# Parameter Estimation

- Maximum-Likelihood Estimation
- Bayesian Estimation

*Generative approach*

- We could design an optimal classifier if we knew:
  - $P(\omega_i)$ (priors)
  - $p(x \mid \omega_i)$ (class-conditional densities)
- Unfortunately, we rarely have this complete information!
- We have some knowledge and training data

  Training data set $\{(\mathbf{x}_i, \omega_i)\}$

- Design a classifier from training data
- Use the samples to estimate the unknown probability distributions

1

- Difficult to estimate an unknown p(x)
  especially in high dimensional case

- Assume a priori information about the problem: e.g., the parametric families of probability distributions $p(x|\theta)$

- E.g., Normality of $p(\mathbf{x} \mid \omega_i)$

  $$p(\mathbf{x} \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - Characterized by parameters $\mu_i$, $\Sigma_I$

- This knowledge significantly simplifies the problem, from one of estimating an unknown function p(x) to one of estimating the parameters $\theta$

2

## Parameter estimation

- Use a set D of training samples drawn independently from the probability distribution $p(x|\theta)$ to estimate the unknown parameter vector $\theta$

- A classic problem in statistics
  - Maximum-Likelihood (ML) and the Bayesian estimations

3

- Maximum-Likelihood Estimation
  - Simpler than any other alternative techniques
  - Suppose that D contains n samples, $x_1, x_2, \ldots, x_n$
  - Samples are i.i.d.—independent and identically distributed random variables.

$$p(D \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta) = L(\theta)$$

$p(D \mid \theta)$ is called the likelihood of $\theta$ w.r.t. the set of samples

  - ML estimate of $\theta$ is, by definition the value $\hat{\theta}$ that maximizes $p(D \mid \theta)$

    "It is the value of $\theta$ that best agrees with the actually observed training sample"

4

- Optimal estimation
  - For analytical purposes, it is usually easier to work with the logarithm of the likelihood
  - We define $LL(\theta)$ as the *log-likelihood function*

$$LL(\theta) = \ln p(D \mid \theta)$$

$$LL(\theta) = \sum_{k=1}^{k=n} \ln p(x_k \mid \theta)$$

Log-likelihood is numerically more stable

  - Determine $\theta$ that maximizes the log-likelihood

$$\hat{\theta}_{ML} = \arg\max_{\theta} LL(\theta)$$

5

- Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let $\nabla_\theta$ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^t$$

- Set of necessary conditions for an optimum is

$$\nabla_\theta LL = 0$$

$$\left(\nabla_\theta LL = \sum_{k=1}^{k=n} \nabla_\theta \ln p(x_k \mid \theta)\right)$$

6

- Example of a specific case ML estimation:
  - Univariate Gaussian Case: *unknown* $\mu$ *and* $\sigma$

    $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$\ln p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta \ln p(x_k \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}(\ln p(x_k \mid \theta)) \\ \frac{\partial}{\partial\theta_2}(\ln p(x_k \mid \theta)) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{pmatrix}$$

7

Summation:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & \textbf{(1)} \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & \textbf{(2)} \end{cases}$$

Combining (1) and (2), one obtains:

$$\hat{\mu} = \sum_{k=1}^{k=n} \frac{x_k}{n} \quad ; \quad \hat{\sigma}^2 = \frac{\sum_{k=1}^{k=n} (x_k - \hat{\mu})^2}{n}$$

8

## Parameter Estimation: Discrete Case

- Binary variable

*P(X=1) = θ, P(X=0) = 1- θ*

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

*Bernoulli distribution*

- Let i.i.d. samples D = $\{x_1, x_2, \ldots, x_n\}$

$$P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{\sum_{i=1}^{n} (1-x_i)}$$

9

- *Sufficient statistics*:
  $N_1$: number of 1's in D, $N_0$: number of 0's in D

$$P(D \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

- A *sufficient statistic* is a function of the data that summarizes the relevant information needed to compute the likelihood
- Log-likelihood
  $LL(\theta) = N_1 \ln \theta + N_0 \ln (1 - \theta)$
- ML estimation

$$\hat{\theta}_{ML} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{n}$$

10

- Multi-valued discrete random variables {1, …, k}

$$\theta_i = P(X = i)$$

$$P(x \mid \theta) = \prod_{i=1}^{k} \theta_i^{\delta_{xi}}$$

$$\textit{where} \quad \text{the Kronecker} \quad \text{delta}$$

$$\delta_{xi} = \begin{cases} 1 & \text{x} = \text{i} \\ 0 & \text{x} \neq \text{i} \end{cases}$$

$$P(D \mid \theta) = \prod_{j=1}^{n} P(x_j \mid \theta) = \prod_{i=1}^{k} \theta_i^{N_i}$$

- Sufficient statistics $N_i$: the # of times i appears in D

11

- Multi-valued discrete random variables {1, …, k} using 1-of-*k* representation:
  The variable is represented by a k-dimensional vector ***x*** in which one element equals 1 and all remaining equal 0

$$P(\vec{x} \mid \vec{\theta}) = \prod_{i=1}^{k} \theta_i^{x_i}$$

$$P\left(D \middle| \vec{\theta}\right) = \prod_{j=1}^{n} (\vec{x}_j \mid \vec{\theta}) = \prod_{j=1}^{n} \prod_{i=1}^{k} \theta_i^{x_{ji}} = \prod_{i=1}^{k} \theta_i^{N_i}$$

  - Sufficient statistics $N_i$: the # of times i appears in D

12

- ML estimates

$$\hat{\theta}_{iML} = \frac{N_i}{\sum_j N_j} = \frac{N_i}{n}$$

13

- MLE Summary
  - Intuitively appealing
  - One of the most commonly used estimators in statistics
  - Asymptotically consistent - converges to the true value as the number of examples approaches infinity
  - Problem with ML estimate – unstable when estimating from small samples
    - – assigns zero probability to unobserved values
    - – can lead to difficulties when estimating from small samples

