

# Study Notes

Yangyang Li  
[yangyang.li@northwestern.edu](mailto:yangyang.li@northwestern.edu)

Update on August 15, 2023



# Contents

	1.5.1	Variables Elimination . . . . .	16
	1.5.2	Belief propagation . . . . .	16
	1.5.3	Max-product Algorithm . . . . .	16
	1.5.4	Factor Graph . . . . .	16
	1.6	Expectation Maximum . . . . .	16
	1.7	Hidden Markov Model . . . . .	16
	<b>II</b>	<b>Algorithm and Data Structure</b>	<b>17</b>
	<b>2</b>	<b>Algorithm</b>	<b>19</b>
	2.1	Graph . . . . .	19
	<b>III</b>	<b>Programming</b>	<b>21</b>
	<b>3</b>	<b>C++</b>	<b>23</b>
	<b>4</b>	<b>Rust</b>	<b>25</b>
	<b>IV</b>	<b>Research</b>	<b>27</b>
	<b>5</b>	<b>Paper Reading</b>	<b>29</b>
		<b>Appendices</b>	<b>31</b>
		<b>Appendix A Formulas</b>	<b>31</b>
	A.1	Gaussian distribution . . . . .	31
		<b>Bibliography</b>	<b>33</b>
		<b>Alphabetical Index</b>	<b>35</b>
<b>Acronyms</b>			<b>7</b>
<b>Preface</b>			<b>9</b>
0.1		Features of this template . . . . .	9
0.1.1		crossref . . . . .	9
0.1.2		ToC (Table of Content) . . . . .	9
0.1.3		header and footer . . . . .	9
0.1.4		bib . . . . .	10
0.1.5		preface, index, quote (epi-graph) and appendix . . . . .	10
<b>I</b>		<b>Machine Learning</b>	<b>11</b>
<b>1</b>		<b>Probability</b>	<b>13</b>
1.1		Maximum Likelihood Estimation . . . . .	13
1.2		Maximum A Posteriori Estimation . . . . .	13
1.3		Gaussian Distribution . . . . .	14
1.4		Bayesian Network . . . . .	16
1.5		Probability Graph . . . . .	16



# List of Figures

1.1	A simple Bayesian network.	16
1.2	Belief propagation.	16

# List of Theorems

A.1	Theorem (Central limit theorem)	31
-----	---------------------------------	----

# List of Definitions

A.1	Definition (Gaussian distribution)	31
-----	------------------------------------	----



# Acronyms

MAP	Maximum A Posteriori Estimation <a href="#">13</a>
MLE	Maximum Likelihood Estimation <a href="#">13</a> , <a href="#">14</a>
PDF	Probability Density Function <a href="#">14</a>





# Preface

## Contents

<b>0.1 Features of this template</b> . . . . .	<b>9</b>
--	----------

### 0.1 Features of this template

*TeX, stylized within the system as  $\text{\LaTeX}$ , is a typesetting system which was designed and written by Donald Knuth and first released in 1978. TeX is a popular means of typesetting complex mathematical formulae; it has been noted as one of the most sophisticated digital typographical systems.*

- [Wikipedia](#)

#### 0.1.1 crossref

different styles of clickable definitions and theorems

- nameref: [Gaussian distribution](#)
- autoref: [Definition A.1](#), ??
- cref: Definition [A.1](#),
- hyperref: [Gaussian](#),

#### 0.1.2 ToC (Table of Content)

- mini toc of sections at the beginning of each chapter
- list of theorems, definitions, figures
- the chapter titles are bi-directional linked

#### 0.1.3 header and footer

fancyhdr

- right header: section name and link to the beginning of the section
- left header: chapter title and link to the beginning of the chapter
- footer: page number linked to ToC of the whole document

#### 0.1.4 bib

- titles of reference is linked to the publisher webpage e.g., [Kit+02]
- backref (go to the page where the reference is cited) e.g., [Chi09]
- customized video entry in reference like in [Bab16]

#### 0.1.5 preface, index, quote (epigraph) and appendix

*index* page at the end of this document...

**Part I**

**Machine Learning**



# Chapter 1

## Probability

### Contents

---

<a href="#">1.1 Maximum Likelihood Estimation</a>	13
<a href="#">1.2 Maximum A Posteriori Estimation</a>	13
<a href="#">1.3 Gaussian Distribution</a>	14
<a href="#">1.4 Bayesian Network</a>	16
<a href="#">1.5 Probability Graph</a>	16
<a href="#">1.6 Expectation Maximum</a>	16
<a href="#">1.7 Hidden Markov Model</a>	16

---

### 1.1 Maximum Likelihood Estimation

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)^T, \mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \quad (1.1)$$

in which  $N$  is the number of samples,  $p$  is the number of features. The data is sampled from a distribution  $p(\mathbf{x} | \theta)$ , where  $\theta$  is the parameter of the distribution.

