Parameter Estimation

- Maximum-Likelihood Estimation
- Bayesian Estimation



Generative approach

- We could design an optimal classifier if we knew:
 - P(ω_i) (priors)
 - p(x | ω) (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



- 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(x | ω_i)

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

- Characterized by parameters μ_i , Σ_i
- This knowledge significantly simplifies the problem, from one of estimating an unknown function p(x) to one of estimating the parameters θ

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

Maximum-Likelihood Estimation



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

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

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

ML estimate of θ is, by definition the value $\hat{\boldsymbol{\theta}}$ that maximizes p(D | θ) "It is the value of $\boldsymbol{\theta}$ that best agrees with the actually observed training sample"

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* $LL(\theta) = In p(D \mid \theta)$

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

Log-likelihood is numerically more stable

• Determine θ that maximizes the log-likelihood

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



• Let θ = $(\theta_1,\,\theta_2,\,...,\,\theta_p)^t$ and let ∇_θ be the gradient operator

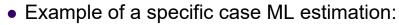
$$\nabla_{\theta} = \left[\frac{\partial}{\partial \theta_{1}}, \frac{\partial}{\partial \theta_{2}}, \dots, \frac{\partial}{\partial \theta_{p}}\right]^{t}$$

· Set of necessary conditions for an optimum is

$$\nabla_{\theta} LL = 0$$

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

6





• Univariate Gaussian Case: unknown μ and σ θ = (θ_1, θ_2) = (μ, σ^2)

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

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

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

Summation:



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

$$-\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$
 (2)

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

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

Parameter Estimation: Discrete Case



Binary variable

$$P(X=1) = \theta$$
, $P(X=0) = 1 - \theta$

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

Bernoulli distribution

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

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

• 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(θ) = N₁ ln θ + N₀ ln (1 θ)
- ML estimation

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

10

• Multi-valued discrete random variables {1, ..., k}

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

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

where the Kronecker delta

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

$$P(D \mid \theta) = \prod_{j=1}^{n} P(x_{j} \mid \theta) = \prod_{i=1}^{k} \theta_{i}^{N_{i}}$$

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

 Multi-valued discrete random variables {1, ..., k} using 1-of-k representation:



The variable is represented by a k-dimensional vector \mathbf{x} in which one element equals 1 and all remaining equal 0

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

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

Sufficient statistics N_i: the # of times i appears in D₁₂



ML estimates

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



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 θ was supposed fix
- In BE θ is a random variable
- Our knowledge about θ 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)

"Compute the posterior density $p(\theta \mid D)$ " then "Derive $p(x \mid D)$ "



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

Using Bayes formula, we have:

$$p(\theta \mid \mathsf{D}) = \frac{p(\mathsf{D} \mid \theta)p(\theta)}{\int p(\mathsf{D} \mid \theta)p(\theta)d\theta}$$

And by i.i.d. assumption:

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

16

The integration to obtain p(x|D) is often difficult to do



Maximum A Posteriori (MAP) Estimators

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{arg\,max}} p(\theta|D)$$

$$= \underset{\theta}{\operatorname{arg\,max}} p(D|\theta)p(\theta)$$

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

Bayesian Parameter Estimation: Discrete Case



Single binary variable

$$P(X=1) = \theta$$
, $P(X=0) = 1 - \theta$

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

• 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: N₁: number of 1's in D,

N₀: 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 $p(\theta \mid D)$

$$p(\theta \mid \mathsf{D}) = c \cdot p(\mathsf{D} \mid \theta) p(\theta)$$
$$= Beta(\theta \mid N_1 + \alpha_1, N_0 + \alpha_0)$$

- The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy
- The parameters α_1 and α_2 are often called hyperparameters

Beta distribution



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

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

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

20

$$p(X = 1 \mid D) = \int p(X = 1 \mid \theta) p(\theta \mid D) d\theta$$
$$= \int \theta p(\theta \mid D) d\theta = \frac{N_1 + \alpha_1}{N_1 + N_0 + \alpha_1 + \alpha_0} \equiv \hat{\theta}_{BE}$$

- 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



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

- α₁ and α₂ can be interpreted as effective number of observations of X=1 and X=0 respectively, "imaginary" counts from our prior experience
- α_1 + α_2 is called *equivalent sample size*

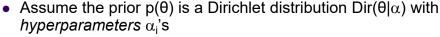
22

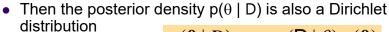
• Multi-valued discrete random variables {1, ..., k}

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

$$P(D \mid \theta) = \prod_{i=1}^{n} P(x_{i} \mid \theta) = \prod_{i=1}^{k} \theta_{i}^{N_{i}}$$

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





$$p(\mathbf{\theta} \mid D) = c \cdot p(\mathsf{D} \mid \theta) p(\mathbf{\theta})$$

 $= Dir(\mathbf{\theta} \mid \mathbf{N} + \boldsymbol{\alpha})$





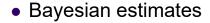
$$\operatorname{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

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

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

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

24





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

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

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)