Machine Learning - Notes

Dom Hutchinson

October 9, 2019

Contents

General

Lecturer - Carl Henrik Ek
Course Website - `http://carlhenrik.com/COMS30007/`
Course Repo - `https://github.com/carlhenrikek/COMS30007`
Course Subreddit - `https://www.reddit.com/r/coms30007/`

1 Introduction

1.1 Motivation

Definition 1.1 - *Deductive Reasoning*
A method of reasoning in which the premieses are viewed as supplying all the evidence for the truth of the conclusion.

Definition 1.2 - *Inductive Reasoning*
A method of reasoning in which the premieses are viewed as supplying some evidence for the truth of the conclusion, rather than all the evidence. This allows for the conclusion of the *Inductive Reasoning* to be false.

Remark 1.1 - *Free-Lunch Theorem*
There are infinite number of hypotheses that perfectly explain the data. Adding a data point removes an infinite number of possibilities, but still leaves infinite possibilities.

Remark 1.2 - *The Task of Machine Learning*
When proposing to use machine learning on a task, one should consider the following questions:

i) How can we formulate beliefes ad assumptions mathematically?

ii) How can we connect our assumptions with data?

iii) How can we update our beliefs?

Remark 1.3 - *Useful Models are not always True*
Our goal is to understand realisations of a system. If we can then we can equate our model to the system. It is important to note thtat our model does not need to be perfectly true to be useful.

1.2 Probability Theory

Definition 1.3 - *Stochastic/Random Variable*
A variable whose value depends on outcomes of random phenomona.
e.g. $x \sim \mathcal{N}(0,1)$.

Definition 1.4 - *Probability Measure,*$\mathbb{P}$
A function with signature $\mathbb{P} : \mathcal{F} \to [0,1]$, where $\mathcal{F}$ is a sample space of rv X, and fulfils $\int_{-\infty}^{\infty} \mathbb{P}(x)dx = 1$.

Definition 1.5 - *Joint Probability Distribution*
A *Probability Measure* for multiple variables, $\mathbb{P} : X \times Y \to [0,1]$.
Let n_{ij} be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{\sum_{i,j} n_{ij}}$$

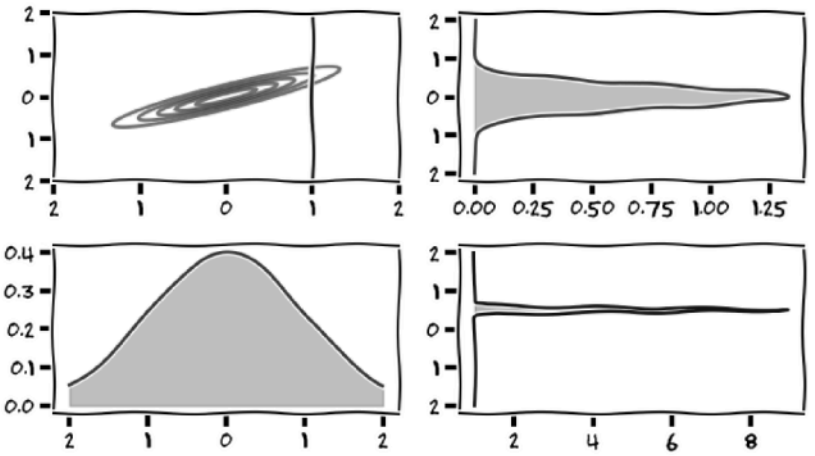
Definition 1.6 - *Marginal Probability Distribution*
A *Probability Measure* for one variable when the sample space is over multiple variables.
Let n_{ij} be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(X = x_i) = \frac{\sum_j n_{ij}}{\sum_{i,j} n_{ij}}$$

Definition 1.7 - *Conditional Probability Distribution*
A *Probability Measure* for a variable, given another variable has a defined value. Let n_{ij} be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(Y = y_j | X = x_i) = \frac{n_{ij}}{\sum_j n_{ij}}$$

Example 1.1 - *Joint, Marginal & Conditional Probability*
The below image shows two marginals distributions in the bottom-left, X, & top-right, Y, their joint distribution in the top-left and a conditional in the bottom right $\mathbb{P}(Y|X = 1)$.



