ML, MAP Estimation and Bayesian

CE-717: Machine Learning
Sharif University of Technology
Fall 2016

Soleymani
Outline

- Introduction
- Maximum-Likelihood (ML) estimation
- Maximum A Posteriori (MAP) estimation
- Bayesian inference
Relation of learning & statistics

- Target model in the learning problems can be considered as a statistical model.

- For a fixed set of data and underlying target (statistical model), the estimation methods try to estimate the target from the available data.
Density estimation

- Estimating the probability density function \( p(x) \), given a set of data points \( \{x^{(i)}\}_{i=1}^N \) drawn from it.

Main approaches of density estimation:

- **Parametric**: assuming a parameterized model for density function
  - A number of parameters are optimized by fitting the model to the data set

- **Nonparametric** (Instance-based): No specific parametric model is assumed
  - The form of the density function is determined entirely by the data
Parametric density estimation

- Estimating the probability density function $p(x)$, given a set of data points $\{x^{(i)}\}_{i=1}^{N}$ drawn from it.

- Assume that $p(x)$ in terms of a specific functional form which has a number of adjustable parameters.

- Methods for parameter estimation
  - Maximum likelihood estimation
  - Maximum A Posteriori (MAP) estimation
Parametric density estimation

- **Goal**: estimate parameters of a distribution from a dataset 
  \[ \mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\} \]
  - \( \mathcal{D} \) contains \( N \) independent, identically distributed (i.i.d.) training samples.

- We need to determine \( \theta \) given \( \{x^{(1)}, \ldots, x^{(N)}\} \)
  - How to represent \( \theta \)?
    - \( \theta^* \) or \( p(\theta) \)?
Example

\[ P(x|\mu) = \mathcal{N}(x|\mu, 1) \]
Example
Maximum Likelihood Estimation (MLE)

- Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given data.

- Likelihood is the conditional probability of observations \( D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \) given the value of parameters \( \theta \)

- Assuming i.i.d. observations:

  \[
p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta)
  \]

  likelihood of \( \theta \) w.r.t. the samples

- Maximum Likelihood estimation

  \[
  \hat{\theta}_{ML} = \arg \max_{\theta} p(D|\theta)
  \]
Maximum Likelihood Estimation (MLE)

\( \hat{\theta} \) best agrees with the observed samples
Maximum Likelihood Estimation (MLE)

\( \hat{\theta} \) best agrees with the observed samples
Maximum Likelihood Estimation (MLE)

\[ \hat{\theta} \text{ best agrees with the observed samples} \]
Maximum Likelihood Estimation (MLE)

\[
\mathcal{L}(\theta) = \ln p(D|\theta) = \ln \prod_{i=1}^{N} p(x^{(i)}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta)
\]

\[
\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}} \mathcal{L}(\theta) = \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \ln p(x^{(i)}|\theta)
\]

Thus, we solve \( \nabla_{\theta} \mathcal{L}(\theta) = 0 \) to find global optimum
MLE
Bernoulli

- Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}, m$ heads (1), $N - m$ tails (0)

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=1}^{N} \theta^{x^{(i)}} (1 - \theta)^{1-x^{(i)}}$$

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x^{(i)}|\theta) = \sum_{i=1}^{N} \{x^{(i)} \ln \theta + (1 - x^{(i)}) \ln (1 - \theta)\}$$

$$\frac{\partial \ln p(\mathcal{D}|\theta)}{\partial \theta} = 0 \Rightarrow \theta_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N} = \frac{m}{N}$$
MLE
Bernoulli: example

- Example: $\mathcal{D} = \{1,1,1\}, \hat{\theta}_{ML} = \frac{3}{3} = 1$
  - Prediction: all future tosses will land heads up

- Overfitting to $\mathcal{D}$
MLE: Multinomial distribution

- Multinomial distribution (on variable with $K$ state):

  Parameter space:
  \[ \theta = [\theta_1, \ldots, \theta_K] \]
  \[ \theta_i \in [0,1] \]
  \[ \sum_{k=1}^{K} \theta_k = 1 \]

  \[ x = [x_1, \ldots, x_K] \]
  \[ x_k \in \{0,1\} \]
  \[ \sum_{k=1}^{K} x_k = 1 \]

  \[ P(x|\theta) = \prod_{k=1}^{K} \theta_k^{x_k} \]
  \[ P(x_k = 1) = \theta_k \]
MLE: Multinomial distribution

\[ \mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \]

\[
P(\mathcal{D}|\mathbf{\theta}) = \prod_{i=1}^{N} P(x^{(i)}|\mathbf{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_{k}^{x^{(i)}_{k}} = \prod_{k=1}^{K} \theta_{k}^{\sum_{i=1}^{N} x_{k}^{(i)}}
\]

\[
\mathcal{L}(\mathbf{\theta}, \lambda) = \ln p(\mathcal{D}|\mathbf{\theta}) + \lambda (1 - \sum_{k=1}^{K} \theta_{k})
\]

\[
\hat{\theta}_{k} = \frac{\sum_{i=1}^{N} x_{k}^{(i)}}{N} = \frac{N_{k}}{N}
\]
MLE
Gaussian: unknown $\mu$

$$\ln p(x^{(i)} | \mu) = -\ln\{\sqrt{2\pi}\sigma\} - \frac{1}{2\sigma^2} (x^{(i)} - \mu)^2$$

$$\frac{\partial L(\mu)}{\partial \mu} = 0 \Rightarrow \frac{\partial}{\partial \mu} \left( \sum_{i=1}^{N} \ln p(x^{(i)} | \mu) \right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2} (x^{(i)} - \mu) = 0 \Rightarrow \hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

