IFT 6390 Fundamentals of Machine Learning
Ioannis Mitliagkas

Homework 1 - Theoretical part

- This homework must be done and submitted to Gradescope individually. You are welcome to discuss with other students but the solution you submit must be your own. Note that we will use Gradescope's plagiarism detection feature. All suspected cases of plagiarism will be recorded and shared with university officials for further handling.

- You need to submit your solution as a pdf file on Gradescope using the homework titled `HW 1 - Theory`.

1. **Probability warm-up: conditional probabilities and Bayes rule** [5 points]

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable Y.

   (b) Consider a biased coin with probability 2/3 of landing on heads and 1/3 on tails. This coin is tossed three times. What is the probability that exactly two heads occur (out of the three tosses) given that the first outcome was a head?

   (c) Give two equivalent expressions of $P(X, Y)$:

   (i) as a function of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$

   (ii) as a function of $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$

   (d) Prove Bayes theorem:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

(e) A survey of certain Montreal students is done, where 55% of the surveyed students are affiliated with UdeM while the others are affiliated with McGill. A student is drawn randomly from this surveyed group.

i. What is the probability that the student is affiliated with McGill?

ii. Now let's say that this student is bilingual, and you know that 80% of UdeM students are bilingual while 50% of McGill students are. Given this information, what is the probability that this student is affiliated with McGill ?

2. **Bag of words and single topic model** [12 points]

We consider a classification problem where we want to predict the topic of a document from a given corpus (collection of documents). The topic of each document can either be *sports* or *politics*. 2/3 of the documents in the corpus are about *sports* and 1/3 are about *politics*.

We will use a very simple model where we ignore the order of the words appearing in a document and we assume that words in a document are independent from one another given the topic of the document.

In addition, we will use very simple statistics of each document as features: the probabilities that a word chosen randomly in the document is either "goal", "kick", "congress", "vote", or any another word (denoted by *other*). We will call these five categories the vocabulary or dictionary for the documents: $V = \{"goal", "kick", "congress", "vote", other\}$.

Consider the following distributions over words in the vocabulary given a particular topic:

	$\mathbb{P}(\text{word} \mid \text{topic} = sports)$	$\mathbb{P}(\text{word} \mid \text{topic} = politics)$
word = "*goal*"	3/200	8/1000
word = "*kick*"	1/200	2/1000
word = "*congress*"	0	1/50
word = "*vote*"	5/1000	2/100
word = *other*	960/1000	950/1000

Table 1:

This table tells us for example that the probability that a word chosen at random in a document is "vote" is only 5/1000 if the topic of the document is *sport*, but it is 2/100 if the topic is *politics*.

(a) What is the probability that a random word in a document is "goal" given that the topic is *politics*?

(b) In expectation, how many times will the word "goal" appear in a document containing 200 words whose topic is *sports*?

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?

(f) Going back to learning, suppose that you do not know the conditional probabilities given a topic or the probability of each topic (i.e. you don't have access to the information in table 1 or the topic distribution), but you have a dataset of N documents where each document is labeled with one of the topics *sports* and *politics*. How would you estimate the conditional probabilities (e.g., $\mathbb{P}(\text{word} = \text{"}goal\text{"} \mid \text{topic} = politics)$) and topic probabilities (e.g., $\mathbb{P}(\text{topic} = politics)$) from this dataset?

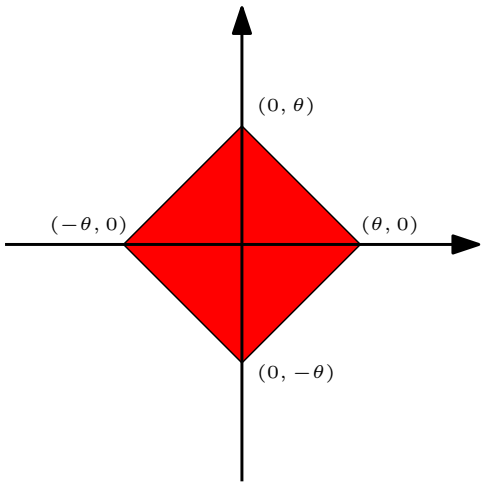
3. **Maximum likelihood estimation** [5 points]

Let $\mathbf{x} \in \mathbb{R}^2$ be uniformly distributed over a diamond area with diagonals 2θ where θ is a parameter as shown in the figure. That is, the pdf of $\mathbf{x}$ is given by

$$f_\theta(\mathbf{x}) = \begin{cases} 1/2\theta^2 & \text{if } ||\mathbf{x}||_1 \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $||\mathbf{x}||_1 = |x_1| + |x_2|$ is the L1 norm.

