Introduction

- Statistical models find the most accurate parameters that fits the data and the belief based on domain knowledge
- E.g. For Bernoulli, we find λ . For Normal Distribution, we find mean $\boldsymbol{\mu}$ and covariance matrix $\sum$. For any distribution, we find $\boldsymbol{\theta}$. All the process which finds the fittest parameter is called *parameter estimation*
- There are 2 ways to determine $\boldsymbol{\theta}$
  - First way is based on dataset, it's called *Maximum Likelihood Estimation* (MLE)
  - Second way is not only based on dataset but the belief the men who has the domain knowledge propose, it's called *Maximum a Posteriori Estimation* (MAP Estimation)

Maximum Likelihood Estimation

Idea

- Assume by somehow, we know n points follow the distribution with parameter of $\boldsymbol{\theta}$
- MLE will find the best $\boldsymbol{\theta}$ such that $\theta = \max_{\theta} p\left(\mathrm{x}_1, \dots, \mathrm{x}_N \mid \theta\right)$, it means we find the distribution to fit the dataset in the best way
- For details, Likelihood means how your model fit the dataset

Independence Assumption and Log-likelihood

- Another issue arises when we try to find out $\boldsymbol{\theta}$ based on $\max_{\theta} p\left(\mathrm{x}_1, \dots, \mathrm{x}_N \mid \theta\right)$ because it's almost impossible to find the distribution which describes the joint probability of the whole dataset
- We can solve this issue by assuming the independence among the points of dataset: $p\left(\mathrm{x}_1, \dots, \mathrm{x}_N \mid \theta\right) \approx \prod_{n=1}^{N} p\left(\mathrm{x}_n \mid \theta\right)$, so the MLE become: find out $\boldsymbol{\theta}$ such that $\theta = \max_{\theta} \prod_{n=1}^{N} p\left(\mathrm{x}_n \mid \theta\right)$
- But, here we will meet another problem here when we try to maximize $\prod_{n=1}^{N} p\left(\mathrm{x}_n \mid \theta\right)$, because it easily $\rightarrow 0$ so to solve it, you need to maximize log: $\theta = \max_{\theta} \sum_{n=1}^{N} \log\left(p\left(\mathrm{x}_n \mid \theta\right)\right)$
- Logarithm Property: Because log is the *monotonic increasing*, $\max_{x} f(x) = \max_{x} \log f(x)$
- E.g. We flip the coins N times and get n times get head. Find the probability of flipping head
  - Intuitively, this probability: $\lambda = \frac{n}{N}$, but now we use MLE to check this probability
  - Put $x_1, x_2, \dots, x_N$ is the output of *head* (1) or *tail* (0) and we have n heads and $m = N - n$ tails: $\begin{cases} \sum_{i=1}^{N} x_i = n \\ N - \sum_{i=1}^{N} x_i = N - n = m \\ p\left(x_i \mid \lambda\right) = \lambda^{x_i}\left(1-\lambda\right)^{1-x_i} \end{cases}$
  - Based on MLE,

$$\lambda = \arg\max_{\lambda}\left[p\left(x_1, x_2, \dots x_N \mid \lambda\right)\right] = \arg\max_{\lambda}\left[\prod_{i=1}^{N} p\left(x_i \mid \lambda\right)\right]$$

$$= \arg\max_{\lambda}\left[\prod_{i=1}^{N} \lambda^{x_i}\left(1-\lambda\right)^{1-x_i}\right] = \arg\max_{\lambda}\left[\lambda^{\sum_{i=1}^{N} x_i}\left(1-\lambda\right)^{N-\sum_{i=1}^{N} x_i}\right]$$

$$= \arg\max_{\lambda}\left[\lambda^{n}\left(1-\lambda\right)^{m}\right] = \arg\max_{\lambda}\left[n\log\lambda + m\log\left(1-\lambda\right)\right] = \arg\max_{\lambda} f\left(\lambda\right)$$

Now, we can take derivative of $f(\lambda)$ to maximize it, $f'(\lambda)=\frac{n}{\lambda}-\frac{m}{1-\lambda}=0 \Leftrightarrow \frac{n}{\lambda}=\frac{m}{1-\lambda} \Leftrightarrow \lambda=\frac{n}{n+m}=\frac{n}{N}$

