Mathematical Minimum (Answers)

Note : these are **math exercises** (nothing programming related) that will help you succeeding in this course and in MT1 !

Fun Begins

Ex 1 : Sums and Products

Sums :

1. $\sum_{i=0}^{N} 1 \ =$

2 . $\sum_{k=1}^{K} \sum_{t=1}^{T} \ 1 =$

3. $\sum_{k=1}^{K} \sum_{t=1}^{T} \ 0.5^k =$

4 . $\sum_{k=1}^{\infty} \sum_{t=1}^{T} 0.5^k \ =$

Solution :

1

$\sum_{i=0}^{N} 1 = \underbrace{1 + \ldots + 1}_{N+1 \text{ terms}} = N + 1$

2. $\sum_{k=1}^{K} \sum_{t=1}^{T} 1 = KT$

3

$\sum_{k=1}^{K} \sum_{t=1}^{T} (0.5)^k = \sum_{t=1}^{T} \left(\sum_{k=1}^{K} (0.5)^k \right) = T \left(0.5 \sum_{k=0}^{K-1} (0.5)^k \right) = T \frac{0.5\left(1-0.5^K\right)}{1-0.5} = T\left(1 - 0.5^K\right)$

where we have used the geometric sequence formula $\sum_{k=0}^{K-1} ar^k = \frac{a\left(1-r^K\right.}{1-r}$

4

$\sum_{k=1}^{\infty} \sum_{t=1}^{T} (0.5)^k = \sum_{t=1}^{T} \left(\sum_{k=1}^{\infty} (0.5)^k \right) = T \frac{0.5}{1-0.5} = T$

Recall the geometric series formula $\sum^{\infty} r^k = \frac{1}{1-r}$

Products :

The notation $\prod_{i=1}^{N} p_i$ denotes the product with N factors:

$$\prod_{i=1}^{N} p_i = p_1 p_2 \cdots p_N$$

1. $\prod_{i=1}^{M} \frac{1}{\theta} =$
2. $\prod_{k=1}^{K} \frac{k}{k+1} =$
3. $\ln\left(\prod_{k=1}^{K} e^{k}\right) =$

Solutions :

1

$$\prod_{i=1}^{M} \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^{M}$$

2. $\prod_{k=1}^{K} \frac{k}{k+1} = \frac{1}{2}\,\frac{2}{3} \cdots \frac{K-1}{K}\,\frac{K}{K+1} = \frac{1}{K+1}$

3

$$\ln\left(\prod_{k=1}^{K} e^{k}\right) = \sum_{k=1}^{K} k = 1 + 2 + \cdots + K = \frac{K(K+1)}{2}$$

If you struggle for Ex 1, watch these :

Arithmetic & Geometric series : https://www.khanacademy.org/math/precalculus

Notes on notations :

- Geometric summation starting from i=0 or i=1 will only have different numerator for finite sum, with one have power to N and one to N+1. How about infinite sum of the geometric series.. What is the difference in the numerator ?
  For geometric summation the index is not important. It is more the first term and the number of terms. The infinite sum is just the limit when K goes to infinity of the finite version.

- You can think of K and T as some given constants whereas k is a dummy variable.
  https://en.wikipedia.org/wiki/1/2_%2B_1/4_%2B_1/8_%2B_1/16_%2B_%E2%8B%AF

Ex 2 : Function Properties

For each of the following functions f(x) below :

Find its limits $\lim_{x \to \pm\infty} f(x)$ as x approachs $\pm\ \infty$.

Choose the values of x where $f\ (x)$ is differentiable, i.e. $f\ '(x)$ exists

Choose the values of x where $f\ (x)$ is also strictly increasing, i.e. $f\ '(x) > 0$

For $f(x)\ = \max(0, x)$

(If the limit diverges to infty, enter inf for ∞, and -inf for $-\ \infty$ lim

$$\lim_{x \to -\infty} f(x) = 0$$

$$\lim_{x \to +\infty} f(x) = \text{inf}$$

Choose the intervals of x where

f(x) differentiable: **f'(x)>0 :**

(The left column is for "f(x) differentiable", and the right one is where "f'(x)>0.)

