Machine Learning: Mathematical Background

Statistics

Dariush Hosseini

dariush.hosseini@ucl.ac.uk
Department of Computer Science
University College London

Lecture Overview

1 Lecture Overview

2 Maximum Likelihood Estimation

3 PAC Setting

4 Bayesian Parameter Estimation

5 Estimation for Gaussian Random Variables

6 Summary

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations ⟶ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions ⟶ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically ⟶ **Probability**
  - Drawing conclusions based on the analysis of data ⟶ **Statistics**

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations ⟶ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data). To this end we are often interested in studying the continuous change of such functions ⟶ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically ⟶ **Probability**
  - Drawing conclusions based on the analysis of data ⟶ **Statistics**

Probabilistic Machine Learning

- The probabilistic setting is common in machine learning
- We will encounter this setting many times, so we should take the time to examine what learning looks like within it
- Typically (although not always) in this setting we suppose that the data is drawn from a **probability distribution** whose properties we seek to learn
- We use data to learn (**infer**) an **estimate** of the properties in question

Learning Outcomes for Today’s Lecture

By the end of this lecture you should:

1. Understand the **frequentist** approach to probabilistic learning
2. Have an awareness of the **PAC** approach to machine learning as one alternative to the traditional probabilistic approach
3. Understand the **Bayesian** approach to probabilistic learning

Lecture Overview

1. Lecture Overview
2. Maximum Likelihood Estimation
3. PAC Setting
4. Bayesian Parameter Estimation
5. Estimation for Gaussian Random Variables
6. Summary

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

[1] *Based on slides by Vibhav Gogate.*

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

- **He says:** I have this coin - what's the probability it lands heads up?

[1] *Based on slides by Vibhav Gogate.*

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

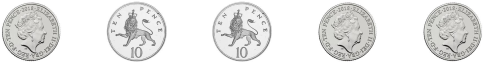
- **He says:** I have this coin - what's the probability it lands heads up?
- **You say:** Please flip it a few times

[1] *Based on slides by Vibhav Gogate.*

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

- **He says:** I have this coin - what's the probability it lands heads up?
- **You say:** Please flip it a few times

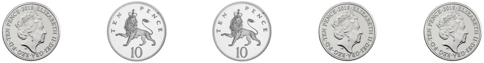


[1] *Based on slides by Vibhav Gogate.*

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

- **He says:** I have this coin - what's the probability it lands heads up?
- **You say:** Please flip it a few times



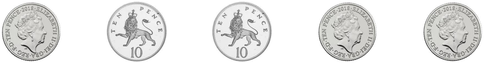
- **You say:** The probability is $\frac{3}{5}$

[1] *Based on slides by Vibhav Gogate.*

Your First Consulting Job [1]

Billionaire with more money than sense asks you a question:

- **He says:** I have this coin - what's the probability it lands heads up?
- **You say:** Please flip it a few times


TEN PENCE 2018 ELIZABETH II·DEI·GRA·REG·FID·DEF
TEN PENCE
10
TEN PENCE
10
TEN PENCE 2018 ELIZABETH II·DEI·GRA·REG·FID·DEF
TEN PENCE 2018 ELIZABETH II·DEI·GRA·REG·FID·DEF

- **You say:** The probability is $\frac{3}{5}$
- **He says:** Why?
- **You say:** Because...

[1] *Based on slides by Vibhav Gogate.*

Bernoulli Distribution

- **Coin flipping**:
  $\mathcal{X}$ is a random variable whose outcomes, $x \sim \mathcal{D}$, are those of a coin flip
  $$\Omega = \{\text{Tails} = 0, \text{Heads} = 1\}$$
  $$\text{Probability of 'Heads' is } \theta$$
- This characterises $\mathcal{D}$ as a **Bernoulli Distribution**

Representation

- **Task**
  - We want to predict the probability that a loaded coin will land on one side or the other, given some data, $\mathcal{S}$
- **Data**
  - $\mathcal{S}$ is a sequence of $n = 5$ coin flips, with outcomes $\{x^{(i)}\}_{i=1}^{5}$:

$$\mathcal{S} = \{x^{(1)} = 1, x^{(2)} = 0, x^{(3)} = 0, x^{(4)} = 1, x^{(5)} = 1\}$$

  - Each $x^{(i)}$ is the outcome of some random variable $\mathcal{X}_i$
- **Representation**
  - Each $\mathcal{X}_i$ is a random variable with outcomes drawn from the same Bernoulli distribution characterised by a single parameter, θ, such that $p_{\mathcal{X}_i}(x = 1; \theta) = \theta$
  - So the hypothesis space is characterised by $\theta \in [0, 1]$

