4 Linear Regression

Consider the following Linear Regression model.

$$Y_i = a x_i + b_i + Z_i$$

$$Z_i \sim \mathcal{N}(0,\sigma^2)$$

What is MLE for a_i, b_i

First we define Likelihood function for a single observation:

$$L(a,b \mid x_i, Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - a x_i - b)^2}{2\sigma^2}}$$

$$L(a,b \mid (x_i, Y_i)_{i=1}^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - a x_i - b)^2}{2\sigma^2}}$$

$$log(L) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - a x_i - b)^2$$

$$\frac{dlog(L)}{da} = 0 \rightsquigarrow \sum_{i=1}^n x_i(Y_i - a x_i - b) = 0$$

$$\sum x_i Y_i = a \sum x_i^2 + b \sum x_i$$

$$\frac{dlog(L)}{db} = 0 \rightsquigarrow \sum_{i=1}^n (Y_i - a x_i - b) = 0$$

$$\sum Y_i = a \sum x_i + nb$$

By solving these two equations we have:

$$\sum x_i Y_i = a \sum x_i^2 + b \sum x_i$$

$$\frac{1}{n} \sum x_i \sum y_i = \frac{a}{n} (\sum x_i)^2 + b \sum x_i$$

$$\frac{1}{n} \sum x_i \sum y_i - \frac{a}{n} (\sum x_i)^2 = \sum x_i Y_i - a \sum x_i^2$$

$$a = \frac{\frac{1}{n} \sum x_i \sum y_i - \sum x_i Y_i}{\frac{1}{n} (\sum x_i)^2 - \sum x_i^2}$$

And $\hat{b} = \frac{1}{n}\sum Y_i - \frac{1}{n}\hat{a}\sum x_i = \overline{Y} - \overline{x}\hat{a}$

4.5 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 $n = O(\sigma^2/\varepsilon^2)$ then with probability at least 3/4 we will reach our goal.

---

let $\overline{X}$ be the sample mean we aim that:

$$P(|\overline{X} - \mu| < \epsilon) \geq \frac{3}{4}$$

Using Chebyshev's inequality we have:

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

$$Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

$$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

$$P(|\overline{X} - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

$$n = \mathcal{O}(\frac{\sigma^2}{\epsilon^2}) \rightsquigarrow n = k\frac{\sigma^2}{\epsilon^2}$$

$P(|\overline{X} - \mu| < \epsilon) \geq 1 - \frac{1}{k} \Rightarrow k = 4$ then with probability at least 3/4 we will reach our goal.

5 Eigenvalues

Assume $A_{2\times 2}$ with λ_1, λ_2 as eignvalues then prove that

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

Let's perform eigndecomposition:

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

$det(A - \lambda I) = 0$ then we will have:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Then we have exponential series as follows:

$$e^{PDP^{-1}} = PP^{-1} + PDP^{-1} + \frac{PDP^{-1}PDP^{-1}}{2!} + \dots$$

$$e^{\mathbf{A}} = \sum_{n=0}^{\infty} \frac{V\Lambda^n V^{-1}}{n!}$$

$$A = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T$$

$$\begin{aligned} A^2 &= (\lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T)(\lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T) \\ &= \lambda_1^2 v_1 v_1^T v_1 v_1^T + \lambda_2^2 v_2 v_2^T v_2 v_2^T + \lambda_1\lambda_2 v_1 v_1^T v_2 v_2^T + \lambda_1\lambda_2 v_2 v_2^T v_1 v_1^T \\ &= \lambda_1^2 v_1 (v_1^T v_1) v_1^T + \lambda_2^2 v_2 (v_2^T v_2) v_2^T + \lambda_1\lambda_2 v_1 (v_1^T v_2) v_2^T + \lambda_1\lambda_2 v_2 (v_2^T v_1) v_1^T \\ &= \lambda_1^2 v_1 v_1^T + \lambda_2^2 v_2 v_2^T + \lambda_1\lambda_2 v_1 (v_1^T v_2) v_2^T + \lambda_1\lambda_2 v_2 (v_2^T v_1) v_1^T \\ &= \lambda_1^2 v_1 v_1^T + \lambda_2^2 v_2 v_2^T + \lambda_1\lambda_2 (v_1 \cdot v_2^T)(v_2 \cdot v_1^T) + \lambda_1\lambda_2 (v_2 \cdot v_1^T)(v_1 \cdot v_2^T) \end{aligned}$$