$\cancel{x < 0}$	☐ $x < 0$
$x = 0$	☐ $x = 0$
$x > 0$	☒ $x > 0$

(Graph this function on a piece of paper!)

For $f(x)\ = \frac{1}{1+e^{-x}}$

(Enter inf for ∞ and similarly -inf for $-\ \infty$ if the limit diverges to infty.)

$$\lim_{x \to -\infty} f(x) = 0$$

$$\lim_{x \to +\infty} f(x) = 1$$

Choose the intervals of x where
f(x) differentiable: f'(x)>0

$x < 0$	$\square x < 0$
$x = 0$	$\square x = 0$
$x > 0$	$\square x > 0$

Notes on functions : you can use https://www.wolframalpha.com/ or https://www.derivativecalculator.net to calculate and plot derivatives !

- The domain of $f(x) = \max(0, x)$ is the entire real line, i.e. $(-\infty, \infty)$. It asks what the function value is when $x \to -\infty$

- The condition of max (0,x) means y=0 for x<0. So what is the slope for x<=0?
  A continuous function would be an unsegmented straight line or unbroken curve. Maybe another way to say it is that the formula that describes its behavior continues across all x. In this case the formula for x<0 is f(x)=0 and the formula for x>0 is f(x)=x. y=0 does not hold for all x. A continuous function would imply that f(x)=0 over all x, but that's not the case.

- why f(x)=max(0,x) isn't differentiable at x>0
  If this function is differentible at x=0, then the left-hand limit and the right-hand limit of the derivative of f(x) at x=0 to must be the same. If you draw a graph of this function, you can see that the graph is not "smooth" at x=0. If a function is differentiable, the its graph should be smooth.

- If a function is not differentiable at certain point x0 , then f′(x0) does not exist. This means you cannot make any statement about f′(x) at x0 , which naturally eliminates itself from answering questions like what x makes f′(x)>0

- A function which is differentiable and strictly increasing is not necessarily has positive derivative at all points: take for example f(x)=x^3, which is strictly increasing and differentiable at all points, but has derivative 0 at the origin

Ex 3 : Points and Vectors (Solution only)

Ex1 :

- Plugging into the equation for norm, we get that the length of $\begin{bmatrix} 0.4 \\ 0.3 \end{bmatrix}$ is equal to $\sqrt{0.4^2 + 0.3^2} = 0.5$. Notice that the ratio of x:y is 3: 4 so we can use 3: 4: 5 triangle to speed up our calculation to find the length of the vector.
- We do the same for $\begin{bmatrix} -0.15 \\ -0.2 \end{bmatrix}$

Using the second expression for dot product and rearranging, we get $\alpha = \cos^{-1} \frac{x \cdot y}{\|x\| \|y\|}$. Using the first expression for dot product and plugging it in we get that $\alpha = \cos^{-1} \frac{(0.4)(-0.15)+(0.3)(0.2)}{\sqrt{(0.4)^2+(0.3)^2}\sqrt{(-0.15)^2+(0.2)^2}}$

Ex 2 :

Based on the previous equations for the dot product, we find that the angle between $x^{(1)}$ and $x^{(2)}$ is:

$$\alpha = \cos^{-1} \frac{x^{(1)} \cdot x^{(2)}}{\left\|x^{(1)}\right\| \left\|x^{(2)}\right\|}$$

$$\alpha = \cos^{-1} \frac{a_1^2 - a_2^2 + a_3^2}{a_1^2 + a_2^2 + a_3^2}$$

$x^{(1)}$ is orthogonal to $x^{(2)}$ when $x^{(1)} \cdot x^{(2)} = 0$ or $a_1^2 - a_2^2 + a_3^2 = 0$

Ex 3 : x/norm(x)

Ex 4 :

- The definition of the projection of one vector onto another is the part of the first vector which points in the same direction as the second vector. Thus the projection of $x^{(1)}$ onto $x^{(2)}$ points in the direction of $x^{(2)}$
The vector has magnitude $\left\|x^{(1)}\right\| \cos \alpha$. From our previous result $\alpha = \cos^{-1} \frac{x^{(1)} \cdot x^{(2)}}{\left\|x^{(1)}\right\| \left\|x^{(2)}\right\|}$, the projection thus has magnitude $\frac{x^{(1)} \cdot x^{(2)}}{\left\|x^{(2)}\right\|}$. Plugging in our values for $x^{(1)}$ and $x^{(2)}$ we get $\frac{a_1^2 - a_2^2 + a_3^2}{\sqrt{a_1^2 + a_2^2 + a_3^2}}$ No
Hence, to find the final vector projection, we scale the unit vector in the direction of the vector projection, which is $\frac{x^{(2)}}{\left\|x^{(2)}\right\|}$ by the