Evaluation

If we assume that the coin flips are **independent and identically distributed** (**i.i.d.**) then:

- Probability theory tells us that the probability of a sequence $\mathcal{S}$, given a hypothesis θ, is:

$$\mathbb{P}(\mathcal{S} = \{x^{(i)}\}_{i=1}^{n}; \theta) = p_{\mathcal{X}}(\mathcal{S}; \theta) = \prod_{i=1}^{n} p_{\mathcal{X}_i}(x^{(i)}; \theta)$$

- Let this **likelihood** be our evaluation function, L (recall $\mathrm{L} : (\mathcal{E}, \mathcal{S}, f_\theta) \mapsto \mathrm{L}(\mathcal{E}, \mathcal{S}, f_\theta) \in \mathbb{R}$)

Optimisation

- So what's the best θ?
- It seems natural to maximise the probability of occurrence (the likelihood) of the $\mathcal{S}$ which we observed
- The **Maximum Likelihood Estimator** (**MLE**), θ_{MLE}, is the one which optimises the likelihood
- The **MLE** has many nice properties, but for our purposes it allows us to write:

$$\theta_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \left(\prod_{i=1}^{n} p_{\mathcal{X}_i}(x^{(i)}; \theta) \right)$$

Optimisation

- Number of Heads $= \alpha_H$
- Number of Tails $= \alpha_T$
- So:

$$\mathsf{L} = \prod_{i=1}^{n} p_{\mathcal{X}_i}(x^{(i)}; \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- And:

$$\begin{aligned}\theta_{\mathsf{MLE}} &= \operatorname*{argmax}_{\theta}\big(\theta^{\alpha_H}(1-\theta)^{\alpha_T}\big)\\ &= \operatorname*{argmax}_{\theta} \ln\big(\theta^{\alpha_H}(1-\theta)^{\alpha_T}\big)\\ &= \operatorname*{argmax}_{\theta}\big(\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\big)\end{aligned}$$

Optimisation

- We can solve this by computing the derivative...

$$\begin{aligned}\frac{d}{d\theta}\ln \mathrm{L} &= \frac{d}{d\theta}\left(\alpha_H \ln\theta + \alpha_T \ln(1-\theta)\right)\\ &= \alpha_H \frac{d}{d\theta}\ln\theta + \alpha_T \frac{d}{d\theta}\ln(1-\theta)\\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{(1-\theta)}\end{aligned}$$

- And setting it equal to zero:

$$\begin{aligned}\frac{\alpha_H}{\theta_{\mathrm{MLE}}} - \frac{\alpha_T}{(1-\theta_{\mathrm{MLE}})} &= 0\\ \frac{\alpha_H}{\theta_{\mathrm{MLE}}} &= \frac{\alpha_T}{(1-\theta_{\mathrm{MLE}})}\\ \theta_{\mathrm{MLE}} &= \frac{\alpha_H}{\alpha_H + \alpha_T}\end{aligned}$$

How Much Data?

- **You say:** ...$\alpha_H = 3$ and $\alpha_T = 2$, so $\theta = \frac{3}{5}$

How Much Data?

- **You say:** ...$\alpha_H = 3$ and $\alpha_T = 2$, so $\theta = \frac{3}{5}$
- **He says:** But what if I flipped 30 Heads and 20 Tails

How Much Data?

- **You say:** ...$\alpha_H = 3$ and $\alpha_T = 2$, so $\theta = \frac{3}{5}$
- **He says:** But what if I flipped 30 Heads and 20 Tails
- **You say:** Same answer

How Much Data?

- **You say:** ...$\alpha_H = 3$ and $\alpha_T = 2$, so $\theta = \frac{3}{5}$
- **He says:** But what if I flipped 30 Heads and 20 Tails
- **You say:** Same answer
- **He says:** But which one's better...And why?
- **You say:** More flips are definitely better!...Because...