For  $N$  i.i.d. samples, the likelihood function is  $p(\mathbf{X} | \theta) = \prod_{i=1}^N p(\mathbf{x}_i | \theta)$

In order to get  $\theta$ , we use [Maximum Likelihood Estimation \(MLE\)](#) to maximize the likelihood function.

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \log p(\mathbf{X} | \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log p(\mathbf{x}_i | \theta) \quad (1.2)$$

### 1.2 Maximum A Posteriori Estimation

In Bayes' theorem, the  $\theta$  is not a constant value, but  $\theta \sim p(\theta)$ . Hence,

$$p(\theta | \mathbf{X}) = \frac{p(\mathbf{X} | \theta)p(\theta)}{p(\mathbf{X})} = \frac{p(\mathbf{X} | \theta)p(\theta)}{\int_{\theta} p(\mathbf{X} | \theta)p(\theta)d\theta} \quad (1.3)$$

In order to get  $\theta$ , we use [Maximum A Posteriori Estimation \(MAP\)](#) to maximize the posterior function.

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} p(\theta | \mathbf{X}) = \underset{\theta}{\operatorname{argmax}} \frac{p(\mathbf{X} | \theta)p(\theta)}{p(\mathbf{X})} \quad (1.4)$$

After  $\theta$  is estimated, then calculating  $\frac{p(\mathbf{X} | \theta) \cdot p(\theta)}{\int_{\theta} p(\mathbf{X} | \theta) p(\theta) d\theta}$  to get the posterior distribution. We can use the posterior distribution to predict the probability of a new sample  $\mathbf{x}$ .

$$p(x_{\text{new}} | \mathbf{X}) = \int_{\theta} p(x_{\text{new}} | \theta) \cdot p(\theta | \mathbf{X}) d\theta \quad (1.5)$$

### 1.3 Gaussian Distribution

Gaussian distribution is also called normal distribution.

$$\theta = (\mu, \sigma^2), \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (1.6)$$

For MLE,

$$\theta = (\mu, \Sigma) = (\mu, \sigma^2), \quad \theta_{\text{MLE}} = \operatorname{argmax}_{\theta} \log p(\mathbf{X} | \theta) = \operatorname{argmax}_{\theta} \sum_{i=1}^N \log p(x_i | \theta) \quad (1.7)$$

Generally, the [Probability Density Function \(PDF\)](#) of a Gaussian distribution is:

$$p(x | \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right) \quad (1.8)$$

in which  $\mu$  is the mean vector,  $\Sigma$  is the covariance matrix,  $\det$  is the determinant of matrix.  $\det$  is the product of all eigenvalues of a matrix.

Hence,

$$\log p(\mathbf{X} | \theta) = \sum_{i=1}^N \log p(x_i | \theta) = \sum_{i=1}^N \log \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right) \quad (1.9)$$

Let's only consider 1 dimension case for brevity, then

$$\log p(\mathbf{X} | \theta) = \sum_{i=1}^N \log p(x_i | \theta) = \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right) \quad (1.10)$$

Let's get the optimal value for  $\mu$ ,

$$\mu_{\text{MLE}} = \operatorname{argmax}_{\mu} \log p(\mathbf{X} | \theta) = \operatorname{argmin}_{\mu} \sum_{i=1}^N \frac{1}{2} (x_i - \mu)^2 \quad (1.11)$$

So,

$$\frac{\partial \log p(\mathbf{X} | \theta)}{\partial \mu} = \sum_{i=1}^N (\mu - x_i) = 0 \rightarrow \mu_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.12)$$

Let's get the optimal value for  $\sigma^2$ ,

$$\begin{aligned}
\sigma_{\text{MLE}} &= \underset{\sigma}{\operatorname{argmax}} \log p(\mathbf{X} \mid \theta) \\
&= \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right) \\
&= \underset{\sigma}{\operatorname{argmax}} \sum_{i=1}^N \left[ -\log \sqrt{2\pi\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
&= \underset{\sigma}{\operatorname{argmin}} \sum_{i=1}^N \left[ \log \sigma + \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\end{aligned}$$

Hence,

$$\frac{\partial}{\partial \sigma} \sum_{i=1}^N \left[ \log \sigma + \frac{(x_i - \mu)^2}{2\sigma^2} \right] = 0 \rightarrow \sigma_{\text{MLE}}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (1.13)$$

$\mathbb{E}_D [\mu_{\text{MLE}}]$  is unbiased.

$$\mathbb{E}_D [\mu_{\text{MLE}}] = \mathbb{E}_D \left[ \frac{1}{N} \sum_{i=1}^N x_i \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_D [x_i] = \frac{1}{N} \sum_{i=1}^N \mu = \mu \quad (1.14)$$