Suppose that n samples $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are drawn independently according to $f_\theta(\mathbf{x})$.



(a) Let $f_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ denote the joint pdf of n independent and identically distributed (i.i.d.) samples drawn according to $f_\theta(\mathbf{x})$. Express $f_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ as a function of $f_\theta(\mathbf{x}_1), f_\theta(\mathbf{x}_2), \ldots, f_\theta(\mathbf{x}_n)$

(b) We define the maximum likelihood estimate by the value of θ which maximizes the likelihood of having generated the dataset D from the distribution $f_\theta(\mathbf{x})$. Formally,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}^+} f_\theta(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n),$$

Find the maximum likelihood estimate of θ.

4. **Maximum likelihood meets histograms** [10 points]

Let $X_1, X_2, \cdots, X_n$ be n i.i.d data points drawn from a piece-wise constant probability density function over N equal size bins between 0 and 1 $(B_1, B_2, \cdots, B_N)$, where the constants are $\theta_1, \theta_2, \cdots, \theta_N$.

$$p(x; \theta_1, \cdots, \theta_N) = \begin{cases} \theta_j & \frac{j-1}{N} \leq x < \frac{j}{N} \text{ for } j \in \{1, 2, \cdots, N\} \\ 0 & \text{otherwise} \end{cases}$$

We define μ_j for $j \in \{1, 2, \cdots, N\}$ as $\mu_j \coloneqq \sum_{i=1}^{n} \mathbb{1}(X_i \in B_j)$.

(a) Using the fact that the total area underneath a probability density function is 1, express θ_N in terms of the other constants.

(b) Write down the log-likelihood of the data in terms of $\theta_1, \theta_2, \cdots, \theta_{N-1}$ and $\mu_1, \mu_2, \cdots, \mu_{N-1}$.

(c) Find the maximum likelihood estimate of θ_j for $j \in \{1, 2, \cdots, N\}$.

5. **Histogram methods** [10 points]

Consider a dataset $\{x_j\}_{j=1}^n$ where each point $x \in [0,1]^d$. Let $f(x)$ be the true unknown data distribution. You decide to use a histogram method to estimate the density $f(x)$ and divide each dimension into m bins.

(a) Show that for a measurable set S, $\mathbb{E}_{x\sim f}[\mathbb{1}_{\{x\in S\}}] = \mathbb{P}_{x\sim f}(x \in S)$, where $\mathbb{1}_{\{x\in S\}} = 1$ if $x \in S$ and 0 otherwise.

(b) Combining the result of the previous question with the Law of Large Numbers, show that the estimated probability of falling in bin i, as given by the histogram method, tends to $\mathbb{P}_{x\sim f}(x \in V_i) = \int_{V_i} f(x)dx$, the true probability of falling in bin i, as $n \to \infty$. V_i denotes the volume occupied by bin i.

(c) Consider the MNIST dataset with 784 dimensions (i.e. $x \in [0,1]^{784}$). We divide each dimension into 2 bins. How many digits (base 10) does the total number of bins have?

(d) Assuming a uniform distribution over all bins, how many data points would you need to get k points per bin on average?

(e) Assuming a uniform distribution over all bins, what is the probability that a particular bin is empty, as a function of d, m and n?

