Principles of Parameter Estimation

Predrag Radivojac
School of Informatics, Computing, and Engineering
Bloomington, Indiana

INDIANA UNIVERSITY BLOOMINGTON

PRELIMINARIES

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^{n}$, $x_i \in \mathcal{X}$

Objective: find a model $\hat{f} \in \mathcal{F}$ that models the phenomenon well

Requirements:

(i) the ability to generalize well

(ii) the ability to incorporate prior knowledge and assumptions

(iii) scalability

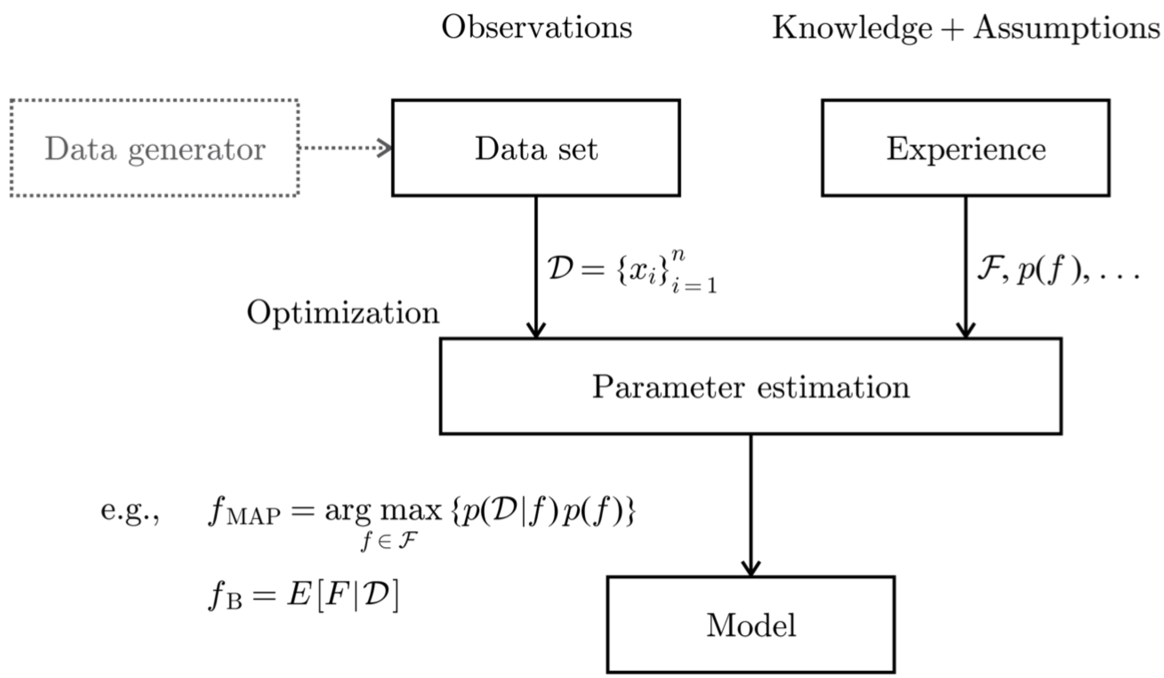
Terminology through an example: $\mathcal{D} = \{3.1, 2.4, -1.1, 0.1\}$

What is the data generator?

$\mathcal{F} = \text{Gaussian}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

Parameter estimation

Statistical Framework



$$\textit{Model inference}: \text{Observations} + \begin{array}{c}\text{Knowledge}\\ \text{and}\\ \text{Assumptions}\end{array} + \text{Optimization}$$

MAXIMUM A POSTERIORI INFERENCE

Idea:

$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} \{p(f|\mathcal{D})\},$$

where $p(f|\mathcal{D})$ is called the posterior distribution.

How do we calculate it?

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}$$

where $p(\mathcal{D}|f)$ = likelihood, $p(f)$ = prior, and $p(\mathcal{D})$ = data distribution.