- E.g. We roll the 6-face dice, probability of each face is same. Assume you roll N times, number of times we get first, second, … face is $n_1,n_2,\dots,n_6$ and $\sum_{i=1}^{6} n_i = N$. Calculate probability of each face. Assume $n_i > 0$
    - Intuitively, this probability: $\lambda=\frac{n_j}{N}$, now use MLE to check this probability
    - Represent each output of dice as the 6-value vector $\mathbf{x}_i \in \{0,1\}^6$ in which 1 respects the value of face you roll, the others are 0, so $p(\mathbf{x}_i \mid \boldsymbol{\lambda})=\prod_{j=1}^{6}\lambda_j^{x_i^j}$ in which λ_j is the probability of face j, x_i^j : j is the value number j in vector x_i, and put $n_j=\sum_{i=1}^{N} x_i^j, \forall j=1,2,\dots,6$
    - Based on MLE,

$$\boldsymbol{\lambda}=\arg\max_{\boldsymbol{\lambda}}\left[\prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\lambda})\right]=\arg\max_{\boldsymbol{\lambda}}\left[\prod_{i=1}^{N}\prod_{j=1}^{6}\lambda_j^{x_i^j}\right]$$

$$=\arg\max_{\boldsymbol{\lambda}}\left[\prod_{j=1}^{6}\lambda_j^{\sum_{i=1}^{N} x_i^j}\right]=\arg\max_{\boldsymbol{\lambda}}\left[\prod_{j=1}^{6}\lambda_j^{n_j}\right]$$

$$=\arg\max_{\boldsymbol{\lambda}}\left[\sum_{j=1}^{6} n_j \log(\lambda_j)\right] \qquad with \sum_{j=1}^{6}\lambda_j=1$$

- Apply Lagrange, we have: $L(\boldsymbol{\lambda},\mu)=\sum_{j=1}^{6} n_j \log(\lambda_j)+\mu\left(1-\sum_{j=1}^{6}\lambda_j\right)$
- To find the solution of Lagrange, we just $\begin{cases}\dfrac{\partial L(\boldsymbol{\lambda},\mu)}{\partial \lambda_j}=0\\ \dfrac{\partial L(\boldsymbol{\lambda},\mu)}{\partial \mu}=0\end{cases}$

$$\rightarrow \begin{cases}\dfrac{\partial L(\boldsymbol{\lambda},\mu)}{\partial \lambda_j}=\dfrac{n_j}{\lambda_j}-\mu=0 \rightarrow \lambda_j=\dfrac{n_j}{\mu}\\ \dfrac{\partial L(\boldsymbol{\lambda},\mu)}{\partial \mu}=1-\sum_{j=1}^{6}\lambda_j=0 \rightarrow \sum_{j=1}^{6}\lambda_j=\sum_{j=1}^{6}\dfrac{n_j}{\mu}=1 \rightarrow \mu=\sum_{j=1}^{6} n_j=N \rightarrow \lambda_j=\dfrac{n_j}{N}\end{cases}$$

- E.g. Assume we need to measure the somebody's height. It's hard to find the exact height in once time. Therefore, we measure many times and find the **expectation** of the data with the assumption which is data is based on Normal Distribution and independent
    - In some cases, expectation of the data, which we need to find out, may not be expectation of the distribution. So here, we need to prove that expectation of the data = expectation of the distribution
    - Assume the height we got is $x_1,x_2,\dots,x_N$. So here we find the distribution with μ and σ^2 such that $x_1,x_2,\dots,x_N$ is the most likely. We know $p(x_i \mid \mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$
    - Based on MLE,

$$\mu,\sigma = \arg\max_{\mu,\sigma}\left[\prod_{i=1}^{N} p\left(x_i \mid \mu,\sigma^2\right)\right]$$

$$= \arg\max_{\mu,\sigma}\left[\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}\exp\left(-\frac{\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{2\sigma^2}\right)\right]$$

$$= \arg\max_{\mu,\sigma}\left[-N\log\left(\sigma\right)-\frac{\sum_{i=1}^{N}\left(x_i-\mu\right)^2}{2\sigma^2} \triangleq J\left(\mu,\sigma\right)\right]$$