So the answer x

Hence, to find the final vector projection, we scale the unit vector in the direction of the vector projection, which

length, $\|p_{x^{(1)}\to x^{(2)}}\|$. So the answer is $\|p_{x^{(1)}\to x^{(2)}}\| \frac{x^{(2)}}{\|x^{(2)}\|}$

Let's be a and b two (not necessary unit) vectors. We want to compute the vector c being the projection of a on b and its $1-2$ norm (or length).
Let's start from the lenght. We know from a well-known trigonometric equation that
$||c|| = ||a|| * \cos(\alpha)$, where α is the angle between the two vectors a and b
But we also know that the dot product $a \cdot b$ is equal to $||a|| * ||b|| * \cos(\alpha)$
By substitution we find that $||c|| = \frac{a \cdot b}{||b||}$. This quantity is also called the component of a in the direction of b.
To find the vector c we now simply multiply $||c||$ by the unit vector in the direction of b, $\frac{b}{||b||}$, obtaining $c = \frac{a \cdot b}{||b||^2} * b$
If b is already a unit vector, the above equations reduce to:
$||c|| = a \cdot b$ and $c = (a \cdot b) * b$

If you reading the https://en.wikipedia.org/wiki/Vector_projectionVector Projection article on Wikipedia, and you may be confused by the different formulas that all seem to be describing the same thing. In the opening paragraph, it says that the scalar projection is defined as:

$$a^1 = \| \mathbf{a} \| \cos\ \theta = \mathbf{a} \cdot \hat{\mathbf{b}} = \mathbf{a} \cdot \frac{\mathbf{b}}{\| \mathbf{b} \|}$$

However, in the section "Definitions in terms of a and b: Vector Projection", it says that the definition of vector projection of a onto b is:

$$\mathbf{a}_1 = \frac{\mathbf{a} \cdot \mathbf{b}}{\| \mathbf{b} \|^2} \mathbf{b} = \frac{\mathbf{a} \cdot \mathbf{b}}{\mathbf{b} \cdot \mathbf{b}} \mathbf{b}$$

The first one is a scalar projection, meaning it is only a number. The second is a vector projection, which represents the projected vector. a_1 is the length of a1.

Ex 4 : Probability Density Functions

BEFORE DOING THE NEXT EXERCICES, READ THESE :

https://ermongroup.github.io/cs228-notes/preliminaries/probabilityreview/

https://tutorial.math.lamar.edu/classes/calcii/probability.aspx

And watch these (-20 min) :

https://www.youtube.com/watch?v=QKA4HNEw3aY

https://www.khanacademy.org/math/statistics-probability/random-variables-statslibrary/random-variables-continuous/v/probability-density-functions

Let X be a **continuous** random variable with probability **density** function (pdf) $f_X(x)$

- Is the value of $f_X(x)$ always $\in$ [0,1]?

  Yes/No

- For $a < b$, $\int_a^b f_X(x)dx \in [0,1]$ and represents the probability that the value of X falls between a and b

  Yes/No

- Is the value of $f_X(x)$ always non-negative?

  Yes/No

- The value of integral $\int_{-\infty}^{\infty} f_X(x)dx$ of $f_X(x)$ from $-\infty$ to ∞ is a finite, undetermined value.

  Yes/No

Solution :

1. While probabilities are always between 0 and 1, the probability density function (PDF) is not the actual probability of observing a particular outcome. This is an important distinction from probability mass functions, the analog for discrete random variables. So the PDF can be greater than 1, but its integral, which gives the probability must always be $\in [0,1]$
2. Yes, by definition.
3. Yes, by definition $f_X(x) \geq 0$
4. The integral across a range (here, from $-\infty$ to ∞) is the total probability that X takes values in that range. since this range contains all possible values any random variable can take, by definition, not only is the integral finite, but since the total probability must be 1, the integral is always 1, i.e. $\int_{-\infty}^{\infty} p_X(x)dx = 1$

Notes :

- Note that PDF values are **probability density**, not **probability.** On top of this, "... probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to 1." Read more here : https://en.wikipedia.org/wiki/Probability_density_function

- The fourth question might be unclear. Providing $f_X(x)$ is valid PDF, you know what this integral is $\int_{-\infty}^{\infty} f_X(x)dx$ integrated to.