Maximum A Posteriori Inference

Finding the data distribution:

$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

We can now simplify the process if we observe that

$$\begin{aligned} p(f|\mathcal{D}) &= \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|f) \cdot p(f) \end{aligned}$$

Maximum Likelihood Inference

Express the posterior distribution as

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})}$$
$$\propto p(\mathcal{D}|f) \cdot p(f)$$

Now, ignore $p(f)$ to get

$$f_{\mathrm{ML}} = \underset{f \in \mathcal{F}}{\arg\max} \{p(\mathcal{D}|f)\}$$

There are technical problems with this approach.

Examples of Maximum Likelihood Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is an i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$

Find λ

Solution: Poisson probability mass function is $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$\lambda_{\text{ML}} = \underset{\lambda \in (0,\infty)}{\arg\max} \left\{ p(\mathcal{D}|\lambda) \right\}.$$

Likelihood:
$$\begin{aligned} p(\mathcal{D}|\lambda) &= p(\{x_i\}_{i=1}^n \,|\lambda) \\ &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Examples of Maximum Likelihood Inference

Likelihood: $$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

Log-likelihood: $$ll(\mathcal{D}, \lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln (x_i!)$$

Optimization:

$$\frac{\partial ll(\mathcal{D}, \lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$
$$= 0$$

Solution:

$$\lambda_{\mathrm{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
$$= 5.5$$

MAP and ML estimates are called the point estimates.

Examples of MAP Inference

Example: $\mathcal{D} = \{2, 5, 9, 5, 4, 8\}$ is i.i.d. sample from Poisson(λ), $\lambda \in \mathbb{R}^+$

Assume λ is taken from $\Gamma(x|k,\theta)$ with parameters $k = 3$ and $\theta = 1$

Find λ

Solution: Poisson: $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

Gamma: $\Gamma(x|k,\theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}$, where $x > 0$, $k > 0$, and $\theta > 0$.

Likelihood:

$$p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

Prior:

$$p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)}.$$

Examples of MAP Inference

Log-likelihood:

$$
\begin{aligned}
\ln p(\lambda|\mathcal{D}) &\propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\
&= \ln \lambda (k - 1 + \sum_{i=1}^{n} x_i) - \lambda (n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)
\end{aligned}
$$

We now obtain

$$
\begin{aligned}
\lambda_{\text{MAP}} &= \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}} \\
&= 5
\end{aligned}
$$

ANOTHER EXAMPLE

Example: $\mathcal{D} = \{x_i\}_{i=1}^n$ is i.i.d. sample from Gaussian(μ, σ^2)

Find μ and σ

Solution: Gaussian: $p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{\mathrm{ML}})^2 .$$

Relationship to Kullback-Leibler (KL) Divergence

The KL divergence between two probability distributions $p(x)$ and $q(x)$ is

$$D_{\mathrm{KL}}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

Assume now the data is generated according to some $p(x|\theta_t)$. We estimated it as $p(x|\theta)$.

Let's look at the KL divergence

$$\begin{aligned} D_{\mathrm{KL}}(p(x|\theta_t)||p(x|\theta)) &= \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{p(x|\theta_t)}{p(x|\theta)} dx \\ &= \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta)} dx - \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta_t)} dx. \end{aligned}$$

$-E\left[\log p(x|\theta)\right]$

Relationship to Kullback-Leibler (KL) Divergence

$$\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|\theta) \quad \xrightarrow{a.s.} \quad E[\log p(x|\theta)]$$

when $n \to \infty$.

Conclusion:

When $n \to \infty$, ML estimation implies $p(x|\theta_{\mathrm{ML}}) = p(x|\theta_t)$

This usually implies $\theta_{\mathrm{ML}} = \theta_t$

Parameter Estimation for Mixtures of Distributions

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^{n}$, $x_i \in \mathcal{X}$

$$p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j).$$

$$w_j \geq 0, \quad \sum_{j=1}^{m} w_j = 1.$$

where $\theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m)$

Example: Consider a mixture of $m = 2$ exponential distributions.