Theorem 1.1 - *Product Rule*
For random variables X & Y

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

Theorem 1.2 - *Sum Rule*
For random variables X & Y

$$\mathbb{P}(X = x) = \sum_j \mathbb{P}(X = x, Y = y_j)$$

Theorem 1.3 - *Baye's Theorem*
For random variables X & Y

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

Definition 1.8 - *Elements of Bayes' Theorem*
The elements of *Bayes' Theory* can be broken down to explain parts of the model.

$$\underbrace{\mathbb{P}(\theta|Y)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(Y|\theta)}^{\text{Likelihood}}\ \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(Y)}_{\text{Evidence}}}$$

Posterior	Which parameters of the model do I belive produce distributions have generated the data Y
Likelihood	How likley is the data to come from the model specifely indexed by θ
Prior	What distribution do I think parameter θ has
Evidence	How likely do I think data Y is for all models.

N.B. The *Evidence* normalises this function.

Definition 1.9 - *Conjugate Prior*
If the *Prior* is the same probability distribution family as the *Posterior* then the *Prior* is called a *Conjugate Prior*.

Definition 1.10 - *Expectaction Value,* $\mathbb{E}$
The mean value a random variable will produce from a large number of samples.

Continuous	Discrete
$\mathbb{E}(X) = \int_{-\infty}^{\infty} x\mathbb{P}(X)dx$	$\mathbb{E}(X) = \sum_{-\infty}^{\infty} x\mathbb{P}(X)dx$
$\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$	$\mathbb{E}(f(X)) = \sum_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$

Definition 1.11 - *Variance*
Describes the amount of spread in the values a single random variable will produce.

$$\text{var}(X) = \mathbb{E}\left(x - \mathbb{E}(x))^2\right) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2$$

Definition 1.12 - *Covariance*
Describes the joint variability between two random variables.

$$\text{cov}(X, Y) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big)$$

Definition 1.13 - *Marginalisation*
The process of summing out the probability of one random variable using its joing probability with another rando variable.

$$\begin{array}{rrcl} \text{Continuous} & \mathbb{P}(X = x) & = & \int (X = x, Y = y)dy \\ \text{Discrete} & \mathbb{P}(X = x) & = & \sum_i \mathbb{P}(X = x, Y = y_i) \end{array}$$

Definition 1.14 - *Likelihood Function*
Define $\mathbf{X} \sim f_n(\cdot; \theta^*)$ for some unknown $\theta^* \in \Theta$ and let $\mathbf{x}$ be an observation of $\mathbf{X}$.
A *Likelihood Function* is any function, $L(\cdot; \mathbf{x}) : \Theta \to [0, \infty)$, which is proportional to the PMF/PDF of the obeserved realisation $\mathbf{x}$.

$$L(\theta; \mathbf{x}) := C f_b(\mathbf{x}; \theta) \ \forall \ C > 0$$

N.B. Sometimes this is called the *Observed* Likelihood Function since it is dependent on observed data.

Definition 1.15 - *Log-Likelihood Function*
Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ for some unknown $\theta^* \in \Theta$ and $\mathbf{x}$ be an observation of $\mathbf{X}$.
The *Log-Likelihood Function* is the natural log of a *Likelihood Function*

$$\ell(\theta; \mathbf{x}) := \ln f_n(\mathbf{x}; \theta) + C, \ C \in \mathbb{R}$$

Definition 1.16 - *Maximum Likelihood Estimation*
The *Maximum Likelihood Estimate* is an estimate for a parameter of a probability distribution which is the value which maximises the *Likelihood Function* (or the *Likelihood Function*).

$$\hat{\theta} := \text{argmax}_\theta L(\theta; \mathbf{x})$$

Definition 1.17 - *Central Limit Theorem*
The distribution of the sum (or mean) of a large number of independent, identically distributied random variables can be approximated to a normal distribution, regardless of the distributions of the random variables.

