Introduction to Pattern Recognition and Data Mining

Lecture 3: Parameter Estimation

Instructor: Dr. Giovanni Seni

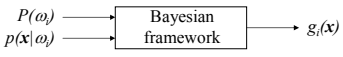
Department of Computer Engineering
Santa Clara University

Overview

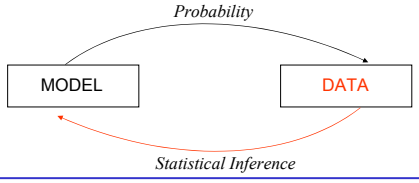
- Introduction
  - Statistical inference
  - Estimator's bias and variance
- Maximum Likelihood (ML) estimation
  - Binomial distribution
  - Normal distribution
  - Simple linear regression
- Bayesian estimation

Introduction
Statistical Inference

- Bayesian classifier



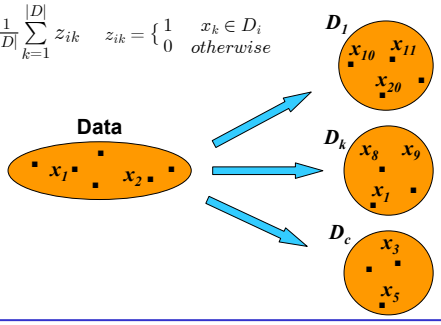
- Dual role of probability and statistical inference



Introduction
Statistical Inference (2)

- Estimation of priors is simple

$$\hat{P}(\omega_i) = \frac{1}{|D|}\sum_{k=1}^{|D|} z_{ik} \qquad z_{ik} = \begin{cases} 1 & x_k \in D_i \\ 0 & otherwise \end{cases}$$



Introduction
Statistical Inference (3)

- Suppose D contains n samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$
  - Assume c separate problems
- Premise 1 – *i.i.d.*
  - Samples have been drawn at random according to $p(\boldsymbol{x})$ – the model
  - Samples are independent
- Premise 2 – known parametric form
  - $p(\boldsymbol{x}|\boldsymbol{\theta})$ is determined uniquely by a parameter vector θ

Introduction
Statistical Inference (4)

- Probability of observed data arising under an implicitly assumed model M

$$p(D|\theta, M) = \prod_{i=1}^{n} p(x_i|\theta, M)$$

  - θ are the parameters of the model
  - when regarded as a function of θ, it is called the likelihood $L(\theta|D)$
- We use $p(D|\theta, M)$ to decide how realistic the assumed model is
  - Reject/change model if the likelihood is low

Introduction
Statistical Inference (5)

- Let $\hat{\theta}$ be an estimator of a θ
  - $\hat{\theta}$ is a random variable, with different values arising as different samples are drawn (e.g., by repeatedly subsampling original data set)
- Measures of quality
  - $Bias(\hat{\theta}) = \varepsilon(\hat{\theta}) - \theta$:
    - reflects any systematic error in our prediction
  - $Var(\hat{\theta}) = \varepsilon(\hat{\theta} - \varepsilon(\hat{\theta}))^2$:
    - measures how much our estimates will vary across different data sets (sensitivity to particular training data set)

Maximum-Likelihood Estimate
θ_{ML}

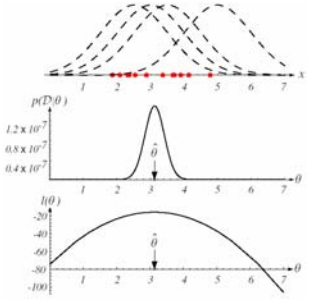
- $\hat{\theta}$ that maximizes $L(\theta|D)$
  - value of θ that best agrees with or supports the observed training samples
- Often more convenient to work in log domain $l(\theta|D)$
- Assuming a well-behaved, differentiable function

$$\hat{\theta} = \arg\max_{\theta} l(\theta|D) \qquad l(\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

Solve $\nabla_\theta l = 0$ $\qquad \nabla_\theta l = \sum_{i=1}^{n} \nabla_\theta \ln p(x_i|\theta)$