$p(x|\theta_j) = \lambda_j e^{-\lambda_j x}$, where $\lambda_j > 0$

$$p(x|\lambda_1, \lambda_2, w_1, w_2) = w_1 \cdot \lambda_1 e^{-\lambda_1 x} + w_2 \cdot \lambda_2 e^{-\lambda_2 x}$$

where $\lambda_1, \lambda_2 > 0$, $w_1, w_2 \geq 0$, and $w_1 = 1 - w_2$

Parameter Estimation for Mixtures of Distributions

Likelihood:

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i|\theta) \\
&= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j p(x_i|\theta_j) \right)
\end{aligned}
$$

$p(\mathcal{D}|\theta)$ has $O(m^n)$ terms. It can be calculated in $O(mn)$ time as a log-likelihood.

How can we find θ? Is there a closed-form solution?

IDEA #1

Suppose we know what data point is generated by what mixing component.

That is, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ is an i.i.d. sample from some distribution $p(x, y)$, where $y \in \mathcal{Y} = \{1, 2, \ldots, m\}$ specifies the mixing component.

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \prod_{i=1}^{n} p(x_i, y_i|\theta) \\
&= \prod_{i=1}^{n} p(x_i|y_i, \theta) p(y_i|\theta) \\
&= \prod_{i=1}^{n} w_{y_i} p(x_i|\theta_{y_i}),
\end{aligned}
$$

where $w_j = P(Y = j)$.

Idea #1

Log-likelihood:

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{n} (\log w_{y_i} + \log p(x_i|\theta_{y_i}))$$
$$= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),$$

where n_j is the number of data points in $\mathcal{D}$ generated by the j-th mixing component.

Constrained optimization: Let's first find $\mathbf{w}$

$$L(\mathbf{w}, \alpha) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left(\sum_{j=1}^{m} w_j - 1 \right)$$

where α is the Lagrange multiplier.

IDEA #1

Set $\frac{\partial}{\partial w_k} L(\mathbf{w}, \alpha) = 0$ for every $k \in \mathcal{Y}$ and $\frac{\partial}{\partial \alpha} L(\mathbf{w}, \alpha) = 0$. Solve it.

It follows that $w_k = -\frac{n_k}{\alpha}$ and $\alpha = -n$.

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k),$$

where $I(\cdot)$ is the indicator function.

To find all θ_j, we need to get concrete.

$$\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i | \lambda_{y_i}) = 0,$$

for each $k \in \mathcal{Y}$.

IDEA #1

We obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} I(y_i = k) \cdot x_i},$$

for each $k \in \mathcal{Y}$.

Recall that

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k)$$

If the mixing component designations $\mathbf{y}$ are known,
the parameter estimation is greatly simplified.

IDEA #2

Suppose we know the θ but not the mixing componen designations..

Looks like clustering, right? Let's see. Express

$$p(\mathbf{y}|\mathcal{D},\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta)$$
$$= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)}$$

and subsequently find the best configuration out of m^n possibilities.

Data is i.i.d. so y_i can be estimated separately. The MAP estimate for y_i

$$\hat{y}_i = \arg\max_{y_i \in \mathcal{Y}} \left\{ \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \right\}$$

CLASSIFICATION EXPECTATION MAXIMIZATION (CEM)

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$
2. Calculate $y_i^{(0)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(0)} p(x_i | \lambda_k^{(0)})}{\sum_{j=1}^{m} w_j^{(0)} p(x_i | \lambda_j^{(0)})} \right\}$ for $\forall i \in \{1, 2, \ldots, n\}$
3. Set $t = 0$
4. Repeat until convergence
   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} I(y_i^{(t)} = j)$
   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} I(y_i^{(t)}=k)}{\sum_{i=1}^{n} I(y_i^{(t)}=k) \cdot x_i}$
   (c) $t = t + 1$
   (d) $y_i^{(t+1)} = \arg\max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(t)} p(x_i | \lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i | \lambda_j^{(t)})} \right\}$
5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$