6. **Gaussian Mixture** [10 points]

Let $\mu_0, \mu_1 \in \mathbb{R}^d$, and let Σ_0, Σ_1 be two $d \times d$ positive definite matrices (i.e. symmetric with positive eigenvalues).
We now introduce the two following pdf over $\mathbb{R}^d$:

$$f_{\mu_0,\Sigma_0}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_0)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma_0^{-1}(\mathbf{x}-\mu_0)}$$

$$f_{\mu_1,\Sigma_1}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}$$

These pdf correspond to the multivariate Gaussian distribution of mean μ_0 and covariance Σ_0, denoted $\mathcal{N}_d(\mu_0, \Sigma_0)$, and the multivariate Gaussian distribution of mean μ_1 and covariance Σ_1, denoted $\mathcal{N}_d(\mu_1, \Sigma_1)$.

We now toss a balanced coin Y, and draw a random variable X in $\mathbb{R}^d$, following this process : if the coin lands on tails ($Y = 0$) we draw X from $\mathcal{N}_d(\mu_0, \Sigma_0)$, and if the coin lands on heads ($Y = 1$) we draw X from $\mathcal{N}_d(\mu_1, \Sigma_1)$.

(a) Calculate $\mathbb{P}(Y = 0|X = \mathbf{x})$, the probability that the coin landed on tails given $X = \mathbf{x} \in \mathbb{R}^d$, as a function of μ_0, μ_1, Σ_0, Σ_1, and $\mathbf{x}$. Show all the steps of the derivation.

(b) Recall that the Bayes optimal classifier is $h_{Bayes}(\mathbf{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \mathbb{P}(Y = y|X = \mathbf{x})$. Show that in this setting if $\Sigma_0 = \Sigma_1$ the Bayes optimal classifier is linear in $\mathbf{x}$.

(b) Recall that the Bayes optimal classifier is $h_{Bayes}(\mathbf{x}) = \underset{y \in \{0,1\}}{\text{argmax}}\, \mathbb{P}(Y = y|X = \mathbf{x})$. Show that in this setting if $\Sigma_0 = \Sigma_1$ the Bayes optimal classifier is linear in $\mathbf{x}$.

Answer 1.a)

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}.$$

Answer 1.b)
Given the formula above, let us first find $\mathbb{P}(X,Y)$, where:

X: Getting exactly two heads

Y: First throw is head

There are only two ways of getting exactly two heads when the first throw is an head: HHT or HTH. These two events each happen with probability $\frac{2}{3} * \frac{2}{3} * \frac{1}{3}$.

Thus, given that $\mathbb{P}(Y) = 2/3$, the answer is:

$$\frac{2 * \frac{2}{3} * \frac{2}{3} * \frac{1}{3}}{\frac{2}{3}} = \frac{4}{9}$$

Answer 1.c)

We can use 1.a and isolate to get the following results:

i)

$$\mathbb{P}(X,Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$$

ii)

$$\mathbb{P}(X,Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$$

Answer 1.d)

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$$

from the definition, and then

$$= \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}$$

from 1.c)

Answer 1.e)
i) 45%

ii) X: the student is from McGill

Y: The student is bilingual

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$$

$$= \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}$$

$$= \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y|X)\mathbb{P}(X) + \mathbb{P}(Y|X^c)\mathbb{P}(X^c)}$$

=

$$= \frac{0.45 * 0.5}{0.45 * 0.5 + 0.8 * 0.55} = 45/133$$

Answer 2.a)

$$\mathbb{P}(goal|politics) = 8/1000 = 1/125$$

Answer 2.b)

$$200 * \mathbb{P}(goal|sports) = 200 * 3/200 = 3 \text{ words on average}$$

Answer 2.c)

$$2/3 * 3/200 + 1/3 * 8/1000 = 19/1500$$

Answer 2.d)

$$\mathbb{P}(sports|kick) = \frac{2/3 * 1/200}{2/3 * 1/200 + 1/3 * 2/1000} = 5/6$$

Answer 2.e)

X: goal is second Y: Kick is first

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)} = \frac{\mathbb{P}(X,Y|sport) * \mathbb{P}(sport) + \mathbb{P}(X,Y|politics) * \mathbb{P}(politics)}{\mathbb{P}(Y)} =$$

$$\frac{2/3 * 1/200 * 3/200 + 1/3 * 2/1000 + 8/1000}{2/3 * 1/200 + 1/3 * 2/1000} = 83/6000$$

Answer 2.f)

P(word = x | topic = y) : Take only the y-labeled docs and divide the total number of times "x" appears across all y-labelled documents by the total number of words in said y-labelled documents.