Maximum-Likelihood Estimate
Example

- Assumed model: $p(x|\theta) \sim N(\theta, \sigma)$



Maximum-Likelihood Estimate
Binomial Distribution

- Assumed model: $P(x|\theta) = \theta^x(1-\theta)^x \;\; ; \;\; x = 0,1$
- Scenario: customers at a supermarket either purchase or don't purchase milk; θ is the probability that milk is purchased by a random customer

$$L(\theta|x_1,\ldots,x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^r(1-\theta)^{n-r}$$

where r is number among n sample customers who bought milk

$$l(\theta) = r\ln\theta + (n-r)\ln(1-\theta)$$

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{r}{\theta} - \left(\frac{n-r}{1-\theta}\right) = 0 \;\; \Rightarrow \;\; \hat{\theta}_{ML} = \frac{r}{n}$$

Maximum-Likelihood Estimate
Normal Density – Unknown μ

- Consider a single point $\boldsymbol{x}_i$

$$p(x_i|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_i-\theta)^t\Sigma^{-1}(x_i-\theta)}$$

$$\ln p(x_i|\theta) = -\frac{1}{2}\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}(x_i-\theta)^t\Sigma^{-1}(x_i-\theta)$$

$$\nabla_\theta \ln p(x_i|\theta) = \Sigma^{-1}(x_i-\theta)$$

For the full log-likelihood: $\nabla_\theta l = \sum_{i=1}^{n} \Sigma^{-1}(x_i-\theta) = 0$

$$\Rightarrow \boxed{\hat{\mu}_{ML} = \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$ the sample mean!

Maximum-Likelihood Estimate
Normal Density – Unknown μ and Σ

- Univariate case $p(x_i|\theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x_i-\theta_1)^2}$

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

$$\nabla_\theta \ln p(x_i|\theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_i-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

For the full log-likelihood: $\frac{\partial l}{\partial\theta_1} = 0 \;\; \Rightarrow \;\; \sum_{i=1}^{n} \frac{1}{\theta_2}(x_i-\theta_1) = 0 \;\; \Rightarrow \;\; \hat{\theta}_1 = \hat{\mu}_{ML}$

$$\frac{\partial l}{\partial\theta_2} = 0 \;\; \Rightarrow \;\; -\sum_{i=1}^{n}\frac{1}{\theta_2} + \sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{\theta_2^2} = 0 \;\; \Rightarrow \;\; \hat{\theta}_2 = \boxed{\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\theta}_1)^2}$$

Maximum-Likelihood Estimate
Normal Density – Unknown μ and Σ (2)

- Multivariate case…

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^t$$

- $\hat{\mu}_{ML}$ is unbiased

$$\mathcal{E}(\hat{\mu}) = \mathcal{E}\left[\frac{1}{n}(x_1 + x_2 + \ldots + x_n)\right] = \frac{1}{n}\mathcal{E}(x_1 + x_2 + \ldots + x_n)$$
$$= \frac{1}{n}(\mathcal{E}(x_1) + \mathcal{E}(x_2) + \ldots + \mathcal{E}(x_n)) = \frac{1}{n}(\mu + \mu + \ldots + \mu)$$
$$= \mu$$

Maximum-Likelihood Estimate
Normal Density – Unknown μ and Σ (3)

- $\hat{\sigma}_{ML}$ is biased

$$\mathcal{E}(\hat{\sigma}) = \mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right]$$
$$= \mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2\right] - \mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu} - \mu)^2\right]$$
$$= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

⇒ but asymptotically unbiased!

Maximum-Likelihood Estimate
Simple Linear Regression

- Assumed model: $Y = a + bX + e$
- Data: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- e error term; a random variable assumed to be $\sim N(0, \sigma)$; we can write $e = Y - (a + bX)$

$$p(e_i \mid \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - (a + bx_i)}{\sigma}\right)^2}$$

- Likelihood

$$L(a,b \mid \theta) = \prod_{i=1}^{n} p(e_i \mid \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2}$$

Maximum-Likelihood Estimate
Simple Linear Regression (2)

- Log-likelihood:

$$l(a,b \mid \theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

- To maximize $l(a,b|D)$ we need to minimize the sum of squared differences

  ⇒ Least Squares Method!