MLE corresponds to many well-known estimation methods.
MLE
Gaussian: unknown $\mu$ and $\sigma$

$$\theta = [\mu, \sigma]$$

$$\nabla_{\theta} \mathcal{L}(\theta) = 0$$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} \Rightarrow \hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma} \Rightarrow \hat{\sigma}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \hat{\mu}_{ML})^2$$
Maximum A Posteriori (MAP) estimation

- **MAP estimation**
  \[
  \hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{D})
  \]

- Since \( p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) \)
  \[
  \hat{\theta}_{MAP} = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta)
  \]

- **Example of prior distribution:**
  \[
  p(\theta) = \mathcal{N}(\theta_0, \sigma^2)
  \]
MAP estimation
Gaussian: unknown \( \mu \)

\[
p(x|\mu) \sim N(\mu, \sigma^2) \quad \mu \text{ is the only unknown parameter}
\]

\[
p(\mu|\mu_0) \sim N(\mu_0, \sigma_0^2) \quad \mu_0 \text{ and } \sigma_0 \text{ are known}
\]

\[
\frac{d}{d\mu} \ln \left( p(\mu) \prod_{i=1}^{N} p(x^{(i)}|\mu) \right) = 0
\]

\[
\Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma^2} (x^{(i)} - \mu) - \frac{1}{\sigma_0^2} (\mu - \mu_0) = 0
\]

\[
\Rightarrow \hat{\mu}_{MAP} = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2} \sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}
\]

\[
\frac{\sigma_0^2}{\sigma^2} \gg 1 \text{ or } N \to \infty \Rightarrow \hat{\mu}_{MAP} = \hat{\mu}_{ML} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}
\]
Maximum A Posteriori (MAP) estimation

- Given a set of observations $\mathcal{D}$ and a prior distribution $p(\theta)$ on parameters, the parameter vector that maximizes $p(\mathcal{D}|\theta)p(\theta)$ is found.

\[
\hat{\theta}_{MAP} \cong \hat{\theta}_{ML} \quad \text{or} \quad \hat{\theta}_{MAP} > \hat{\theta}_{ML}
\]

\[
\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_{ML}
\]
MAP estimation
Gaussian: unknown $\mu$ (known $\sigma$)

$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$

$p(\mu|\mathcal{D}) = N(\mu|\mu_N, \sigma_N)$

$\mu_N = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2} \sum_{i=1}^{N} x^{(i)}}{1 + \frac{\sigma_0^2}{\sigma^2} N}$

$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$

More samples $\Rightarrow$ sharper $p(\mu|\mathcal{D})$
Higher confidence in estimation
Conjugate Priors

- We consider a form of prior distribution that has a simple interpretation as well as some useful analytical properties.

- Choosing a prior such that the posterior distribution that is proportional to $p(\mathcal{D}|\theta)p(\theta)$ will have the same functional form as the prior.

$$\forall \alpha, \mathcal{D} \exists \alpha' \quad P(\theta|\alpha') \propto P(\mathcal{D}|\theta)P(\theta|\alpha)$$

Having the same functional form
Prior for Bernoulli Likelihood

- **Beta distribution** over $\theta \in [0,1]$:

  $\text{Beta}(\theta | \alpha_1, \alpha_0) \propto \theta^{\alpha_1-1} (1 - \theta)^{\alpha_0-1}$

  $\text{Beta}(\theta | \alpha_1, \alpha_0) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \theta^{\alpha_1-1} (1 - \theta)^{\alpha_0-1}$

- Beta distribution is the conjugate prior of Bernoulli:

  $P(x|\theta) = \theta^x (1 - \theta)^{1-x}$

$$E[\theta] = \frac{\alpha_1}{\alpha_0 + \alpha_1}, \quad \hat{\theta} = \frac{\alpha_1 - 1}{\alpha_0 - 1 + \alpha_1 - 1}$$

most probable $\theta$
Beta distribution

$p(\theta)$

$\theta$

Beta$(1, 1)$

$p(\theta)$

$\theta$

Beta$(2, 2)$

$p(\theta)$

$\theta$

Beta$(10, 10)$

$p(\theta)$

$\theta$

Beta$(3, 2)$

$p(\theta)$

$\theta$

Beta$(15, 10)$

$p(\theta)$

$\theta$

Beta$(0.5, 0.5)$
Benoulli likelihood: posterior

Given: $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, $m$ heads ($1$), $N - m$ tails ($0$)