P(topic) : just count how many documents are labeled with this topic and divide by the number of documents.

Answer 3a))

Since the variables are i.i.d., we can simply take the product.

$$f_\theta(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{2\theta^2} \mathbb{1}_{\{\|x_i\|_1 \leq \theta\}}$$

Which is equivalent to :

$$0 \text{ if } \exists i \mid \|x_i\|_1 > \theta$$

and otherwise:

$$\frac{1}{2^n \theta^{2n}} \text{ if } !\exists i \mid \|x_i\|_1 > \theta$$

Answer 3.b)

Given question 3.a), it's easy to see that we need our maximising θ to be no smaller than the biggest norm, or else the joint pdf evaluated on our sample points would be zero. We also have a strictly decreasing function in terms of θ for the interval of values of θ bigger than the biggest norm, which can trivially be determined with the derivative, which is negative for all positive

values of θ in that interval (the derivative being $\frac{-2n}{2^n \theta_i^{2n+1}}$ on that interval). Combining these two facts, we get that θ maximizes the likelihood when it is as small as it can be without being smaller than any of the norms of $\mathbf{x}_i$:

$$argmax_\theta(f_\theta) = max_i \|\mathbf{x}_i\|_1$$

Answer 4.a)
We need the area under the curve summing up to 1. Since it's a piece-wise constant function where every step is of width 1/N, that's just going to be the sum of a bunch of rectangles of width 1/N and of height θ_j:

$$\sum_{j=1}^{N-1} (\theta_j * \frac{1}{N}) + \theta_N * \frac{1}{N} = 1$$

Multiply by N on both sides, and send the sum to the right, and you get the result.

$$\theta_N = N - \sum_{j=1}^{N-1} \theta_j$$

Answer 4.b)

$$P(D_n|\theta) = \prod_{j=1}^{N} (\theta_j^{\mu_j})$$

$$= (N - \sum_{j=1}^{N-1} \theta_j)^{\mu_N} \prod_{j=1}^{N-1} (\theta_j^{\mu_j})$$

$$= (N - \sum_{j=1}^{N-1} \theta_j)^{n - \sum_{i=1}^{N-1} \mu_i} \prod_{j=1}^{N-1} (\theta_j^{\mu_j})$$

Taking the log, we get:

$$\log(P(D_n|\theta))$$

$$= (n - \sum_{j=1}^{N-1} \mu_j) \log(N - \sum_{j=1}^{N-1} \theta_j) + \sum_{j=1}^{N-1} \mu_j \log(\theta_j)$$

Answer 4.c)
A necessary condition is for the gradient to be zero. Taking the partial derivatives, we have:

$$\text{For all positive } i < N:$$
$$\frac{\partial \log(P(D_n|\theta))}{\partial \theta_i} = \frac{(n - \sum_{j=1}^{N-1} \mu_j)}{(N - \sum_{j=1}^{N-1} \theta_j)} * (-1) + \frac{\mu_i}{\theta_i} = 0$$

Rearranging, we get:
$$\frac{(N - \sum_{j=1}^{N-1} \theta_j)}{(n - \sum_{j=1}^{N-1} \mu_j)} = \frac{\theta_i}{\mu_i}$$

Using our answer to a), as well as the fact that the sum of all μ_j is n, we get:
$$\frac{\theta_N}{n - (n - \mu_N)} = \frac{\theta_i}{\mu_i}$$

Which, after rearranging and simplifying, gets us:
$$\mu_i \frac{\theta_N}{\mu_N} = \theta_i$$

Note that the sum of all θ is N, as one can easily see in a). Using that fact, we can find our answer:
$$N = \theta_N + \sum_{i=1}^{N-1} \theta_i = \theta_N + \sum_{i=1}^{N-1} \mu_i \frac{\theta_N}{\mu_N}$$

Putting it on the same denominator:
$$N = \frac{\theta_N}{\mu_N} \sum_{i=1}^{N} \mu_i = \frac{\theta_N}{\mu_N} n$$

And finally, rearranging, we get:
$$\theta_N = \frac{N \mu_N}{n}$$

By symmetry, or by substitution, we can easily see that:
$$\theta_i = \frac{N \mu_i}{n}$$

.