14

## Bayesian Estimation

- In MLE $\theta$ was supposed fix
- In BE $\theta$ is a random variable
- Our knowledge about $\theta$ is assumed to be contained in a known prior density $p(\theta)$
- Compute posterior density $p(\theta|D)$
- Our goal is to compute $p(x|D)$ which is as close as we can come to obtaining the unknown $p(x)$

15

"Compute the posterior density p(θ | D)" then "Derive p(x | D)"

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta$$

Using Bayes formula, we have:

$$p(\theta|\mathrm{D}) = \frac{p(\mathrm{D}|\theta)p(\theta)}{\int p(\mathrm{D}|\theta)p(\theta)d\theta}$$

And by i.i.d. assumption:

$$p(\mathrm{D}|\theta) = \prod_{k=1}^{k=n} p(x_k|\theta)$$

16

- The integration to obtain p(x|D) is often difficult to do
- Maximum A Posteriori (MAP) Estimators

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$
$$= \arg\max_{\theta} p(D|\theta)p(\theta)$$

- If p(θ|D) peaks very sharply at $\theta_{MAP}$, then p(x|D) can be approximated by p(x| $\theta_{MAP}$), that is, use $\theta_{MAP}$ as the estimate for the true parameter
- If p(D| θ) peaks sharply at $\theta_{ML}$, then $\theta_{MAP}$ is close to $\theta_{ML}$

17

## Bayesian Parameter Estimation: Discrete Case

- Single binary variable

  *P(X=1) = θ,   P(X=0) = 1- θ*

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

- Let i.i.d. samples D = {$x_1$, $x_2$, …, $x_n$}

$$P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

*Sufficient statistics*: $N_1$: number of 1's in D, $N_0$: number of 0's in D

18

Assume the prior is a Beta distribution

$$p(\theta) = Beta(\theta \mid \alpha_1, \alpha_0) = c\,\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}$$

The posterior density p(θ | D)

$$\begin{aligned} p(\theta \mid \mathrm{D}) &= c \cdot p(\mathrm{D} \mid \theta)\, p(\theta) \\ &= Beta(\theta \mid N_1 + \alpha_1, N_0 + \alpha_0) \end{aligned}$$

- The property that the posterior distribution follows the same parametric form as the prior distribution is called *conjugacy*
- The parameters $\alpha_1$ and $\alpha_2$ are often called *hyperparameters*

19

Beta distribution

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

$$0 \le \theta \le 1$$

$$E(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Gamma Function

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(1) = 1, \Gamma(x) = (x-1)! \quad \text{for interger } x$$



$$p(X=1 \mid D) = \int p(X=1 \mid \theta) p(\theta \mid D) d\theta$$

$$= \int \theta p(\theta \mid D) d\theta = \frac{N_1 + \alpha_1}{N_1 + N_0 + \alpha_1 + \alpha_0} \equiv \hat{\theta}_{BE}$$

- It can be proved that:
  If the prior is well-behaved – i.e. does not assign 0 density to any *feasible* parameter value, then both MLE and Bayesian estimate converge to the same value in the limit
- Both *almost surely* converge to the underlying distribution *P*(X)
- But the ML and Bayesian approaches behave differently when the number of samples is small

$$\hat{\theta}_{BE} = \frac{N_1 + \alpha_1}{N_1 + N_0 + \alpha_1 + \alpha_0}$$

- $\alpha_1$ and $\alpha_2$ can be interpreted as *effective number of observations* of X=1 and X=0 respectively, “*imaginary*” counts from our prior experience

- $\alpha_1 + \alpha_2$ is called *equivalent sample size*

22

- Multi-valued discrete random variables {1, …, k}

$$\theta_i = P(X = i)$$

$$P(D \mid \theta) = \prod_{j=1}^{n} P(x_j \mid \theta) = \prod_{i=1}^{k} \theta_i^{N_i}$$

Sufficient statistics $N_i$: the # of times i appears in D

- Assume the prior p(θ) is a Dirichlet distribution Dir(θ|α) with *hyperparameters* $\alpha_i$'s
- Then the posterior density p(θ | D) is also a Dirichlet distribution

$$p(\boldsymbol{\theta} \mid D) = c \cdot p(\mathrm{D} \mid \theta)\, p(\boldsymbol{\theta}) \\ = Dir(\boldsymbol{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})$$

23

Dirichlet distribution with hyperparameters $\alpha_i$'s

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

$$0 \leq \theta_i \leq 1, \quad \sum_{i=1}^{k} \theta_i = 1, \quad \alpha = \sum_{i=1}^{k} \alpha_i$$

$$E(\theta_i) = \frac{\alpha_i}{\alpha}$$

24

- Bayesian estimates

$$P(X = i \mid D) = \int p(X = i \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

$$= \int \theta_i Dir(\boldsymbol{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})\, d\boldsymbol{\theta} = \frac{N_i + \alpha_i}{n + \alpha} \equiv \hat{\theta}_{iBE}$$

- The hyperparameters $\alpha_i$ can be thought of as "*imaginary*" counts from our prior experience
- $\alpha$: imaginary *equivalent sample size*
- Let $p_i$ be prior belief about $\theta_i$ : $\alpha_i = \alpha\, p_i$
- The larger the equivalent sample size, the more confident we are in our prior
- *Laplace estimates:* $\alpha = k$, $\alpha_i = 1$

25

## Summary of Bayesian estimation

- Treat the unknown parameters as random variables
- Assume a prior distribution for the unknown parameters
- Update the distribution of the parameters based on data
- Finally compute $p(x|D)$