$$A^2 = \lambda_1^2 v_1 v_1^T + \lambda_2^2 v_2 v_2^T$$

$$\begin{aligned} e^A &= I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots \\ &= I + \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \frac{1}{2!}(\lambda_1^2 v_1 v_1^T + \lambda_2^2 v_2 v_2^T) + \frac{1}{3!}(\lambda_1^3 v_1 v_1^T + \lambda_2^3 v_2 v_2^T) + \cdots \\ &= (1 + \frac{\lambda_1}{1!} + \frac{\lambda_1^2}{2!} + \frac{\lambda_1^3}{3!} + \cdots) v_1 v_1^T + (1 + \frac{\lambda_2}{1!} + \frac{\lambda_2^2}{2!} + \frac{\lambda_2^3}{3!} + \cdots) v_2 v_2^T \\ &= e^{\lambda_1} v_1 v_1^T + e^{\lambda_2} v_2 v_2^T \end{aligned}$$

$$B_i = \begin{bmatrix} \frac{P_{1,1}P_{2,2}\lambda_1^i - P_{1,2}P_{2,1}\lambda_2^i}{P_{1,1}P_{2,2} - P_{1,2}P_{2,1}} & \frac{-P_{1,1}P_{2,2}\lambda_1^i + P_{1,2}P_{2,1}\lambda_2^i}{P_{1,1}P_{2,2} - P_{1,2}P_{2,1}} \\ \frac{-P_{1,1}P_{2,2}\lambda_2^i + P_{1,2}P_{2,1}\lambda_1^i}{P_{1,1}P_{2,2} - P_{1,2}P_{2,1}} & \frac{P_{1,1}P_{2,2}\lambda_2^i - P_{1,2}P_{2,1}\lambda_1^i}{P_{1,1}P_{2,2} - P_{1,2}P_{2,1}} \end{bmatrix}$$

$$B = \sum_{i=0}^{\infty} B_i = \frac{1}{|P|} \begin{bmatrix} P_{1,1}P_{2,2}\sum \lambda_1^i - P_{1,2}P_{2,1}\sum \lambda_2^i & -P_{1,1}P_{2,2}\sum \lambda_1^i + P_{1,2}P_{2,1}\sum \lambda_2^i \\ -P_{1,1}P_{2,2}\sum \lambda_2^i + P_{1,2}P_{2,1}\sum \lambda_1^i & P_{1,1}P_{2,2}\sum \lambda_2^i - P_{1,2}P_{2,1}\sum \lambda_1^i \end{bmatrix}$$

$$= \frac{1}{|P|} \begin{bmatrix} P_{1,1}P_{2,2}e^{\lambda_1} - P_{1,2}P_{2,1}e^{\lambda_2} & -P_{1,1}P_{2,2}e^{\lambda_1} + P_{1,2}P_{2,1}e^{\lambda_2} \\ -P_{1,1}P_{2,2}e^{\lambda_2} + P_{1,2}P_{2,1}e^{\lambda_1} & P_{1,1}P_{2,2}e^{\lambda_2} - P_{1,2}P_{2,1}e^{\lambda_1} \end{bmatrix}$$

$$B = Pe^{\Lambda}P^T \rightsquigarrow P^{-1}BP = e^{\Lambda}$$

$$e^{\Lambda} = \frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_2} e^{\Lambda} = \frac{1}{\lambda_1 - \lambda_2} \begin{bmatrix} \lambda_1 e^{\lambda_1} - \lambda_2 e^{\lambda_2} & 0 \\ 0 & \lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_2} \end{bmatrix}$$

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

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

$$Pe^{\Lambda}P^{-1} = \frac{\lambda_1 e^{\lambda_2} - \lambda_2 e^{\lambda_1}}{\lambda_1 - \lambda_2} I + \frac{e^{\lambda_1} - e^{\lambda_2}}{\lambda_1 - \lambda_2} A$$

$$PP^{-1} = I$$

Which is the desired identity

6 SVD Decomposition

If the SVD decomposition of matrix A is defined as

$A = U\Sigma V^T$, then the Moore-Penrose pseudoinverse is $A^\dagger$ defined as $A^\dagger = V\Sigma^\dagger U^T$

6.1 Show that if A has full row rank, then we have:

$A^\dagger = A^T(AA^T)^{-1}$ and it has full column rank, then prove $A^\dagger = (A^TA)^{-1}A^T$

We know that if $A^\dagger$ is Moore-Penrose pseudoinverse of matrix A then it satisfies these conditions:

- A has full rank
- $AA^\dagger A = A$
- $A^\dagger AA^\dagger = A^\dagger$
- $(AA^\dagger)^T = AA^T$
- $(A^\dagger A)^T = A^\dagger A$

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

Now we start by the given definition:

A full rank then

AA^T is invertible

$A^T = V\Lambda U^{-1}$

$AA^T = U\Lambda V^TV\Lambda U^T = U\Lambda^2U^T$

$(AA^T)^{-1} = U\Lambda^{-2}U^T$

$A^T(AA^T)^{-1} = V\Lambda^{-1}U^T = A^\dagger$

$A^TA = V\Lambda U^TU\Lambda V^T = V\Lambda^2V^T$

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

6.2 Find the SVD decomposition of the matrix

$$A = \begin{bmatrix} 1 & 3 & 1 \\ 2 & -1 & 2 \end{bmatrix}$$

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

$$A^T A = \begin{bmatrix} 5 & 1 & 5 \\ 1 & 10 & 1 \\ 5 & 1 & 5 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 11 & 1 \\ 1 & 9 \end{bmatrix}$$

$$|A^T A - \lambda I| = 0$$

$$|AA^T - \lambda I| = 0$$

$$\lambda = 0, 10 + \sqrt{2}, 10 - \sqrt{2}$$

$$(AA^T - \lambda I)\nu = 0 \rightsquigarrow \nu = [\sqrt{2} + 1, 1]^T, [-\sqrt{2} + 1, 1]^T$$

$$(A^T A - \lambda I)\zeta = 0 \rightsquigarrow \zeta = [-1, 0.1]^T, [1, \sqrt{2}, 1]^T, [1, -\sqrt{2}, 1]^T$$

$$\Sigma = \begin{bmatrix} \sqrt{10 - \sqrt{2}} & 0 & 0 \\ 0 & \sqrt{10 + \sqrt{2}} & 0 \end{bmatrix}_{2\times 3}$$

$$V = \begin{bmatrix} 1/2 & 1/2 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 1/2 & 1/2 & \sqrt{2}/2 \end{bmatrix}_{3\times 3}$$

$$u_i = \frac{1}{\sigma_i} A v_i$$

$$U = \begin{bmatrix} \frac{2-3\sqrt{2}}{2\sqrt{10-\sqrt{2}}} & \frac{2+3\sqrt{2}}{2\sqrt{10+\sqrt{2}}} \\ \frac{4+\sqrt{2}}{2\sqrt{10-\sqrt{2}}} & \frac{4-\sqrt{2}}{2\sqrt{10+\sqrt{2}}} \end{bmatrix}_{2\times 2}$$

Sanity check shows that $A = U\Sigma V^T$

7 Vector differentiation

Prove the following vector differentiation formulas.

7.1 $\nabla_x(a^T x) = \nabla_x(x^T a) = a$

$$a^T x = x^T a = \sum_{i=1}^{n} a_i x_i$$

$$\nabla_x(a^T x) = \frac{\partial a^T x}{\partial x_j} = \frac{\partial}{\partial x} \sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} \frac{\partial}{\partial x_j} a_i x_i = a$$

7.2 $\nabla_x(Tr(xx^TA)) = \nabla_x(x^TAx) = (A + A^T)x$

$$xx^T = \begin{bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & \dots & x_1x_n \\ x_2x_1 & x_2x_2 & x_2x_3 & \dots & x_2x_n \\ x_3x_1 & x_3x_2 & x_3x_3 & \dots & x_3x_n \\ \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & & \\ x_nx_1 & x_nx_2 & x_nx_3 & \dots & x_nx_n \end{bmatrix}$$

$$Tr(xx^TA) = Tr(x^TAx) = x^TAx$$

$$Tr(xx^TA) = x^TAx = \sum_{j=1}^{n}\sum_{i=1}^{n} x_i a_{ij} x_j$$

$$\nabla_x Tr(xx^TA) = \nabla_x(x^TAx) = \frac{\partial x^TAx}{\partial x_k} = \frac{\partial(\sum_{j=1}^{n}\sum_{i=1}^{n} x_i a_{ij} x_j)}{\partial x_k} = \frac{\sum_{j=1}^{n}\partial\sum_{i=1}^{n} x_i a_{ij} x_j}{\partial x_i} + \frac{\sum_{i=1}^{n}\partial\sum_{j=1}^{n} x_i a_{ij} x_j}{\partial x_j}$$

$$\nabla_x^i = \sum a_{ij}x_j + \sum a_{ji}x_j = Ax + A^Tx = (A + A^T)x$$

7.3 Gradient without explicit differentiation!

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

For symmetric matrix X prove that:

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

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

$X=V\Lambda V^{-1}$

$X^{-1}=V\Lambda^{-1}V^{-1}$

$|X^{-1}|=\frac{1}{|X|}=\prod_{i=1}^{n}\lambda_i^{-1}$

$log(|X^{-1}|)=-\sum_{i=1}^{n}log(\lambda_i)$

$f(X+\Delta X)=-log(det(X+\Delta X))$

$X=Q\Lambda Q^T$

$X+\Delta X=Q(X+\Delta X)Q^T$

$det(X+\Delta X)=det(Q(\Lambda+\Delta\Lambda)Q^T)$

$=det(QQ^T)det(\Lambda+\Delta\Lambda)=det(\Lambda+\Delta\Lambda)$

$f(X+\Delta X)=-log(det(X+\Delta X))=-log(det(\Lambda+\Delta\Lambda))=-log\prod_{i=1}^{n}(\lambda_i+\delta\lambda_i)$

$=-\sum_{i=1}^{n}log(\lambda_i+\delta_i)$

$det(X)=det(Q\Lambda Q^T)=det(\Lambda)$

$f(x)=-log(det(\Lambda))=-log\prod_{i=1}^{n}\lambda_i=-\sum_{i=1}^{n}log(\lambda_i)$

$f(x+\Delta x)-f(x)=\sum_{i=1}^{n}log(\frac{\lambda_i}{\lambda_i+\delta\lambda_i})=-\sum_{i=1}^{n}log(1+\frac{\delta\lambda_i}{\lambda_i})$

$log(1+x)\sim x$

$$f(x+\Delta x) - f(x) = -\sum_{i=1}^{n} \frac{\delta\lambda_i}{\lambda_i}$$

$$< \nabla f, \Delta X > = tr(\nabla f^T Q \Lambda Q^T) = tr(\nabla f^T \Lambda)$$

$$-\sum_{i=1}^{n} \frac{\delta\lambda_i}{\lambda_i} = tr(\nabla f^T \Lambda)$$

$$\nabla_\Lambda f = \frac{-1}{\lambda} = \Lambda^{-1}$$

$$f = -log(det(X))$$

$$\nabla_x f = Q \nabla_\Lambda f Q^T$$

$$\nabla_x f = \nabla_x - log(det(X)) = Q(-\Lambda^{-1})Q^T = -X^{-1}$$

7.4 Gradient without explicit differentiation! Part 2

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

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

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

$$f(X) = tr(X^{-1}A)$$

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

$$= tr(X^{-1}A) - tr(X^{-1}\Delta X X^{-1}A)$$

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

$$< \nabla f, \Delta X > = tr(\nabla f^{T}\Delta X)$$

$$f(X+\Delta X) - f(X) = tr(-X^{-1}A X^{-1}\Delta X) = tr(\nabla f^{T}\Delta X)$$

$$\nabla f = (-X^{-1}A X^{-1})^{T}$$

$$\nabla_x tr(X^{-1}A) = -X^{-T}A^{T}X^{-T}$$

8 Matrix Frobenius Norm

8.1 Frobenius norm

First prove that:

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|A_{ij}|^2} = \sqrt{\sum_{i=1}^{r}\sigma_i(A)^2}$$

$A^T_{n\times m}A_{m\times n} = M_{n\times n}$

$$Tr\{A^TA\} = \sum_{i=1}^{n} m_{i,i} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ji}a_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ji}^2$$

Also we know that if matrix A is complex, the statement is true for $tr\{A^HA\}$

8.2

$A = U\Sigma V^T$

Where $\Sigma = diag(\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r)$j

Then we have from the properties of Frobenius norm:

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

$$||X||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|X_{ij}|^2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|U\Sigma V^T|_{jj}^2}$$

$$= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|(U\Sigma V^T)_{ij}|^2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|\sum_{k=1}^{r}(U_{ik}\Sigma_{kk}V_{kj}^T)|^2}$$