Answer 5.a)
By the definition of expectation, we sum over all the possible event values multiplied by their probability. The possible events of the indicator functions are 1 and 0.

$$\mathbb{E}_{x\sim f}[\mathbb{1}_{\{x\in S\}}] = \sum_{i=0}^{1} i * \mathbb{P}_{x\sim f}(\mathbb{1}_{\{x\in S\}} = i)$$

By expanding the sum, the i=0 term disappears, leaving us with:

$$\mathbb{E}_{x\sim f}[\mathbb{1}_{\{x\in S\}}] = \mathbb{P}_{x\sim f}(\mathbb{1}_{\{x\in S\}} = 1)$$

The probability that the indicator function of $x \in S$ is 1 is the same thing as asking the probability that x is in S

$$\mathbb{E}_{x\sim f}[\mathbb{1}_{\{x\in S\}}] = \mathbb{P}_{x\sim f}(\mathbb{1}_{\{x\in S\}} = 1) = \mathbb{P}_{x\sim f}(x \in S)$$

Answer 5.b)

Let our sample be called S. Let A be the event of an item falling in bin i. Let N be the sample size, and n_i be the number of items in bin i in our sample. The estimated probability of a new item falling in bin i is given by $\overline{A} = n_i/N$. Alternatively, we can rewrite this average like so: $\overline{A} = \frac{1}{N}\sum_{x\in S} \mathbb{1}_{\{x\in Vi\}}$. This is the average over all the sample of an indicator function acting as a random variable based on whether or not item x from the sample falls in V_i. By the law of large numbers, when the sample size N gets bigger, this becomes equal to the expectation of the random variable:

$$\lim_{N\to\infty} \overline{A} = \lim_{N\to\infty} \frac{1}{N} \sum_{x\in S} \mathbb{1}_{\{x\in Vi\}}$$

$$= \mathbb{E}_{x\sim f}[\mathbb{1}_{\{x\in Vi\}}]$$

By 5a), this is equivalent to our desired result:

$$= \mathbb{P}_{x\sim f}(x \in Vi)$$

Answer 5.c)
If we divide each bin in two, we will get 2^{784} dimensions. Changing the base exponent, we get:

$$10^{log_{10}(2^{784})} = 10^{784 log_{10}(2)} \approx 10^{236.008}$$

Thus, we see that the number of bins has 237 digits.

Answer 5.d)
To get k points per bin on average, we need

$$k * m^d$$

points, clearly. Proof (?): If $n = k * m^d$ is the number of points and $N = m^d$ is the number of bins, the average number of points per bin is : $n/N = k * m^d/m^d = k$

Answer 5.e)
The probability that a bin receives a specific item is :

$$1/m^d$$

Conversely, the probability that a bin does not receive a specific item is:

$$(m^d - 1)/m^d$$

If we have n items, the probability of all items not falling into our given bin is :

$$((m^d - 1)/m^d)^n$$

Answer 6.a)
We have that

$$f_x(x) = \sum_{y=0}^{1} f_{x,y}(X = x, Y = y) = f_{X|Y=0}(X|Y = 0) * \mathbb{P}_(Y = 0) + f_{X|Y=1}(X|Y = 1) * \mathbb{P}_(Y = 1)$$

$$= \frac{1}{2} f_{\mu_0,\Sigma_0}(\mathbf{x}) + \frac{1}{2} f_{\mu_1,\Sigma_1}(\mathbf{x})$$

