# Maximum-likelihood Parameter Estimation

# Parameter Estimation

- Bayes classifier is the best classifier.
- But, we should know about the prior probabilities $P(\omega_i)$ and class-conditional densities $p(X|\omega_i)$.
- That means the probabilistic structure of the problem should be known.
- In general, what is given to us is only a training set, not the probabilistic structure !
- Never-the-less we can assume some thing like: *the distribution is Normal* or so, based on the domain. And then, estimate its parameters based on the training set (eg: mean, covariance matrix can be estimated from the sample).

# Density Estimation

- There are two broad ways in which the probability densities can be estimated from the training set.
- These are :
  1. *Parametric methods*
  2. *Non-parametric methods*

# Parametric Methods

- We assume the form of the distribution (eg: Normal) and estimate its parameters (eg: mean and covariance matrix).

  *Two broad parameter estimation methods are:*

  1. maximum-likelihood estimation, and
  2. Bayesian estimation.

# Non-parametric Methods

- We do not assume any thing about the form of the distribution, but we use the training examples directly to estimate the density at a given point.

  *Two broad ways are:*

  1. Parzen window based, and
  2. nearest neighbor based.

# Maximum-likelihood method

- We study about maximum-likelihood parameter estimation.
- Here, we assume that the parameters are unknown but fixed.
- The other parameter estimation method, viz., Bayesian parameter estimation method assumes that *the parameters are unknown and random variables*.
- It is found that, both methods, frequently gives same results.
- Maximum-likelihood method is simpler than the Bayes method.

# Maximum-likelihood method

- Training set is divided class-wise.
- We consider only one class's training set at a time.
- Let the parameters we are trying to estimate, for the class, be $\theta$. For example $\theta = (\mu, \Sigma)^t$, if the distribution is assumed to be a Normal one.

# Maximum-likelihood: General Principle

- Let $\mathcal{D}$ be the training set for the class.
- Let the patterns in $\mathcal{D}$ are independently and identically drawn(i.i.d).
- Suppose that $\mathcal{D}$ contains n samples, $X_1, \ldots, X_n$.
- Then,

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

# Maximum-likelihood: General Principle

- $p(\mathcal{D}|\theta)$, when viewed as a function of $\theta$, is called likelihood of $\theta$ with respect to the set of samples.
- The *maximum-likelihood estimate* of $\theta$ is, by definition, the value $\hat{\theta}$ that maximizes $p(\mathcal{D}|\theta)$.
- Intuitively, this estimate corresponds to the value of $\theta$ that in some sense best agrees with the training set.

# To simplify analytically

- It is usually easier to work with the logarithm of the likelihood than with the likelihood itself.
- Because the logarithm is monotonically increasing, the $\hat{\theta}$ that maximizes the log-likelihood also maximizes the likelihood.
- If $p(\mathcal{D}|\theta)$ is well-behaved, differentiable function of $\theta$, $\hat{\theta}$ can be found by the standard methods of differential calculus.

# Maximum-likelihood ...

- Let the parameter vector $\theta = (\theta_1, \ldots, \theta_p)^t$. That is, there are $p$ parameters to be estimated.
- Let $\nabla_\theta$ be the gradient operator,

$$\nabla_\theta \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

# Maximum-likelihood ...

- We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln p(\mathcal{D}|\theta)$$

$$= \ln \left(\prod_{k=1}^{n} p(X_k|\theta)\right)$$

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

# Maximum-likelihood ...

- $$\nabla_\theta \, l = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial l}{\partial \theta_p} \end{bmatrix}$$

- Thus, a set of necessary conditions for the maximum-likelihood estimate for $\theta$ can be obtained from a set of $p$ equations

$$\nabla_\theta \, l = \mathbf{0}$$

# Maximum-likelihood ...

- Let solution to $\nabla_\theta \, l = \mathbf{0}$ be $\hat{\theta}$.
- $\hat{\theta}$ could represent a true global maximum, a local maximum or minimum, or (rarely) an inflection point of $l(\theta)$.
- One must be careful regarding the above aspect. One remedy is, to find all solutions and findout from them which is the actual solution.
- In case of Normal distribution, we do not get these problems.

# The Gaussian Case: Unknown $\mu$

- Assume that, only $\mu$ is unknown, and we want to find the maximum-likelihood estimate for this.
- $\theta = [\mu]$.
- $$\ln p(X_k|\mu) \; = \; -\frac{1}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}(X_k - \mu)^t\Sigma^{-1}(X_k - \mu)$$

  and

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

# The Gaussian Case: Unknown $\mu$

- The log-likelihood is,

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

Hence,

$$\nabla_\mu \, l \;=\; \sum_{k=1}^{n} \nabla_\mu \ln p(X_k|\mu) \;=\; \sum_{k=1}^{n} \Sigma^{-1}(X_k - \mu)$$

- When we equate the above to zero, we get

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

# The Gaussian Case: Unknown $\mu$

- $$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

  is a very satisfying result.
- It says that the *sample mean* is the maximum-likelihood estimate for the mean.
- Sample mean is nothing but *centroid* of the set of patterns.

# Unknown $\mu$ and $\sigma^2$

- Consider Univariate case.
- $\theta = (\theta_1, \theta_2)^t = (\mu, \sigma^2)^t$
  We know,

$$p(X_k|\theta) = \frac{1}{\sqrt{2\pi\theta_2}} \text{exp} \left[ -\frac{1}{2} \frac{(X_k - \theta_1)^2}{\theta_2} \right]$$

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

# Unknown $\mu$ and $\sigma^2$

$$\nabla_\theta \ln p(X_k|\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \end{bmatrix} \left[ -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(X_k - \theta_1)^2 \right]$$

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

# Unknown $\mu$ and $\sigma^2$

- Maximum likelihood estimate for $\theta$ is obtained at

$$\sum_{k=1}^{n} \nabla_\theta \ln p(X_k|\theta) = \mathbf{0}$$

- That is,

$$\sum_{k=1}^{n} \frac{1}{\theta_2}(X_k - \theta_1) \;=\; 0 \qquad (1)$$

$$-\sum_{k=1}^{n} \frac{1}{\theta_2} + \sum_{k=1}^{n} \frac{(X_k - \theta_1)^2}{\theta_2^2} \;=\; 0 \qquad (2)$$

# Unknown $\mu$ and $\sigma^2$

- We get,

$$\theta_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$$\theta_2 = \hat{\sigma^2} = \frac{1}{n}\sum_{k=1}^{n} (X_k - \hat{\mu})^2$$

# Multivariate case: Unknown $\mu$ and $\Sigma$

- It can be found similar to univariate case.
- We get,

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

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (X_k - \hat{\mu})(X_k - \hat{\mu})^t$$

# A problem

Consider univariate case. Let $X$ have an exponential density

$$p(X|\theta) = \begin{cases} \theta e^{-\theta X} & X \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Given $\{X_1, \ldots, X_k\}$, the i.i.d drawn training set, find the maximum-likelihood estimate of $\theta$.