Example :

PMT9

Recall that the probability density of a *normal* or *Gaussian* distribution with mean µ and variance σ^2 is,

$$g(x) \equiv \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

While σ^2 denotes the variance, the *standard deviation* is σ. You may assume σ>0.

In statistics, one technique to fit a function to data is a procedure known as *maximum likelihood estimation (MLE)*. At the heart of this method, one needs to calculate a special function known as the *likelihood function*, or just the *likelihood*. Here is how it is defined :

Let $x_0, x_1, \dots, x_{n-1}$ denote a set of nn input data points. The likelihood of these data, $L(x_0, \dots, x_{n-1})$ is defined to be

$$L(x_0, \dots, x_{n-1}) \equiv \prod_{k=0}^{n-1} p(x_i)$$

Ex 5: Univariate Gaussians

Before doing these take a look at :

- For quantile qs - https://www.youtube.com/watch?v=TzKeCv4S7nY
- For normal distri qs
  - https://web.stanford.edu/class/archive/cs/cs109/cs109.1178/lectureHandouts/110-normaldistribution.pdf
- For last qs, calculator. READ the qs carefully. Standard deviation is not the same as variance
  - http://onlnestatbook.com/2/calculators/normal_dist.html (from Joseph_Yiin)
- https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring2014/readings/MIT18_05S14_Reading6a.pdf

A univariate Gaussian or normal distributions can be completely determined by its mean and variance.

Gaussian distributions can be applied to a large numbers of problems because of the central limit theorem (CLT). The CLT posits that when a large number of independent and identically distributed ((i.i.d.) random variables are added, the cumulative distribution function (cdf) of their sum is approximated by the cdf of a normal distribution.

Recall the probability density function of the univariate Gaussian with mean μ and variance σ^2, $\mathcal{N}(\mu, \sigma^2)$

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

- Probability review: PDF of Gaussian distribution

In practice, it is not often that you will need to work directly with the probability density function (pdf) of Gaussian variables. Nonetheless, we will make sure we know how to manipulate the (pdf) in the next two problems.

The pdf of a Gaussian random variable X is given by

$$f_X(x) = \frac{n}{3\sqrt{2\pi}} \exp\left(-\frac{n^2(x-2)^2}{18}\right)$$

then what is the mean μ and variance σ^2 of X?

μ =

σ^2 =

Solution :

Comparing

$$f_X(x) = \frac{n}{3\sqrt{2\pi}} \exp\left(-\frac{n^2(x-2)^2}{18}\right) = \frac{1}{(3/n)\sqrt{2\pi}} \exp\left(-\frac{(x-2)^2}{2(3/n)^2}\right)$$

with

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

yields $\mu = 2$ and $\sigma^2 = \frac{9}{n^2}$

Let $X \sim \mathcal{N}\ (\mu, \sigma^{\,2})$, i.e. the pdf of X is

$$f_X \qquad (x) = \frac{1}{} \exp\left(-\frac{(x-\mu)^2}{2}\right) \qquad \sigma\sqrt{2\pi}\ \ 2\sigma$$

Let Y=2X. Write down the pdf of the random variable Y. (Your answer should be in terms of y, σ and μ.) fY(y)=

Solution :

If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $Y = 2X \sim \mathcal{N}\left(2\mu, 4\sigma^2\right)$ by the following general properties of expectations and variance:

$$\begin{aligned} \mathbf{E}[2X] &= 2\mathbf{E}[X] \\ \mathrm{Var}[2X] &= 2^2\,\mathrm{Var}[X] = 4\,\mathrm{Var}[X] \end{aligned}$$

Therefore,

$$f_Y(y) = \frac{1}{2\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-2\mu)^2}{2\left(4\sigma^2\right)}\right)$$

- Argmax

Let $f_X(x; \mu, \sigma^2)$ denote the probability density function of a normally distributed variable X with mean μ and variance σ^2 What value of x maximizes this function?