We're interested in :

$$\mathbb{P}(Y = 0|X)$$

We can use the usual conditional density formula for this (even though Y is not continuous):

$$\mathbb{P}(Y = 0|X) = \frac{f_{X,Y}(X = x, Y = 0)}{f_X(X = x)}$$

$$= \frac{f_{X|Y=0}(X|Y=0) * \mathbb{P}_(Y=0)}{\frac{1}{2}f_{\mu_0,\Sigma_0}(\mathbf{x}) + \frac{1}{2}f_{\mu_1,\Sigma_1}(\mathbf{x})}$$
$$= \frac{\frac{1}{2}f_{\mu_0,\Sigma_0}(\mathbf{x})}{\frac{1}{2}f_{\mu_0,\Sigma_0}(\mathbf{x}) + \frac{1}{2}f_{\mu_1,\Sigma_1}(\mathbf{x})}$$
$$= \frac{f_{\mu_0,\Sigma_0}(\mathbf{x})}{f_{\mu_0,\Sigma_0}(\mathbf{x}) + f_{\mu_1,\Sigma_1}(\mathbf{x})}$$

We can simplify a bit, by substituing and removing the $(2\pi)^{d/2}$, but it doesn't add much to the solution.

Answer 6.b)
First, let's note that the classifier will pick y that gives the maximum value for:

$$P(Y = y|X = x)$$

Which, given 6.a above, combined with the statement of 6.b) about $\Sigma_0 = \Sigma_1$, and simplifying the determinants, we can change into:

$$argmax_{y\in 0,1}(\frac{e^{-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma_0^{-1}(\mathbf{x}-\mu_y)}}{e^{-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma_0^{-1}(\mathbf{x}-\mu_y)} + e^{-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma_0^{-1}(\mathbf{x}-\mu_y)}})$$

Crossing out the identical denominators (which are always positive, and thus do not affect the answer) we get:

$$argmax_{y\in 0,1}(e^{-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma_0^{-1}(\mathbf{x}-\mu_y)})$$

Next, taking the log, a monotonically increasing function which will not change the argmax, we get the following relationship:

$$argmax_{y\in 0,1}(-\frac{1}{2}(\mathbf{x}-\mu_y)^T\Sigma_0^{-1}(\mathbf{x}-\mu_y))$$

Expanding the equation, we get:

$$argmax_{y\in 0,1}(-\frac{1}{2}(\mathbf{x})^T\Sigma_0^{-1}(\mathbf{x})+\frac{1}{2}(\mathbf{x})^T\Sigma_0^{-1}(\mu_y)+\frac{1}{2}(\mu_y)^T\Sigma_0^{-1}(\mathbf{x})-\frac{1}{2}(\mu_y)^T\Sigma_0^{-1}(\mu_y))$$

Note that taking the transpose of the two middle terms would necessarily give the same singular value since the covariance matrix is symmetric*. Note also that the leading term is only in function of x, and can thus be removed, since it is the same for both values of y. With that in mind, we get this:

$$argmax_{y\in 0,1}((\mathbf{x})^T\Sigma_0^{-1}(\mu_y) - \frac{1}{2}(\mu_y)^T\Sigma_0^{-1}(\mu_y))$$

Which we can see, is linear in x. Using the fact that the covariance matrix is symmetric, We can rewrite it so that x is to the right, like so:

$$argmax_{y\in 0,1}(\mu_y^T\Sigma_0^{-1}(\mathbf{x}) - \frac{1}{2}\mu_y^T\Sigma_0^{-1}\mu_y)$$

*Proof that since the covariance matrix is always symmetric, the middle terms are equivalent:

$(\mathbf{x})^T\Sigma_0^{-1}(\mu_y) = K$, a scalar. Taking the transpose will not change the value, since it is a scalar. Taking the transpose on both sides, we get:

$(\mu_y)^T(\Sigma_0^{-1})^T(\mathbf{x}) = K$

And since the matrix Σ_0^{-1} is the inverse of a symmetric matrix (and thus symmetric itself), then we simply have: $(\mu_y)^T\Sigma_0^{-1}(\mathbf{x}) = K = (\mu_y)^T(\Sigma_0^{-1})^T(\mathbf{x})$