

Study Notes

Yangyang Li
yangyang.li@northwestern.edu

Update on August 12, 2023

Contents

	1.3	Gaussian Distribution	14
	1.4	Hidden Markov Model	15
	II	Algorithm and Data Structure	17
	2	Algorithm	19
	2.1	Graph	19
	III	Programming	21
	3	C++	23
	4	Rust	25
	IV	Research	27
	5	Paper Reading	29
		Appendices	31
		Appendix A Formulas	31
	A.1	Gaussian distribution	31
	Bibliography		33
	Alphabetical Index		35
Acronyms			7
Preface			9
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
I	Machine Learning		11
1	Probability		13
1.1		Maximum Likelihood Estimation . . .	13
1.2		Maximum A Posteriori Estimation . .	13

List of Figures

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 13
MLE	Maximum Likelihood Estimation 13 , 14
PDF	Probability Density Function 14

Preface

Contents

0.1 Features of this template	9
--	----------

0.1 Features of this template

TeX, stylized within the system as \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

1.1 Maximum Likelihood Estimation	13
1.2 Maximum A Posteriori Estimation	13
1.3 Gaussian Distribution	14
1.4 Hidden Markov Model	15

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 θ 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 θ , 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 θ 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 θ , 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 θ is estimated, then calculating $\frac{p(\mathbf{X} | \theta)p(\theta)}{\int_{\theta} p(\mathbf{X} | \theta)p(\theta)d\theta}$ 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,

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

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 μ is the mean vector, Σ 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 μ ,

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

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 σ^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.

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

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

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

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

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

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

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

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

$$(1.23)$$

1.4 Hidden Markov Model

Part II

Algorithm and Data Structure

Chapter 2

Algorithm

Contents

2.1 Graph	19
---------------------	----

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