Solution :

The answer is μ, the mean of the distribution. If you look at the graph of the standardized normal distribution, you see that the maximum is at 0, its mean. Any normal distribution with different mean or variance is simply a shifted (different mean) or stretched (different variance) version of this distribution, so our result holds for any normally distributed variable. Alternatively, you can differentiate the PDF and determine the maximum, which gives you the same result.

- Maximum of pdf

As above, let $f_X(x; \mu, \sigma^2)$denote the probability density function of a normally distributed variable X with mean μ and variance σ^2

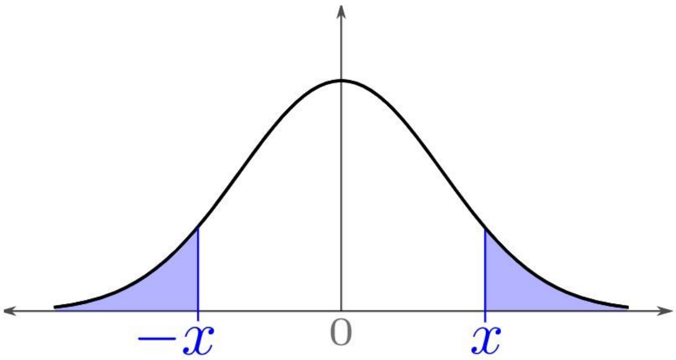
What is the maximum value of $f_X(x; \mu, \sigma^2)$

Solution :

From the question above, we know that the maximum value occurs when x=μ. Observe the PDF of a normal variable: setting x=μ forces the exponent of e to 0, leaving us with the answer above

- Quantile :

The quantile of order $1 - \alpha$ of a variable X, denoted by q_α i X) (specific to a particular X), is the number such that $\mathbf{P}(X \leq q_\alpha) = 1 - \alpha$



The total area of the two shaded regions is 0.03.

P(|X|≤0.03)

☐ P(|X|≤0.015)

☐ 0.97 ☐

0.985 ☐

q0.03

☐ q0.015

Let $X \sim \mathcal{N}(1,2)$, i.e., the random variable X is normally distributed with mean 1 and variance 2. What is the probability that

$X \in [0.5,2]$?

Solution :

The total area of the two shaded regions equals $\mathbf{P}(|X| \geq x) = 0.03$. By symmetry, the probability in the positive tail is $\mathbf{P}(X \geq x) = 0.015$; hence $x = q_\alpha$ with $\alpha = 0.015$

For the wrong choices:

- The first pair of choices mixed up the values of probability with the value of the variable.
- The third and fourth choices "0.97" and "0.985" are meant to play the role resembling $1 - \alpha$ in this example, but these are wrong for the same reasons as the first pair of choices. In any case, to give a particular numerical value of x, the answer must depend on σ.
- The fifth choice would have been correct again if the area of one of the tails is 0.03 .

Notes :

- There is no need to do integration in any of the problems. Recall what the Normal distribution is parametrized by and then do some pattern matching on the provided PDF.

- How to find the maximum of a normal PDF?
  One method would be to take find the derivative of the pdf and set it equal to zero. This gives you a critical point and then you can look at the sign of the derivative for other values and see if it is the maximum.
- On Quantiles. The curve is a probability density, it integrates to one (100%). If they say the blue area is 0.03 (3%), what is the probability that a random variable distributed like that falls into that blue region? x is the number that limits that slices that region. For example, if it was a standard normal, and the blue are integrated to 5%, x would result in 1.96, meaning that the probability of a r.v. X being less than 1.96 or higher than 1.96 is 5%. We call 1.96 the 2.5% quantile, given its a two tailed test. Use this to fiddle around
  http://onlinestatbook.com/2/calculators/normal_dist.html

- Notice the definition of $q\alpha$ is defined so that the **right** tail has probability α. This is different from how the quantile function is defined in python or R