Lecture Overview

1. Lecture Overview
2. Maximum Likelihood Estimation
3. PAC Setting
4. Bayesian Parameter Estimation
5. Estimation for Gaussian Random Variables
6. Summary

Hoeffding's Inequality

- Recall Hoeffding's Inequality:
- Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent random variables, such that each $\mathcal{X}_i$ is bounded by the interval $[a_i, b_i]$.
  Let the empirical mean of these random variables be defined as: $\overline{\mathcal{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i$.
  Then:

  $$\mathbb{P}\left(|\overline{\mathcal{X}} - \mathbb{E}[\overline{\mathcal{X}}]| \geqslant t\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

  Where $t \geqslant 0$

Hoeffding's Inequality

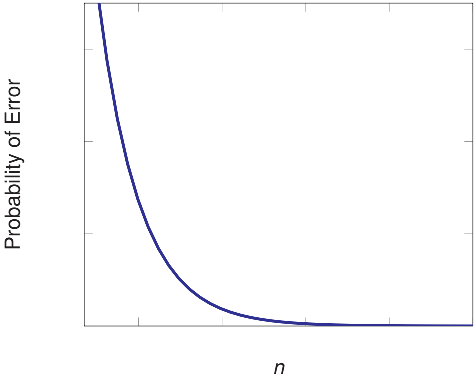
- If we identify each $\mathcal{X}_i$ with our coin flipping random variables, then: $\mathbb{E}[\overline{\mathcal{X}}] = \theta$; $a_i = 0$; $b_i = 1$, and θ_{MLE} is an outcome of the random variable $\overline{\mathcal{X}}$. Thus:

- The probability that the error in our parameter estimate, θ_{MLE}, versus the true parameter, θ, is greater than some constant, ϵ, can be bounded as follows:

$$\mathbb{P}(|\theta_{\text{MLE}} - \theta| \geqslant \epsilon) \leqslant 2e^{-2n\epsilon^2}$$

Hoeffding’s Inequality

- In other words, as $n = \alpha_H + \alpha_T$ increases, the probability of error decreases exponentially:



PAC Learning

- **He says:** OK, so I want to know θ, within $\epsilon = 10\%$, with probability of mistake, $\delta \leqslant 5\%$. How many flips do I need to make?

PAC Learning

- **He says:** OK, so I want to know θ, within $\epsilon = 10\%$, with probability of mistake, $\delta \leqslant 5\%$. How many flips do I need to make?
- **You say:** You need at least 184 flips in order to know that you are:
  - **probably** (with probability greater that $1 - \delta$)
  - **approximately** (with error less than ϵ) **correct**

PAC Learning

- **From Hoeffding and from the PAC requirement:**

$$\mathbb{P}(|\theta - \theta_{\text{MLE}}| \geqslant \epsilon) \leqslant 2e^{-2n\epsilon^2} \leqslant \delta$$

- **Combining these:**

$$\begin{aligned}
\delta &\geqslant 2e^{-2n\epsilon^2} \\
\ln\delta &\geqslant \ln 2 - 2n\epsilon^2 \\
n &\geqslant \frac{\ln(\frac{2}{\delta})}{2\epsilon^2} \\
&= \frac{\ln(\frac{2}{0.05})}{2(0.1)^2} \\
&\approx 184
\end{aligned}$$

PAC Learning

- This is the **PAC** or **Probably Approximately Correct** setting:
  - Here we try to generate a **probabilistic bound** for some quantity of interest such that it is **approximately** known with **high probability.**
  - This whole approach is a branch of **Computation Learning Theory.**
  - It defines a principled framework for algorithm generation:

PAC Learning

1. Develop bounds on certain evaluation functions applied to unseen data.
   - Often we assume nothing about the underlying probability distribution of the data but that it is i.i.d.!
2. Optimise these bounds and thus define a learning algorithm.

Lecture Overview

1 Lecture Overview

2 Maximum Likelihood Estimation

3 PAC Setting

4 Bayesian Parameter Estimation

5 Estimation for Gaussian Random Variables

6 Summary

Prior Beliefs

- **He says:** Actually, I have a strong **prior belief** that θ is ‘close’ to 0.5. Can you incorporate this somehow?

Prior Beliefs

- **He says:** Actually, I have a strong **prior belief** that θ is ‘close’ to 0.5. Can you incorporate this somehow?

- **You say:** Of course: We could apply **Bayesian Learning** and instead of regarding θ as a quantity which is **fixed** but **unknown**, we could regard it as a **random variable**, Θ, with its own probability distribution, which we seek to **update** as we receive more data...

Prior Beliefs

- **He says:** Actually, I have a strong **prior belief** that θ is 'close' to 0.5. Can you incorporate this somehow?

- **You say:** Of course: We could apply **Bayesian Learning** and instead of regarding θ as a quantity which is **fixed** but **unknown**, we could regard it as a **random variable**, Θ, with its own probability distribution, which we seek to **update** as we receive more data...

- **You say:** ...We can then use this updated **posterior belief** to make a point estimate θ...