For each value of k, Σ_{kk} is a scalar so we can re write this equation as below:

$$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|\sum_{k=1}^{r}(U_{ik}V_{kj}^T\Sigma_{kk})|^2} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(\sum_{k=1}^{r}(U_{ik}V_{kj}^T\Sigma_{kk}))^2}$$

$$\sqrt{\sum_{k=1}^{r}(\sum_{j=1}^{n}\sum_{i=1}^{m}(U_{ik}V_{kj}^T\Sigma_{kk}))^2} = \sqrt{\sum_{k=1}^{r}\Sigma_{kk}^2(\sum_{j=1}^{n}\sum_{i=1}^{m}(U_{ik}V_{kj}^T))^2}$$

$$\rightsquigarrow U_{ik}^2 = 1, V_{jk}^2 = 1 \rightarrow \sqrt{\sum_{k=1}^{r}\Sigma_{kk}^2} = ||\Sigma||_F$$

8.3

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

$||A||_F^2 = \sigma_1^2 + \sigma_2^2 + \ldots \sigma_r^2$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_r \geq 0$

$\sigma_1^2 \leq \sum_{i=1}^{r} \sigma_i^2$

$\sigma_2 \leq \sigma_1$

$\sigma_3 \leq \sigma_1$

...

$\sigma_r \leq \sigma_1$

$||A||_F^2 \leq r\sigma_1$

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

9 Right or wrong!

9.1 TRUE

$A \in R^{m \times n}$, full rank

$B \in R^{n \times p}$, full rank

$AB = 0_{m \times p}$

Imgine $n \leq m$ then we have:

$rank(A) = min(m, n) = n$

$dim(C(A)) = min(n, m) = n$

$dim(N(A)) = 0$ and $AB = 0 \rightsquigarrow dim(N(A)) \neq 0$ and $B \neq 0$

So we conclude

$m \leq n$

$rank(A) = m$

$dim(N(A)) = n - m$

$n \leq p$

$C(B) \in N(A)$

$dim(C(B)) \leq dim(N(A))$

$rank(B) \leq n - m$

Then we assume $n \leq p$ and we have:

$n \leq p$

$rank(B) = n \leq n - m$

$m \leq 0$ and $m > 0$

So $p \leq n$ and

$rank(B) = p$

$p \leq n - m$

$p + m \leq n$

9.2 TRUE

Any matrix that has inverse is full rank:

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

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

$|A - I| \neq 0$

And $A - I$ is indeed full rank

9.3FALSE

in general eignvectors of AB and BA are not equal

$ABv = \lambda v$

$BABv = B\lambda v$

$BA(Bv) = \lambda(Bv)$

So we proved that eignvector of BA is Bv and $Bv = v$ does not always hold.

10 Calculating normalized eigenvectors from eigenvalues!

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

$\mathbf{S} = Q\Lambda Q^{-1}$ by eignvalue decomposition (we can perform it because of problem statement about A)

$$det(\mathbf{S}) = det(QQ^{-}1)det(\Lambda) = det(\Lambda) = \prod_{i=1}^{n} \lambda_i$$

10.2 prove $\mathbf{A}adj(\mathbf{A}) = det(\mathbf{A})I = adj(\mathbf{A})\mathbf{A}$ (Cramer's rule)

$$\mathbf{A}^{-1} = \frac{adj(\mathbf{A})}{det(\mathbf{A})} \rightsquigarrow \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \frac{adj(\mathbf{A})\mathbf{A}}{det(\mathbf{A})} = \frac{\mathbf{A}adj(\mathbf{A})}{det(\mathbf{A})}$$

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

10.3 prove $adj(\mathbf{A}) = \sum_{i=1}^{n} \{ \prod_{k=1,k\neq i}^{n} \lambda_k(\mathbf{A})\} v_i(\mathbf{A}) v_i(\mathbf{A})^H$

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} \rightarrow \mathbf{A}^{-1} = \mathbf{Q}^{-1}\Lambda^{-1}\mathbf{Q} = \mathbf{Q}^{-1}diag(\frac{1}{\lambda_i})\mathbf{Q} \rightsquigarrow \mathbf{Q}^{-1} = \mathbf{Q}^H : \mathbf{A}^{-1} = \mathbf{Q}^H diag(\frac{1}{\lambda_i})\mathbf{Q}$$

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

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

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

10.4 Now prove the identity:

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

$\lambda_i \in \mathbf{A}$ eignvalues set

$\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_n$

$adj(A) = ((-1)^{i+j} det(M_{ij}))$

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

$$A = \sum_{i=1}^{n} \lambda_i(A) v_i v_i^*$$

$$adj(A) = \sum_{i=1}^{n} (\prod_{k=1,j\neq i}^{n} \lambda_k(A)) v_i v_i^*$$

$|\lambda I - A - \hat{\lambda} I| = 0$

$|(\lambda + \hat{\lambda})I - A| = 0$

Which means eignvalues of $\lambda I - A$ would be $\lambda_A + \lambda$

And it's eign vectors:

$(\lambda I - A - (\lambda + \lambda_A)I)v = 0$

$(\lambda_A I - A)v = 0$

Which means it's eignvectors are identical to matrix A,

$$adj(\lambda I_n - A) = \sum_{i=1}^{n} (\prod_{k=1,k\neq i}^{n} (\lambda - \lambda_k(A)) v_i v_i^*$$

By setting $\rightsquigarrow \lambda = \lambda_i(A)$ we have:

$$adj(\lambda_i(A) I_n - A) = (\sum_{i=1}^{n} (\prod_{k=1,k\neq i}^{n} (\lambda_i(A) - \lambda_k(A)) v_i v_i^*$$

If one specializes to the case $\lambda = \lambda_i(A)$ for some $i = 1,2,...,n$ all but one summands of the right hand side vanish

$$adj(\lambda_i(A) I_n - A) = (\prod_{k=1,k\neq i}^{n} (\lambda_i(A) - \lambda_k(A))) |v_{i,j}|^2$$

By extracting out the jj component we have:

$$det(\lambda_i(A)I_{n-1} - M_j) = (\prod_{k=1,k\neq i}^{n} (\lambda_i(A) - \lambda_k(A))) \mid v_{i,j} \mid^2$$

$$\prod_{k=1,k\neq i}^{n} (\lambda_i(A) - \lambda_k(A))) \mid v_{i,j} \mid^2 = \prod_{k=1}^{n-1} (\lambda_i(A) - \lambda_k(M_{jj}))$$

Also by using characteristic polynomial of A we have

$$p_A : p_A(\lambda) = det(\lambda I_n - A) = \prod_{k=1}^{n} (\lambda - \lambda_k(A))$$

And $p_{M_j}(\lambda) = det(\lambda I_{n-1} - M_j) = \prod_{k=1}^{n-1} (\lambda - \lambda_k(M_j))$

Then by taking derivative at $\lambda_i(A)$

$$p_A^{'}(\lambda_i(A)) = \prod_{k=1,k\neq i}^{n} (\lambda_i(A) - \lambda_k(A)))$$

Then we have:

$$\mid v_{ij} \mid^2 p_A^{'}(\lambda_i(A)) = p_{M_j}(\lambda_i(A))$$

11 Optimization

Solving soft SVM problem:

The primal optimization problem of soft SVM is:

$$\text{minimize}_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1,2,...,n$$

$$\xi_i \geq 0, \quad i = 1,2,...,n$$

$$\mathscr{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n}\mu_i\xi_i$$

With $\alpha_i, \mu_i \geq 0$ as multipliers,

And KKT conditions are:

1. Stationary

$$\nabla_w \mathscr{L} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0$$

$$\frac{\partial \mathscr{L}}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial \mathscr{L}}{\partial \xi_i} = C - \alpha_i - \mu_i = 0, \quad i = 1,2,...,n$$

2. Primal feasibility

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1,2,...,n$$

$$\xi_i \geq 0, \quad i = 1,2,...,n$$

3. Dual Feasibility

$$\alpha_i \geq 0, \quad i = 1,2,...,n$$

$$\mu_i \geq 0, \quad i = 1,2,...,n$$

4. Complementary Slackness

$\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0, \quad i = 1,2,...,n$

$\mu_i \xi_i = 0, \quad i = 1,2,...,n$

If $\mu_i = 0$ or $\zeta_i = 0$ or $\alpha_i = 0$ $y_i(W^T x_i + w_o) + \zeta_i - 1 = 0$

If $\mu_i > 0$ then $\zeta_i = C$ and if $0 < \alpha_i < C$ $y_i(w^T x_i + w_0) - 1 = 0$

$\alpha_i = C$ $\mu_i = 0$ $\zeta_i \geq 0$ $y_i(w^T x_i + w_0) \leq 1$

$0 < \alpha_i < C$ $\mu_i > 0$ $\zeta_i = 0$ $y_i(w^T x_i + w_0) = 1$

$\alpha_i = 0$ $\mu_i = C$ $\zeta_i = 0$ $y_i(w^T x_i + w_0) \geq 1$