- 2π is ignored because it does not impact on result
- Now, we can take partial derivative of $J\left(\mu,\sigma\right)$ to maximize it

$$\begin{cases} \dfrac{\partial J}{\partial \mu} = \dfrac{1}{\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right) = 0 \\ \dfrac{\partial J}{\partial \sigma} = -\dfrac{N}{\sigma} + \dfrac{1}{\sigma^3}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 = 0 \end{cases} \rightarrow \begin{cases} \mu = \dfrac{1}{N}\sum_{i=1}^{N} x_i \\ \sigma^2 = \dfrac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \end{cases}$$

Maximum a Posterior

Idea

- Assume we flip the coin 5000 times, we got 1000 heads, so $\text{probability}_{\text{head}} = 0.2$ and this probability may be reliable because of large data point (5000). On the contrary, assume we flip the coin 5 times, we just got 1 head, so $\text{probability}_{\text{head}} = 0.2$, but because the small data points (5) – low training, this probability may be unreliable (or overfitting)
- Therefore, when we got low-training problem, we need to consider the belief (assumption of parameter), in above case, we believe that $\text{probability}_{\text{head}} \approx 0.5$
- Maximum a Posterior (MAP) can solve such problem. MAP introduces the constraint for parameter $\boldsymbol{\theta}$, *the prior*.
- Instead of finding out $\theta = \arg\max_{\theta} p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N \mid \theta\right)$, we find out $\theta = \arg\max_{\theta} p\left(\theta \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)$
- $p\left(\theta \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)$ is called *posterior probability*
- However, $p\left(\theta \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)$, this probability is hard to find out because it's more common sense to find out $p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N \mid \theta\right)$ which constructs the distribution when given parameter θ and after that, compare the distribution of $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$ we construct from θ and distribution of real data. To solve it, apply Bayes Theorem

$$\theta = \arg\max_{\theta} p\left(\theta \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right) = \arg\max_{\theta}\left[\frac{\overbrace{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N \mid \theta\right)}^{\textit{Likelihood}}\overset{\textbf{Prior}}{p\left(\theta\right)}}{\underbrace{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\right)}_{\textit{Evidence}}}\right]$$

$$= \arg\max_{\theta}\left[p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N \mid \theta\right)p\left(\theta\right)\right]$$

$$= \arg\max_{\theta}\left[\prod_{i=1}^{N} p\left(\mathbf{x}_i \mid \theta\right)p\left(\theta\right)\right]$$

- **Posterior is directly proportional to the multiplication of likelihood and prior**
- *Prior* is hyper-parameter, so How to determine Prior, *Conjugate Prior* may solve it

Conjugate Error

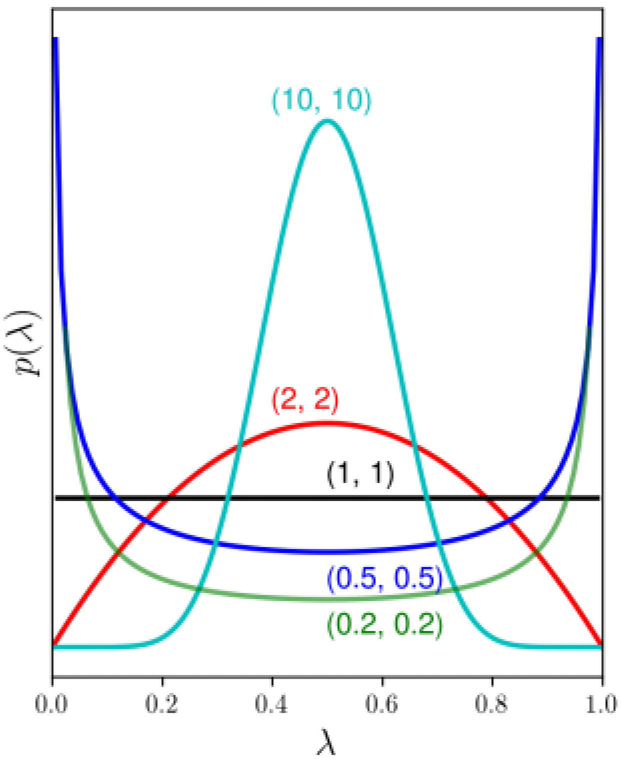
- If posterior $p(\theta \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ is in the *same family* with prior $p(\theta)$, prior and posterior are *conjugate distributions*
- $p(\theta)$ is called *conjugate prior* of likelihood $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \mid \theta)$, so MAP and MLE have the same distribution
- Some couples of *conjugate distributions*:
  - If likelihood function is Gaussian, prior needs to be Gaussian, so the posterior is also Gaussian. We call it *Gaussian Conjugate* or *self-conjugate*
  - If likelihood function is Gaussian, its prior (for variance) is Gamma Distribution, posterior is Gaussian.
    Note: The variance may be used to measure the accuracy of model, the less variance is, the more accuracy the model is
  - Beta is conjugate of Bernoulli Distribution
  - Dirichlet is conjugate of Categorical Distribution

