Estimation of Unknown Prob. Density Functions

- Lecture 4 -

- Until now, we assumed that pdf's are known
- This is not the common case:
  - ↳ in many problems, the underlying pdf has to be estimated from the available data.
- suppose that we can reasonably assume $p(x|\omega_i)$ is a Normal Density with mean μ_i and cov. matrix Σ_i (although we don't know the exact values of these quantities)
  - ↳ the problem is simplified then from estimating an unknown function $p(x|\omega_i)$ to the one of estimating the parameters μ_i and Σ_i

Maximum Likelihood Parameter Estimation

- views the parameters as quantities whose values are fixed but unknown!
- maximizes the probability of obtaining the samples actually observed.

Suppose we have a collection of samples from c classes:

$$D = \{D_1, D_2, \ldots D_c\} \quad \text{: dataset in c classes}$$

- samples from D_j have been drawn independently according to $p(\vec{x} \mid \omega_j)$

such samples are i.i.d → independent and identically distri

Assume $p(\vec{x} \mid \omega_j)$ has a known parametric form,
determined uniquely by $\vec{\theta}_j$

<u>ex</u>: $p(\vec{x} \mid \omega_j) \sim N(\mu_j, \Sigma_j) \rightarrow \vec{\theta}_j = \{\mu_j, \Sigma_j\}$

to show the dependence of $p(\vec{x} \mid \omega_j)$ on θ_j:

$$p(\vec{x} \mid \omega_j, \theta_j)$$

<u>AIM:</u> use the information provided by the training samples to obtain good estimates for the unknown vectors:

$$\{\theta_1, \theta_2, \ldots \theta_c\}$$

Now, assume that samples D_i give no information about θ_j.

↳ Hence, parameters for the different classes are functionally independent.

↳ So, we can work with each class separately:

call θ to params of a class (not using subscript θ_j anymore)

Suppose D contains n samples: $\{x_1, x_2, \ldots x_n\}$. Since samples are i.i.d:

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

likelihood function of θ w.r.t the set of samples D

Find the value $\hat{\theta}$, that maximizes $p(D|\theta)$

Find $\hat{\theta}$ that maximizes $p(D|\theta)$:

For analytic simplicity, use logarithm of the likelihood

↳ monotonically increasing function : no problem.

if the # of params is p: $\vec{\theta} = [\theta_1, \theta_2, \dots \theta_p]^t$

$\nabla \rightarrow$ gradient op. $\quad \nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \dots \frac{\partial}{\partial\theta_p}\right]^t$

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ l(\theta)$$

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

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

$$\nabla_\theta \, l(\theta) = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k|\theta)$$

$\nabla_\theta \, l(\theta) = 0 \quad \rightarrow$ set of p equations.

- solution to $\hat{\theta}$ could represent a true global max, or a local max } check each sln. individually.

Example: The Gaussian Case: Unknown μ

Assume that samples are drawn from a multivariate normal population with mean μ and cov. Σ. Assume only μ unknown!

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

$$\nabla_\mu \ln p(x_k|\mu) = \Sigma^{-1}(x_k-\mu)$$

the maximum likelihood est. of μ must satisfy:

$$\sum_{k=1}^{n} \Sigma^{-1}(x_k-\hat{\mu}) = 0$$

$$\sum_{k=1}^{n} (x_k-\hat{\mu}) = 0$$

$$\sum_{k=1}^{n} \hat{\mu} = \sum_{k=1}^{n} x_k$$

$$n.\hat{\mu} = \sum_{k=1}^{n} x_k$$

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k}$$

arithmetic avg. of the observed samples

Example 2: The Gaussian Case: Unknown μ and σ^2 → Univariate Normal Density

$$\theta = \{\underset{\theta_1}{\mu}, \underset{\theta_2}{\sigma^2}\}$$ → parameters to be estimated

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

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

$\nabla_\theta l = 0$ (max. likelihood cond)

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

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

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