$$p(\theta | \mathcal{D}) \propto p(\mathcal{D} | \theta)p(\theta)$$

$$= \left( \prod_{i=1}^{N} \theta^{x^{(i)}} (1 - \theta)^{(1 - x^{(i)})} \right) \text{Beta}(\theta | \alpha_1, \alpha_0)$$

$$\propto \theta^{m + \alpha_1 - 1} (1 - \theta)^{N - m + \alpha_0 - 1}$$

$$\Rightarrow p(\theta | \mathcal{D}) \propto \text{Beta}(\theta | \alpha'_1, \alpha'_0)$$

$$\alpha'_1 = \alpha_1 + m$$

$$\alpha'_0 = \alpha_0 + N - m$$
Example

Prior Beta: $\alpha_0 = \alpha_1 = 2$

Posterior Beta: $\alpha'_1 = 5, \alpha'_0 = 2$

Given: $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$:
$m$ heads (1), $N - m$ tails (0)

$\alpha_0 = \alpha_1 = 2$

$D = \{1,1,1\} \Rightarrow N = 3, m = 3$

$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D) = \frac{\alpha'_1 - 1}{\alpha'_1 - 1 + \alpha'_0 - 1} = \frac{4}{5}$
Toss example

- MAP estimation can avoid overfitting
  - $\mathcal{D} = \{1,1,1\}, \hat{\theta}_{ML} = 1$
  - $\hat{\theta}_{MAP} = 0.8$ (with prior $p(\theta) = \text{Beta}(\theta|2,2)$)
Bayesian inference

- Parameters $\theta$ as random variables with a priori distribution
  - Bayesian estimation utilizes the available prior information about the unknown parameter
  - As opposed to ML and MAP estimation, it does not seek a specific point estimate of the unknown parameter vector $\theta$

- The observed samples $\mathcal{D}$ convert the prior densities $p(\theta)$ into a posterior density $p(\theta|\mathcal{D})$
  - Keep track of beliefs about $\theta$’s values and uses these beliefs for reaching conclusions
  - In the Bayesian approach, we first specify $p(\theta|\mathcal{D})$ and then we compute the predictive distribution $p(x|\mathcal{D})$
Bayesian estimation: predictive distribution

- Given a set of samples $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$, a prior distribution on the parameters $P(\theta)$, and the form of the distribution $P(x|\theta)$

- We find $P(\theta|\mathcal{D})$ and then use it to specify $\hat{P}(x) = P(x|\mathcal{D})$ as an estimate of $P(x)$:

$$P(x|\mathcal{D}) = \int P(x, \theta|\mathcal{D}) d\theta = \int P(x|\mathcal{D}, \theta) P(\theta|\mathcal{D}) d\theta = \int P(x|\theta) P(\theta|\mathcal{D}) d\theta$$

Predictive distribution

If we know the value of the parameters $\theta$, we know exactly the distribution of $x$

- Analytical solutions exist for very special forms of the involved functions
Benoulli likelihood: prediction

- Training samples: $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$

$$P(\theta) = \text{Beta}(\theta|\alpha_1, \alpha_0)$$
$$\propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}$$

$$P(\theta|\mathcal{D}) = \text{Beta}(\theta|\alpha_1 + m, \alpha_0 + N - m)$$
$$\propto \theta^{\alpha_1+m-1}(1-\theta)^{\alpha_0+(N-m)-1}$$

$$P(x|\mathcal{D}) = \int P(x|\theta) P(\theta|\mathcal{D})d\theta$$

$$= \mathbb{E}_{P(\theta|\mathcal{D})}[P(x|\theta)]$$

$$\Rightarrow P(x = 1|\mathcal{D}) = \mathbb{E}_{P(\theta|\mathcal{D})}[\theta] = \frac{\alpha_1 + m}{\alpha_0 + \alpha_1 + N}$$
ML, MAP, and Bayesian Estimation

- If $p(\theta|\mathcal{D})$ has a sharp peak at $\theta = \hat{\theta}$ (i.e., $p(\theta|\mathcal{D}) \approx \delta(\theta, \hat{\theta})$), then $p(x|\mathcal{D}) \approx p(x|\hat{\theta})$

  In this case, the Bayesian estimation will be approximately equal to the MAP estimation.

- If $p(\mathcal{D}|\theta)$ is concentrated around a sharp peak and $p(\theta)$ is broad enough around this peak, the ML, MAP, and Bayesian estimations yield approximately the same result.

- All three methods asymptotically ($N \to \infty$) results in the same estimate
Summary

- ML and MAP result in a single (point) estimate of the unknown parameters vector.
  - More simple and interpretable than Bayesian estimation

- Bayesian approach finds a predictive distribution using all the available information:
  - expected to give better results
  - needs higher computational complexity

- Bayesian methods have gained a lot of popularity over the recent decade due to the advances in computer technology.

- All three methods asymptotically \((N \to \infty)\) results in the same estimate.
Resource