Bayesian Learning

- Uses **Bayes' Rule**:

$$\begin{aligned} p_{\Theta}(\theta|\mathcal{S}) &= \frac{p_{\mathcal{X}}(\mathcal{S}|\theta)p_{\Theta}(\theta)}{p_{\mathcal{X}}(\mathcal{S})} \\ &\propto p_{\mathcal{X}}(\mathcal{S}|\theta)p_{\Theta}(\theta) \end{aligned}$$

- $p_{\Theta}(\theta)$ is the **Prior Probability** distribution function:
  - Represents a prior degree of belief about θ
- $p_{\mathcal{X}}(\mathcal{S}|\theta)$ is the **Likelihood** (i.e. $p_{\mathcal{X}}(\mathcal{S};\theta)$):
  - Which we encountered in MLE

Bayesian Learning

- $p_{\mathcal{X}}(\mathcal{S})$ is the **Evidence**:
  - A normalisation factor which is constant wrt θ

$$p_{\mathcal{X}}(\mathcal{S}) = \int p_{\mathcal{X}}(\mathcal{S}|\theta')p_{\Theta}(\theta')d\theta'$$

- $p_{\Theta}(\theta|\mathcal{S})$ is the **Posterior Probability** distribution function:
  - Represents a degree of belief about θ *after* observing $\mathcal{S}$
  - We use Bayes' Rule as a mechanism to generate this quantity
  - We will use it to perform point estimation of θ

Which Probability Distributions?

- The Likelihood has a **Binomial Distribution** (up to a constant)
- For the Prior and the Posterior we would like a closed form representation
- Certain forms of the Likelihood, when combined with **Conjugate Prior** distributions, guarantee closed form representations for the Posterior distribution which share the same form as the Prior...
- ...For a Binomial Likelihood, the Conjugate Prior is the **Beta Distribution**, which guarantees that the Posterior is also Beta distributed.

The Beta Distribution

- Let $\mathcal{Y}$ be a continuous random variable, taking values $y \in [0, 1]$, with a Beta distribution:

$$y \sim \text{Beta}(a, b) \qquad \text{where:} \qquad a, b > 0$$

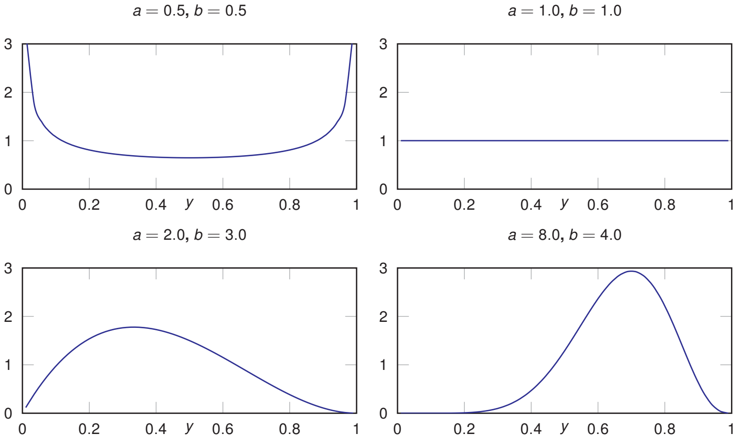
- This has a characteristic pdf $f_{\mathcal{Y}}$:

$$f_{\mathcal{Y}}(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \quad \text{where:} \quad \Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{Y}] = \frac{a}{a+b}$$

$$\text{Var}_{\mathcal{D}}[\mathcal{Y}] = \frac{ab}{(a+b)^2(a+b+1)}$$

The Beta Distribution



Posterior Distribution

- We model the Likelihood with a Binomial distribution:

$$p_{\mathcal{X}}(\mathcal{S}|\theta) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- We model the Prior with a Beta(β_H, β_T) distribution:

$$\theta \sim \text{Beta}(\beta_H, \beta_T)$$

$$p_{\Theta}(\theta) = \frac{\Gamma(\beta_H + \beta_T)}{\Gamma(\beta_H)\Gamma(\beta_T)}\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

Posterior Distribution

- Using Bayes' Rule we derive the following Posterior:

$$
\begin{aligned}
p_{\Theta}(\theta|\mathcal{S}) &\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\frac{\Gamma(\beta_H+\beta_T)}{\Gamma(\beta_H)\Gamma(\beta_T)}\theta^{\beta_H-1}(1-\theta)^{\beta_T-1} \\
&\propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}
\end{aligned}
$$