2 Distributions

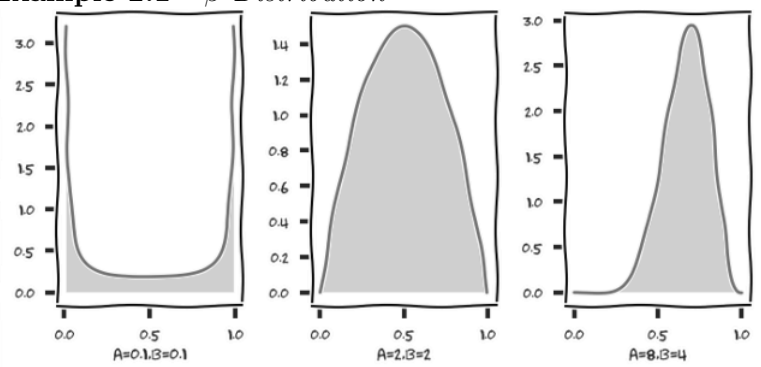
Definition 2.1 - *Bernoulli Distribution*
Models an event with a binary outcome (0 or 1) with parameter p st $\mathbb{P}(X = 1) = p$ Let $X \sim \text{Bernoulli}(p)$. Then

$$\begin{aligned} f_X(x) &= \begin{cases} p & ,x = 1 \\ 1-p & ,x = 0 \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &= \begin{cases} 0 & ,x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases} \\ \mathbb{E}(X) &= p \\ \text{Var}(X) &= p(1-p) \end{aligned}$$

Definition 2.2 - *β-Distribution*
A *β-Distribution* is a continuous distribution over interval $[0, 1]$ which is parameterised by two positive *shape parameters*, α & β. A *β-Distribution* can be used to encode assumptions as a *Prior*.
Let $X \sim \beta(\alpha, \beta)$. Then

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Example 2.1 - *β-Distribution*



Definition 2.3 - *Direchlet Distribution*
Let $X \sim \text{Dir}(\boldsymbol{\alpha})$. Then

$$f_X(x) := \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \times \cdots \times \Gamma(\alpha_N)} \prod_{i=1}^{N} x_i^{\alpha_{i-1}}$$

Definition 2.4 - *Exponential Distribution Family*
The *Exponential Distribution Family* is a set of probability distributions which fit the form.

$$\mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})e^{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x})}$$

With conjugate prior

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\chi})^{\nu} e^{\nu \boldsymbol{\theta}^T \boldsymbol{\chi}}$$

Definition 2.5 - *Multivariate Normal Distribution*
Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

N.B. Also known as *Gaussian Distribution*.

Remark 2.1 - *Conjugate Prior for Normal Distribution*
For a *Normal Distribution* the conjugate prior of $\boldsymbol{\mu}$ is *Gaussian* and the conjugate prior of $\boldsymbol{\Sigma}$ is *Inverse-Wishard*.

0 Appendix

Proof 0.1 - *Deriving Gaussian Marginal Distribution*
NOTE - This is dense as fuck & uses quite a bit of bullshit.

Let $\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}^{-1}\right)$.
$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ can be considered as two parts of the mean vector $\boldsymbol{\mu}$.
Let $\mathbf{x}$ be a realisation of $\mathbf{X}$ where $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2)$ with $\mathbf{x}_1$ & $\mathbf{x}_2$ representing the same partition as $\boldsymbol{\mu}_1$ & $\boldsymbol{\mu}_2$ respecitvely.
Define $D := \dim(\mathbf{x}),\ D_1 := \dim(\mathbf{x}_1)$ & $D_2 := \dim(\mathbf{x}_2)$.