- For the $Y=2X$ question, Note that the random variable is Y not X. So you should use Y instead of X. What kind of distribution does Y follow given by the expression Y = 2X? What is the description of the distribution? (i.e. parameters). Question 2 is actually quite easy to answer, given that X is a gaussian variable (mu, sigma). This, provided you went through a Probability course or read material regarding normal distributions somewhere else.
  You could also start with a CDF, convert it to a known CDF, then convert it to a PDF. You could, but there is no need to do something that sophisticated. Knowing that X is a gaussian variable of mean μ and variance σ2 has a direct implication on the nature of the distribution of Y=2X and the values of Y's mean and variance. Just remember that a linear combination of a gaussian is still a gaussian, and that if you multiply a random variable by a constant, its expected value is multiplied accordingly and its variance is multiplied by the square of that constant.

- For the last question you can use scipy :

```
from scipy.stats import norm
X = norm(3, scale=sqrt(9.))
X.cdf(10)
```

Ex 6 : Polynomials

Recall a degree n polynomial in x1,x2,…,xk are all linear combinations of monomials in x1,x2,…,xk , where monimials in x1,x2,…,xk are unordered words using x1,x_2,…, xk as the letters.

A degree 2, also known as quadratic, polynomial in the 1 variable x is of the form :

$$ax^2 + bx + c$$

for some numbers a,b,c . The polynomial is determined by the 3 coefficients a,b,c , and different choices of (a,b,c) result in different polynomials.

In linear algebraic terms, the space of degree 2 polynomials in 1 variable is of dimension 3 since it consists of all linear combinations of 3 linearly independent vectors x^2,x, and 1 .

A degree 2 polynomial in 2 variables x1,x2 is of the form :

$$ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f$$

or some numbers a,b,c,d,e,f . Different choices of (a,b,c,d,e,f) result in different polynomials.

In linear algebraic terms, the space of degree 2 polynomials in 2 variables is of dimension 6 since it consists of all linear combinations of 6 linearly independent vectors x_1^2, x_2^2, x_1x_2, x_1, x_2, and 1

Consider degree 2 polynomials in 3 variables x1,x2,x3 . How many coefficients are needed to completely determine such a polynomial? Equivalently, what is the dimension of the space of polynomials in 3 variables such polynomials?

Solution :

We count the number of monomials of length 2,1,0 :
The monomials of length 2 are unordered pairs of x_1, x_2, x_3, hence there are $\binom{3}{2}$ This list consists of $x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$
- The monomials of length 1 are x_1, x_2, x_3
- The monomial of length 0 is the constant term, i.e. 1 .

Notes :

- Here you want to collect all the ways you can combine the variables to create terms of degree ≤ n. Looking at example number 2, we have degree =n=2. We collect the following terms

  $n = 2$: $ax_1^2 = ax_1x_1$, $bx_2^2 = bx_2x_2$, cx_1x_2

  $n = 1$: dx_1, ex_2

  $n = 0$: f

  The equation for the given polynomial is the sum of all of these terms. I know that's not a resource so much as my own process for solving the problem, but hopefully it helps a bit. See Khan Academy videos on polynomials.

- Distributing the degree across the variables is analogous to distributing balls among urns, which is a classical combinatorial setting. https://artofproblemsolving.com/wiki/index.php/Ball-and-urn will give you a good intro. Be careful with the mapping of variables between the exercise and their explanation, in particular, consider how to account for unused parts of the degree in terms like x_1 or the constant term.

Check also these : https://math.stackexchange.com/questions/2705636/formula-for-discriminant-of-apolynomial-of-degree-2-in-3-variables
More intuitive : https://en.wikipedia.org/wiki/Triangular_number

- What is dimension of the polynomials of degree N in K variables? - proof versus heuristic patterning ?

By writing out cases up to N=3 and K=4, you can arrive at a simple combinatoric formula that matches all the cases, including the given examples and the graded exercise. What are the ways to show that it is true for arbitrary N and K?

There are a few ways to do this. One is to first restrict yourself to looking at degree-j polynomials for certain j, and notice that there is a constraint on the exponents of the variables. You can count the number of satisfying combinations using the stars and bars technique. Then you can add these up for all degree possibilities j, yielding a combinatorial sum which you may or may not recognize (you can simplify the sum by writing out a few terms and guessing the sum+using induction, or maybe a combinatorial argument, but I'm not aware how to do that). The wikipedia page for "complete homogeneous symmetric polynomial" also has the answer hidden in a lot of math and notation. But the answer is nice in terms of binomial coefficients.