- From our knowledge of the Beta distribution:

$$
p_{\Theta}(\theta|\mathcal{S}) = \frac{\Gamma(\alpha_H+\beta_H+\alpha_T+\beta_T)}{\Gamma(\alpha_H+\beta_H)\Gamma(\alpha_T+\beta_T)}\theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}
$$

$$
\theta|\mathcal{S} \sim \text{Beta}(\alpha_H+\beta_H, \alpha_T+\beta_T)
$$

Posterior Distribution

- So our **Bayesian Update** is:

$$\text{Prior Distribution} \longrightarrow \text{Posterior Distribution}$$

$$\theta \sim \text{Beta}(\beta_H, \beta_T) \longrightarrow \theta|\mathcal{S} \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Bayesian Inference

- So how do we learn θ?
- Well, we don't need to:
  - Remember θ is now a random variable, Θ
  - We should now perform our reasoning in a different way:
- We should reason using the full posterior distribution:
  - Take an expectation over whatever quantity we are interested in, e.g. $g(\Theta)$:

$$\mathbb{E}_\Theta[g(\Theta)|\mathcal{S}] = \int_{-\infty}^{\infty} g(\theta')p_\Theta(\theta'|\mathcal{S})d\theta'$$

Bayesian Inference

- This makes sense conceptually...
    - But this integral is often hard to compute...
- Is there an alternative?
- **Maximum A Posteriori** (or **MAP**) Approximation

MAP Approximation

- This is a sort of halfway house in Bayesian inference
- We seek a point estimate of θ...
- ...And we find it by using the posterior distribution and looking for its maximal value.
- The result is the **MAP estimator**, θ_{MAP}:

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} \left(p_{\Theta}(\theta|\mathcal{S}) \right)$$

- And we proxy our Bayesian reasoning as follows:

$$\mathbb{E}_{\Theta}[g(\Theta)|\mathcal{S}] \approx g(\theta_{\text{MAP}})$$

MAP for the Beta Distribution

- Recall, the form for our posterior distribution:

$$p_{\Theta}(\theta|\mathcal{S}) = \text{const.} \times \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

- So (recalling our MLE optimisation):

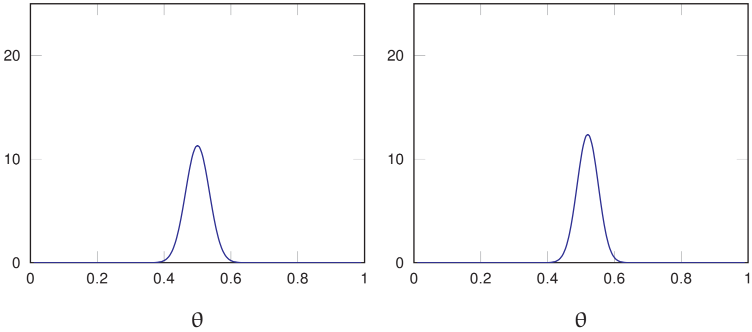
$$\theta_{\text{MAP}} = \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

MAP for the Beta Distribution

- If we have a strong prior belief, then $\beta_H + \beta_T$ will be large, and the form of the prior distribution will dominate the posterior initially...
- ...But as more data is gathered, then α_H, α_T will begin to dominate.
  - This phenomenon is referred to as the **washing out of the priors**
- If we have no strong prior belief, then we might select a **uniform prior**, for which $\beta_H = \beta_T = 1$.
  - For this case: $\theta_{\text{MAP}} = \theta_{\text{MLE}}$