Here we want to get from $\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2)$ to $\mathbb{P}(\mathbf{x}_1)$.
Consider the exponent of the joint distribution

$$E = -\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\Lambda_1 1(\mathbf{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\Lambda_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) - \frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Lambda_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Lambda_{22}(\mathbf{x}_2-\boldsymbol{\mu}_2)$$

To produce the marginal for x_1 we want to isolate the terms involving x_2 so they are easy to remove.

$$\begin{aligned}
E &= -\tfrac{1}{2}\Big[\left(\mathbf{x}_2^T\Lambda_{22}\mathbf{x}_2 - 2\mathbf{x}_2^T\Lambda_2 2(\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1))\right)\\
&- 2\mathbf{x}_1^T\Lambda_{12}\boldsymbol{\mu}_2 + 2\boldsymbol{\mu}_1^T\Lambda_{12}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T\Lambda_{22}\boldsymbol{\mu}_2 + \mathbf{x}_1^T\Lambda_{11}\mathbf{x}_1\\
&- 2\mathbf{x}_1^T\Lambda_1 1\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\Lambda_{11}\boldsymbol{\mu}_1\Big]\\
&= \underbrace{-\frac{1}{2}\left(\mathbf{x}_2 - (\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1))\right)^T\Lambda_{22}\left(\mathbf{x}_2 - (\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1))\right)}_{E_1}\\
&+ \underbrace{\frac{1}{2}\left(\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1\right)}_{A}\\
&- \underbrace{\frac{1}{2}\left(\mathbf{x}_1^T\Lambda_1 1\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{11}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1\Lambda_{11}\boldsymbol{\mu}_1\right)}_{B}
\end{aligned}$$

Note that A & B do not contain any x_2 terms.
Since the co-variance matrix is symmetric we have $\Lambda_{12} = \Lambda_{21}^T$ we have

$$\mathbf{x}_1^T\Lambda_{12}\boldsymbol{\mu}_2 = \mathbf{x}_1^T\Lambda_{21}^T\boldsymbol{\mu}_2 = (\Lambda_{21}\mathbf{x}_1)^T\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^T\Lambda_{21}\mathbf{x}_1$$

We shall not rewrite A & B as quadratic expressions

$$\begin{aligned}
A &= \frac{1}{2}\left(\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1\right)\\
&= \frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
B &= \frac{1}{2}\left(\mathbf{x}_1^T\Lambda_{11}\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{11}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1\Lambda_{11}\boldsymbol{\mu}_1\right)\\
&= \frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\Lambda_{11}(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
\implies A - B &= \frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} - \Lambda_{11})(\mathbf{x}_1-\boldsymbol{\mu}_1)\\
\text{Let } E_2 &:= A - B
\end{aligned}$$

Now the exponent has been orgainised we can consider the whole gaussian expression.

$$\begin{aligned}
\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) &= \frac{e^{E_1} e^{E_2}}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \\
\mathbb{P}(\mathbf{x}_1) &= \int \mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2 \\
&= \int \frac{e^{E_1} e^{E_2}}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} d\mathbf{x}_2 \\
&= \frac{e^{E_2}}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \int e^{E_1} d\mathbf{x}_2 \quad \text{Since } E_2 \text{ is independent of } \mathbf{x}_2
\end{aligned}$$

Now we consider $\int e^{E_1} d\mathbf{x}_2$.
Since we know a gaussian must intergrate to 1 over the whole domain we deduce that

$$\begin{aligned}
\int \frac{1}{(2\pi)^{\frac{D_2}{2}} |\Lambda_{22}^{-1}|^{\frac{1}{2}}} e^{E_1} d_{\mathbf{x}_2} &= 1 \\
\implies \int e^{E_1} d\mathbf{x}_2 &= (2\pi)^{\frac{D_2}{2}} |\Lambda_{22}^{-1}|^{\frac{1}{2}}
\end{aligned}$$