However,  $\mathbb{E}_D [\sigma_{\text{MLE}}^2]$  is biased.

$$\mathbb{E}_D [\sigma_{\text{MLE}}^2] = \mathbb{E}_D \left[ \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right] \quad (1.15)$$

$$= \mathbb{E}_D \left[ \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right] \quad (1.16)$$

$$= \mathbb{E}_D \left[ \frac{1}{N} \sum_{i=1}^N (x_i^2 - 2x_i\mu_{\text{MLE}} + \mu_{\text{MLE}}^2) \right] = \mathbb{E}_D \left[ \sum_{i=1}^N x_i^2 - 2\frac{1}{N} \sum_{i=1}^N x_i\mu_{\text{MLE}} + \mu_{\text{MLE}}^2 \right] \quad (1.17)$$

$$= \mathbb{E}_D \left[ \frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu^2) + \mu^2 - \mu_{\text{MLE}}^2 \right] \quad (1.18)$$

$$= \sigma^2 - \mathbb{E}_D [\mu_{\text{MLE}}^2 - \mu^2] \quad (1.19)$$

$$= \sigma^2 - (\mathbb{E}_D [\mu_{\text{MLE}}^2] - \mathbb{E}_D [\mu_{\text{MLE}}^2]) \quad (1.20)$$

$$= \sigma^2 - \operatorname{Var} [\mu_{\text{MLE}}] = \sigma^2 - \operatorname{Var} \left[ \frac{1}{N} \sum_{i=1}^N x_i \right] \quad (1.21)$$

$$= \sigma^2 - \frac{1}{N^2} \sum_{i=1}^N \operatorname{Var} [x_i] = \frac{N-1}{N} \sigma^2 \quad (1.22)$$

$$(1.23)$$

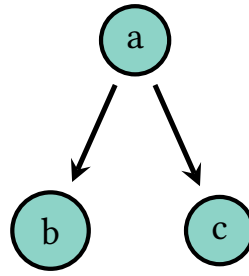


Figure 1.1: A simple Bayesian network.

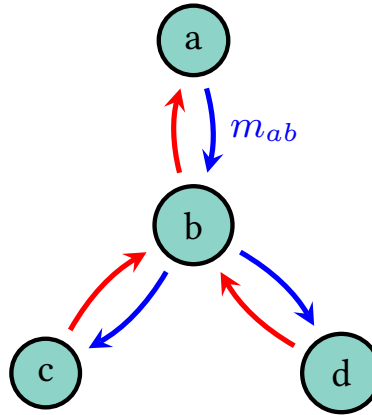


Figure 1.2: Belief propagation.

## 1.4 Bayesian Network

## 1.5 Probability Graph

this section is not finished yet. Need to be reviewed p54

### 1.5.1 Variables Elimination

### 1.5.2 Belief propagation

Belief propagation is mainly used for tree data structure, and equals Section 1.5.1 with caching.

### 1.5.3 Max-product Algorithm

### 1.5.4 Factor Graph

## 1.6 Expectation Maximum

$$\Theta^{(t+1)} = \operatorname{argmax}_{\Theta} \int_{\mathbf{z}} \log P(\mathbf{x}, \mathbf{z} \mid \theta) \cdot P(\mathbf{z} \mid \mathbf{x}, \Theta^{(t)}) d\mathbf{z}$$

continue on p60

## 1.7 Hidden Markov Model



## **Part II**

# **Algorithm and Data Structure**



## Chapter 2

# Algorithm

### Contents

---

<a href="#">2.1 Graph</a> . . . . .	<b>19</b>
-------------------------------------	-----------

---

## 2.1 Graph



# **Part III**

# **Programming**



## Chapter 3

### C++





## Chapter 4

# Rust



**Part IV**

**Research**



## **Chapter 5**

# **Paper Reading**



# Appendix A

## Formulas

### A.1 Gaussian distribution

**Definition A.1** (Gaussian distribution). *Gaussian distribution*

**Theorem A.1** (Central limit theorem).





# Bibliography

- [Bab16] László Babai. “Graph Isomorphism in Quasipolynomial Time”. Jan. 19, 2016. arXiv: [1512.03547](#) [[cs](#), [math](#)] (cit. on p. [10](#)). [ONLINE VIDEO](#)
- [Chi09] Andrew M. Childs. *Universal Computation by Quantum Walk*. Physical Review Letters 102.18 (May 4, 2009), p. 180501. arXiv: [0806.1972](#) (cit. on p. [10](#)).
- [Kit+02] Alexei Yu Kitaev et al. *Classical and quantum computation*. 47. American Mathematical Soc., 2002 (cit. on p. [10](#)).



# Alphabetical Index

**G**

Gaussian distribution . . . . . 31

**I**

index . . . . . 10