Example: Strong Prior Belief

- $\beta_H = 100, \beta_T = 100$
  $\alpha_H = 30, \alpha_T = 20$



Prior: $\theta \sim \text{Beta}(100, 100)$

Posterior: $\theta \sim \text{Beta}(130, 120)$

$\theta_{\text{MAP}} = 0.52$

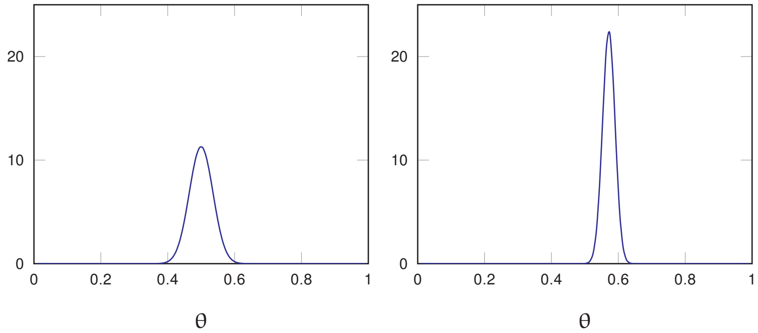
Example: Washing Out Of Priors

- $\beta_H = 100$, $\beta_T = 100$
  $\alpha_H = 300$, $\alpha_T = 200$

Prior: $\theta \sim \text{Beta}(100, 100)$

Posterior: $\theta \sim \text{Beta}(400, 300)$

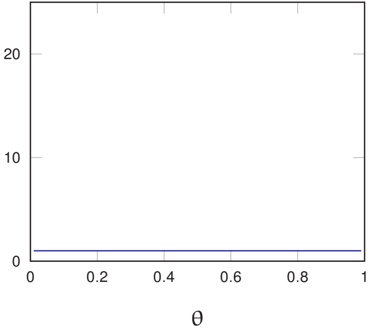
$\theta_{\text{MAP}} = 0.57$



Example: Uniform Prior

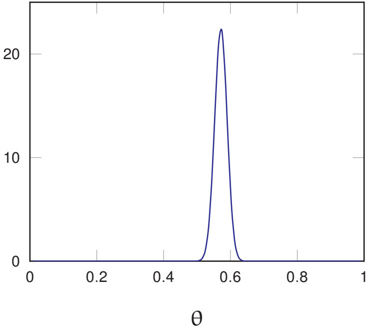
- $\beta_H = 1, \beta_T = 1$
  $\alpha_H = 3, \alpha_T = 2$

Prior: $\theta \sim \text{Beta}(1, 1)$



Posterior: $\theta \sim \text{Beta}(4, 3)$

$\theta_{\text{MAP}} = 0.60$



Lecture Overview

1 Lecture Overview

2 Maximum Likelihood Estimation

3 PAC Setting

4 Bayesian Parameter Estimation

5 Estimation for Gaussian Random Variables

6 Summary

Univariate Gaussian Distribution

- Let $\mathcal{X}$ be a continuous random variable, with outcomes $x \in \mathbb{R}$, distributed according to a **Gaussian** or **Normal** distribution:

$$x \sim \mathcal{N}(\mu, \sigma^2) \qquad \text{where:} \qquad \mu \in \mathbb{R},\ \sigma \in (0, \infty)$$

Likelihood

- Given $\mathcal{S} = \{x^{(i)}\}_{i=1}^{n}$, comprising outcomes of the random variables $\{\mathcal{X}_i\}_{i=1}^{n}$, all drawn i.i.d. according to such a normal distribution, the likelihood is given by:

$$\mathbb{P}(\mathcal{S}; \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right)$$

- And the log-likelihood is given by:

$$\begin{aligned}\ln\left(\mathbb{P}(\mathcal{S}; \mu, \sigma)\right) &= -n \ln(\sigma\sqrt{2\pi}) + \ln\left(\prod_{i=1}^{n} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right)\right) \\ &= -n \ln \sigma - \sum_{i=1}^{n} \left(\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) + \text{const.}\end{aligned}$$

MLE for μ

$$\mu_{\text{MLE}} = \operatorname*{argmax}_{\mu} \ln\left(\mathbb{P}(\mathcal{S}; \mu, \sigma)\right)$$

- Compute the derivative:

$$\begin{aligned}\frac{d}{d\mu} \ln\left(\mathbb{P}(\mathcal{S}; \mu, \sigma)\right) &= -\frac{d}{d\mu}\left(\sum_{i=1}^{n}\left(\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n}\left(\frac{(x^{(i)} - \mu)}{\sigma^2}\right)\end{aligned}$$

MLE for μ

- And set it equal to zero:

$$\sum_{i=1}^{n} \left(\frac{\left(x^{(i)} - \mu_{\mathrm{MLE}}\right)}{\sigma^2} \right) = 0$$

$$\sum_{i=1}^{n} x^{(i)} - n\mu_{\mathrm{MLE}} = 0$$

$$\mu_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$$

MLE for σ

$$\sigma_{\text{MLE}} = \operatorname*{argmax}_{\sigma} \ln\left(\mathbb{P}(\mathcal{S}; \mu, \sigma)\right)$$

- Compute the derivative:

$$\begin{aligned}\frac{d}{d\sigma}\ln\left(\mathbb{P}_{\mathcal{D}}(\mathcal{S}; \mu, \sigma)\right) &= -n\frac{d}{d\sigma}\ln\sigma - \frac{d}{d\sigma}\left(\sum_{i=1}^{n}\left(\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right)\right)\\ &= -\frac{n}{\sigma} + \sum_{i=1}^{n}\left(\frac{(x^{(i)}-\mu)^2}{\sigma^3}\right)\end{aligned}$$