There is also a very clever way to do it that bypasses the need to do a summation of binomial coefficients. The beauty of this way is that you don't have to look at degree-j polynomials one at a time and sum them up, you can count all the possibilities right from the beginning.

Ex 7 : Matrices and Vectors (not necessary for MT1)

(Chances that there are matrices in MT1 are very low but the subject is VITAL beyond MT1, so this just an appetizer) :

Let $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 2 & 1 \end{bmatrix}$

Let $g = [2 \quad 1 \quad 3]$

Can we compute $a\mathbf{A}$?

Solution :

The dimension of g is 1×3 and the dimension of A is 3×3. Since the number of columns in g equals the number of rows in A, the product exists.

Let g and A be as above. Can we compute Ag?

Let $\mathbf{B} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 4 \\ 5 & 6 & 4 \end{bmatrix}$

Solution :

Unlike part c), the dimension of A is 3×3 and the dimension of g is 1×3. Since the number of columns in A does not equal the number of rows in g, the product does not exist. Note that this example shows that matrix multiplication is not commutative, i.e., $\mathrm{AB}\neq\mathrm{BA}$.

Determine the rank of B. Recall that the rank of a matrix is the number of linearly independent rows or columns.

Solution :

Note that the first two rows of B are linearly independent since they are not multiples of each other. Now solve the system $\begin{bmatrix} 2a+b=5c \\ a+4b=6c \\ 4b=4c \end{bmatrix}$. Recall that these three vectors will be linearly independent if the only solution to this set of equations is the zero vector. since we find that this system has the solution $\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$, these vectors are not linearly independent and the rank of the matrix is 2 .

Let $\mathbf{M}^{-1}$ denote the inverse of a matrix M. Let A be as defined above. Compute $\mathbf{A}^{-1}$.

What matrix does the product $\mathbf{AA}^{-1}$ produce?

Solutions : For any matrix A, $\mathrm{AA}^{-1}=\mathrm{A}^{-1}\mathrm{A}=\mathrm{I}$, where I is the identity matrix.

Notes :

- The matrix A defined above is invertible, you could try to compute it via python or R.

- **How do quickly verify linear independence ?**
  For instance in Python you can use numpy with `np.np.linalg.matrix_rank(A)`), which is the simplest "mental model" to recognise linear independence in matrix rows/columns. However See there are others methods. First see if you can find a row or column that can be obtained by a linear combination of other columns.

For example if you had the matrix $\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 3 & 4 & 7 \end{pmatrix}$

you could (with some checking) notice that the last row is the sum of the first row and twice the second row (or that the third column is the sum of the first two columns) and thus the row or column vectors aren't independent. If the pattern isn't obvious with just short visual inspection, you could do it formally by setting up a system of equations: $\alpha[1,2,3]+\beta[1,1,2]+\gamma[3,4,7]=0$. If the only possible solution is $\alpha=\beta=\gamma=0$ then the vectors are independent, otherwise they're dependent (and in this case we have a solution $\alpha=1$, $\beta=2$, $\gamma=1$ so the vectors are dependent).

Another way to test for independence, which only works if you have a square matrix, is to compute the determinant of the matrix. If the determinant is 0 the vectors are dependent, otherwise they're independent.

Another method, in case it's helpful, is to reduce the matrix B to it's row echelon form (pretty much an upper triangular matrix with non zero entries on the diagonal). Once you've completed this task you count the pivots, which are the leftmost nonzero entry in each row where this entry has a 0 below it. The number of pivots equal the rank of the matrix.

This method is slower but can be used to get a definitive answer when you are having trouble confirming the existence or non-existence of linearly dependent rows/columns.

There are 3 pivots in the matrix below and it has a rank of 3.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 3 & 4 & 5 \\ 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Check UT Austin's LAFF linear algebra course, this is one of the best LA courses available !

To be continued….

1. Matrix Multiplication
2. Linear Independence, Subspaces and Dimension
3. Determinant
4. Eigenvalues, Eigenvectors and Determinants
5. 1D Optimization via Calculus
6. Gradients and Optimization