Hyper-parameter

- Given the Bernoulli pdf: $p(x \mid \lambda) = \lambda^x (1-\lambda)^{1-x}$ and its conjugate, Beta pdf: $p(\lambda) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$
- If we ignore the constant parameter $\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ which purpose is to make sure integration of Beta pdf is 1, we can easily realize that Beta distribution is in same family with Bernoulli distribution, so $p(\lambda \mid x) \propto p(x \mid \lambda)\, p(\lambda) \propto \lambda^{x+\alpha-1}(1-\lambda)^{1-x+\beta-1}$ is also in Bernoulli Distribution
- E.g. Back to flipping coin problem, we flip the coin N times, we got n heads and m = N – n tails. If applying the MLE, $\lambda = \dfrac{n}{N}$. How about MAP in which prior is Beta[α, β] ?
  - Based on MAP,

$$
\begin{aligned}
\lambda &= \arg\max_{\lambda}\left[p(x_1, \ldots, x_N \mid \lambda)\, p(\lambda) \right] \\
&= \arg\max_{\lambda}\left[\left(\prod_{i=1}^{N} \lambda^{x_i} (1-\lambda)^{1-x_i} \right) \lambda^{\alpha-1}(1-\lambda)^{\beta-1} \right] \\
&= \arg\max_{\lambda}\left[\lambda^{\alpha-1+\sum_{i=1}^{N} x_i} (1-\lambda)^{\beta-1+N-\sum_{i=1}^{N} x_i} \right] \\
&= \arg\max_{\lambda}\left[\lambda^{n+\alpha-1}(1-\lambda)^{m+\beta-1} \triangleq f(\lambda) \right]
\end{aligned}
$$

  - We maximize $f(\lambda)$ like the way in MLE, so $\lambda = \dfrac{n+\alpha-1}{N+\alpha+\beta-2}$, because Posterior and Likelihood is in the same family, we can easily maximize MAP
  - Remaining issue: How to choose hyper-parameter α and β



Hình 4.1: Đồ thị hàm mật độ xác suất của phân phối Beta khi $\alpha = \beta$ và nhận các giá trị khác nhau. Khi cả hai giá trị này lớn, xác suất để λ gần 0.5 sẽ cao hơn.

- When α = β > 1, Beta pdf is symmetric at $x = 0.5$ and get maximum at $x = 0.5$, so λ is more likely ≈0.5
- When α = β = 1, We got uniform distribution, at this time, probability of every λ is the same. Therefore, when we apply MAP in this case, $\lambda = \frac{n}{N}$ → Conclusion: **MLE is the special case of MAP when Prior is uniform distribution**
- If we choose α = β = 2, we got $\lambda = \frac{n+1}{N+2}$. e.g. choosing N = 5, n = 1, MAP got $\lambda = \frac{2}{7}$ more ≈0.5 than $\frac{1}{5}$ MLE results
- If we choose α = β = 10, we got $\lambda = \frac{n+9}{N+10}$. e.g. choosing N = 5, n = 1, MAP got $\lambda = \frac{10}{23}$ →Conclusion:

$$\alpha = \beta \to \infty, \lambda \to \frac{1}{2}$$

MAP helps to avoid overfitting

- The analogy in MAP and Regularization
  - MAP

$$\theta = \arg\max_{\theta} p(X \mid \theta)\, p(\theta)$$

$$= \arg\max_{\theta} \left[\log \underbrace{p(X \mid \theta)}_{Likelihood} + \log \underbrace{p(\theta)}_{\textbf{Prior}} \triangleq f(\theta) \right]$$

  - f(θ) is very identical with $L(\theta) + \lambda R(\theta)$ in the regularization. So we can say MAP is the method to avoid overfitting in statistical learning, especially when low-training