N.B. Λ_{22}^{-1} is the variance of $\mathbf{x}_2$.
Using the result of this intergal we have

$$\begin{aligned}
\mathbb{P}(\mathbf{x}_1) &= (2\pi)^{\frac{D_2}{2}} |\Lambda_{22}^{-1}|^{\frac{1}{2}} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{E_2} \\
&= \frac{e^{E_2}}{(2\pi)^{\frac{D-D_2}{2}} |\Lambda_{22}^{-1}|^{-\frac{1}{2}} |\Sigma|^{\frac{1}{2}}}
\end{aligned}$$

The Schur complement of Λ_{22} is $\Lambda_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.
Thus

$$\begin{aligned}
|\Lambda_{22}^{-1}|^{-\frac{1}{2}} |\Sigma|^{\frac{1}{2}} &= |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{-\frac{1}{2}} |\Sigma_{11}|^{\frac{1}{2}} |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{\frac{1}{2}} \\
&= |\Sigma_{11}|^{\frac{1}{2}}
\end{aligned}$$

Now we have a full expression

$$\begin{aligned}
\mathbb{P}(\mathbf{x}_1) &= \frac{e^{E_2}}{(2\pi)^{\frac{D-D_2}{2}} |\Lambda_{22}^{-1}|^{-\frac{1}{2}} |\Sigma|^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{D_1}{2}} |\Sigma_{11}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)}
\end{aligned}$$

■

Proof 0.2 - *Deriving Gaussian Conditional Distribution*

Let $\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right)$.
$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ can be considered as two parts of the mean vector $\boldsymbol{\mu}$.
Let $\mathbf{x}$ be a realisation of $\mathbf{X}$ where $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2)$ with $\mathbf{x}_1$ & $\mathbf{x}_2$ representing the same partition as $\boldsymbol{\mu}_1$ & $\boldsymbol{\mu}_2$ respecitvely.
Define $D := \dim(\mathbf{x})$.

We want to find the distribution of $\mathbb{P}(\mathbf{x}_1 | \mathbf{x}_2)$.
From the product rule we know that $\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{P}(\mathbf{x}_1 | \mathbf{x}_2)\mathbb{P}(\mathbf{x}_2)$ and we already know the joint & marginal distributions for a gaussian.
We have that

$$\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) \propto e^{-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}^T \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2\end{pmatrix}}$$

We now want to factor the marginal distribution out of this expression.

$$\mathbb{P}(\mathbf{x}_2) \propto e^{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)}$$

Lets look at the exponent of the joint distribution.
N.B. About to use a lot of Schur Complements

$$\begin{aligned}
E &= -\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\\
&= -\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T\begin{pmatrix}I & 0\\ \Sigma_{22}^{-1}\Sigma_{21} & I\end{pmatrix}^T\begin{pmatrix}(\Sigma/\Sigma_{22})^{-1} & 0\\ 0 & \Sigma_{22}^{-1}\end{pmatrix}\begin{pmatrix}I & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{pmatrix}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\\
&= -\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T\begin{pmatrix}(\Sigma/\Sigma_{22})^{-1} & -(\Sigma/\Sigma_{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{21}\Sigma_{22}^{-1}(\Sigma/\Sigma_{22})^{-1} & \Sigma_{22}^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\\
&= -\frac{1}{2}\Big[\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\Big]^T(\Sigma/\Sigma_{22})^{-1}\Big[\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\Big]\\
&\underbrace{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)}_{E_2}
\end{aligned}$$

Note that E_2 is exactly the exponent for the marginal distribution of $\mathbf{x}_2$ and thus what we want to factory out in order to get to the conditional distribution.

$$\mathbb{P}(\mathbf{x}_1|\mathbf{x}_2) \propto e^{-\frac{1}{2}\left[\mathbf{x}_1-\underbrace{(\boldsymbol{\mu}_1+\Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))}_{\text{mean}}\right]^T(\underbrace{\Sigma/\Sigma_{22}}_{\text{covariance}})^{-1}\left[\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\right]}$$

■