MLE for σ

- And set it equal to zero:

$$-\frac{n}{\sigma_{\mathrm{MLE}}} + \sum_{i=1}^{n}\left(\frac{(x^{(i)}-\mu)^2}{\sigma_{\mathrm{MLE}}^3}\right) = 0$$

$$\sum_{i=1}^{n}\left(\frac{(x^{(i)}-\mu)^2}{\sigma_{\mathrm{MLE}}^2}\right) = n$$

$$\sigma_{\mathrm{MLE}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\mu)^2$$

Multivariate Gaussian Distribution

- Let $\mathcal{X}$ be a set of continuous random variables, with outcomes $\mathbf{x} \in \mathbb{R}^m$, distributed according to a **multivariate Gaussian** distribution:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{where:} \qquad \boldsymbol{\mu} \in \mathbb{R}^m,\ \boldsymbol{\Sigma}^T = \boldsymbol{\Sigma},\ \boldsymbol{\Sigma} \succ 0$$

Likelihood

- Given $\mathcal{S} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$ comprising outcomes of the random variable $\mathcal{X}$, all drawn i.i.d. according to such a multivariate normal distribution, the likelihood is given by:

$$\mathbb{P}_{\mathcal{D}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\right)$$

- And the log-likelihood is given by:

$$\begin{aligned} \ln\left(\mathbb{P}_{\mathcal{D}}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) &= -\frac{mn}{2}\ln(2\pi) - \frac{n}{2}\ln(|\boldsymbol{\Sigma}|) + \ln\left(\prod_{i=1}^{n} \exp\left(-\frac{(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})}{2}\right)\right) \\ &= -\frac{n}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) + \text{const.} \end{aligned}$$

MLE for $\boldsymbol{\mu}$

$$\boldsymbol{\mu}_{\text{MLE}} = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} \ln\left(\mathbb{P}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$$

- Compute the gradient:

$$\begin{aligned}\nabla_{\boldsymbol{\mu}} \ln\left(\mathbb{P}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) &= \nabla_{\boldsymbol{\mu}} \left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu})\right) \\ &= \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu})\end{aligned}$$

MLE for μ

- And set it equal to zero:

$$\sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\text{MLE}}) = \mathbf{0}$$

$$\sum_{i=1}^{n} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_{\text{MLE}}) = \mathbf{0}$$

$$\boldsymbol{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} \mathbf{x}^{(i)}}{n}$$

MLE for Σ

$$\boldsymbol{\Sigma}_{\text{MLE}} = \underset{\boldsymbol{\Sigma}}{\operatorname{argmax}} \ln\left(\mathbb{P}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$$

- Compute the gradient:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} \ln\left(\mathbb{P}(\mathcal{S}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) &= \frac{\partial}{\partial \boldsymbol{\Sigma}} \left(-\frac{n}{2} \ln(|\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) \right) \\
&= -\frac{n}{2} \left(\boldsymbol{\Sigma}^T\right)^{-1} + \frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{\Sigma}^T\right)^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \left(\boldsymbol{\Sigma}^T\right)^{-1} \\
&= -\frac{n}{2} \left(\boldsymbol{\Sigma}^T\right)^{-1} + \left(\boldsymbol{\Sigma}^T\right)^{-1} \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \left(\boldsymbol{\Sigma}^T\right)^{-1} \\
&= -\frac{n}{2} \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}
\end{aligned}$$

MLE for Σ

- And set it equal to zero:

$$-\frac{n}{2}\mathbf{\Sigma}_{\text{MLE}}^{-1} + \mathbf{\Sigma}_{\text{MLE}}^{-1}\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T\mathbf{\Sigma}_{\text{MLE}}^{-1} = \mathbf{0}$$

$$-\mathbf{\Sigma}_{\text{MLE}}^{-1} + \mathbf{\Sigma}_{\text{MLE}}^{-1}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T\mathbf{\Sigma}_{\text{MLE}}^{-1} = \mathbf{0}$$

$$-\mathbf{I} + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T\mathbf{\Sigma}_{\text{MLE}}^{-1} = \mathbf{0}$$

$$-\mathbf{\Sigma}_{\text{MLE}} + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T = \mathbf{0}$$

$$\mathbf{\Sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T$$

Bayesian Estimation

- Let us now assume that $\boldsymbol{\Sigma}$ is known, but $\boldsymbol{\mu}$ is the outcome of some random variable, $\mathcal{M}$, itself drawn from a multivariate normal distribution:

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

Where:
$\boldsymbol{\mu}_0 \in \mathbb{R}^m$
$\boldsymbol{\Sigma}_0^T = \boldsymbol{\Sigma}_0,\ \boldsymbol{\Sigma}_0 \succ 0$

Bayes' Rule

- Using **Bayes' Rule**:

$$
\begin{aligned}
p_{\mathcal{M}}(\boldsymbol{\mu}|\mathcal{S};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}) &= \frac{p_{\mathcal{X}}(\mathcal{S}|\boldsymbol{\mu};\boldsymbol{\Sigma})p_{\mathcal{M}}(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}{p_{\mathcal{X}}(\mathcal{S};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma})} \\
&\propto p_{\mathcal{X}}(\mathcal{S}|\boldsymbol{\mu};\boldsymbol{\Sigma})p_{\mathcal{M}}(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)
\end{aligned}
$$

- $p_{\mathcal{M}}(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ is the **Prior Probability** distribution function

- $p_{\mathcal{X}}(\mathcal{S}|\boldsymbol{\mu};\boldsymbol{\Sigma})$ is the **Likelihood** (i.e. $\mathbb{P}_{\mathcal{D}}(\mathcal{S};\boldsymbol{\mu},\boldsymbol{\Sigma})$)

- $p_{\mathcal{M}}(\boldsymbol{\mu}|\mathcal{S};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma})$ is the **Posterior Probability** distribution function

Prior Distribution & Likelihood

- In particular:

$$p_{\mathcal{M}}(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \frac{1}{(2\pi)^{\frac{m}{2}}} \frac{1}{|\boldsymbol{\Sigma}_0|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right)$$

$$\begin{aligned} p_{\mathcal{X}}(\mathcal{S}|\boldsymbol{\mu}; \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{\frac{mn}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\right) \end{aligned}$$

Aside: Completing the Square

- For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b}, \mathbf{x} \in \mathbb{R}^n$, where $\mathbf{A} = \mathbf{A}^T$:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} = (\mathbf{x} + \mathbf{A}^{-1}\mathbf{b})^T\mathbf{A}(\mathbf{x} + \mathbf{A}^{-1}\mathbf{b}) - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$$

Posterior Distribution

- Thus, using Bayes' rule:

$$
\begin{aligned}
p_{\mathcal{M}}(\boldsymbol{\mu}|\mathcal{S};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}) &\propto p_{\mathcal{X}}(\mathcal{S}|\boldsymbol{\mu};\boldsymbol{\Sigma})p_{\mathcal{M}}(\boldsymbol{\mu};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&\propto \exp\left(-\frac{1}{2}\left[\left(\sum_{i=1}^{n}(\mathbf{x}^{(i)}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})\right)+(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}-2\sum_{i=1}^{n}\mathbf{x}^{(i)T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}+n\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}-2n\overline{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\Big(\boldsymbol{\mu}-(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}n\overline{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)\Big)^T\Big(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\Big)\right. \\
&\qquad\left.\Big(\boldsymbol{\mu}-(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}n\overline{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)\Big)\right)
\end{aligned}
$$

Where: $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}^{(i)}$

Posterior Distribution

- Thus:

$$p_{\mathcal{M}}(\boldsymbol{\mu}|\mathcal{S};\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)\right)$$

Where:
$\boldsymbol{\mu}_n = \left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}n\bar{\mathbf{x}}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$
$\boldsymbol{\Sigma}_n = \left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$

- Recognising the functional form of a multivariate Gaussian allows us to express the posterior distribution of $\boldsymbol{\mu}$ as:

$$\boldsymbol{\mu}|\mathcal{S} \sim \mathcal{N}(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$$

Lecture Overview

1 Lecture Overview

2 Maximum Likelihood Estimation

3 PAC Setting

4 Bayesian Parameter Estimation

5 Estimation for Gaussian Random Variables

6 Summary

Lecture Summary

1. The **frequentist** framework is an approach to probabilistic machine learning, in which parameter estimators - for example the **MLE** - are learnt

2. In the **PAC** framework we attempt to form probabilistic bounds on quantities of interest rather than to infer characteristics of the data generating distribution

3. The **Bayesian** framework is an alternative approach to probabilistic machine learning, where parameters are treated as random variables. The **MAP** estimator is a way of generating a point estimate.