- LS method arises naturally from the choice of a Normal distribution for the error term in the model

Bayesian Estimation
Overview

- *Frequentist* view of probability:
  - Probability is an objective property of the outside world
  - Probability of an event as a "limiting proportion"
    - Tossing a coin
    - Customer buying milk
    - Not one-off events
  - Intrinsic variability lies in the data D
  - θ is fixed but unknown
- Subjective (Bayesian) probability
  - Probability is an individual belief that event will occur
  - Subjective component given as a prior – initial belief event will happen

Bayesian Estimation
Overview (2)

- Subjective (Bayesian) probability
  - θ is a random variable having a distribution of possible values
  - i.e., Known prior density $p(\theta)$
    - Broad and flat if we aren't' very sure
  - Information in D leads to a modification of this distribution to a posterior density $p(\theta|D)$
    - Which, we hope, is sharply peaked about the true value of θ
  - Maximum a posteriori method (MAP)
    - Pick the mode of the distribution
    - ML estimator is MAP estimator for a uniform $p(\theta)$

Bayesian Estimation
General Theory

- To obtain $p(x|D) = p(x|\omega_j, D_j)$ (to build our classifier)
- Compute $p(x \mid D) = \int p(x|\theta)\, p(\theta \mid D)\, d\theta$

  where form of $p(x|\theta)$ is assumed known (as before)

  and $p(\theta|D) \propto \underbrace{p(D|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{parameter prior}}$

Bayesian Estimation
Normal Density – Unknown μ

- Assumed model: $p(x|\theta) \sim N(\theta_\mu, \sigma^2)$
- Assumed prior: $p(\theta_\mu) \sim N(\mu_0, \sigma_0^2)$

$$\Rightarrow \quad p(\theta_\mu|D) = \alpha \prod_{i=1}^{n} p(x_i \mid \theta_\mu)\, p(\theta_\mu)$$

- Easily shown that $p(\theta_\mu|D) \sim N(\mu_n, \sigma_n^2)$ where

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \quad ; \text{ where } \hat{\mu}_n \text{ is sample mean}$$

  i.e., μ_n represents our best guess for θ_μ after observing n samples

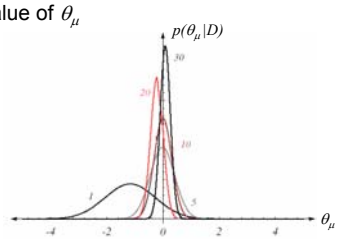
  Consider $n \to \infty$, $\sigma_0 \approx 0$, and $\sigma_0 >> \sigma$

Bayesian Estimation
Normal Density – Unknown μ (2)

and

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2} \quad \Rightarrow \quad \lim_{n \to \infty} \sigma_n^2 = \frac{\sigma^2}{2}$$

- i.e., each additional observation decreases our uncertainty about the true value of θ_μ



Bayesian Estimation
Normal Density – Unknown μ (3)

- We now can compute the class-conditional density

$$p(x \mid D) = \int p(x \mid \theta_\mu) p(\theta_\mu \mid D) d\theta_\mu$$
$$= \int N(\theta_\mu, \sigma^2) N(\mu_n, \sigma_n^2) d\theta_\mu$$
$$\sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- i.e., in $p(x|\theta) \sim N(\theta_\mu, \sigma^2)$ we set $\theta_\mu = \mu_n$ and replace σ^2 with $\sigma^2 + \sigma_n^2$
  - Treat $\mu_n = \alpha \cdot \hat{\mu}_n + \beta \cdot \mu_0$ as if it were the true mean
  - Increase the known variance σ^2 to account for the additional uncertainty resulting from our lack of exact knowledge of the mean

Parameter Estimation
When ML and Bayesian Methods Differ?

- Equivalent in the asymptotic limit of infinite training data or with a "flat" or uniform prior
- Computational complexity
  - ML uses Differential Calculus or gradient search for $\hat{\theta}$
  - B requires complex multidimensional integration
- Interpretability
  - ML returns a single best model/parameter
  - B gives a weighted average
- Confidence in prior information
  - ML solution is of assumed parametric form… not necessarily so in B approach