# MLE and KLD

CS6015 – LINEAR ALGEBRA AND RANDOM PROCESSES

- We could design an optimal classifier if we knew the prior probabilities $P(w_i)$ and the class conditional densities $p(\mathbf{x}|w_i)$.
- Unfortunately, we rarely have this kind of complete knowledge about the probabilistic structure of the problem.
- **One approach to this problem:** Use the samples to estimate the unknown probabilities and probability densities, and then use the resulting estimates as if they were the true values.

- In supervised pattern classification problems, the estimation of the **prior probabilities** presents no serious difficulties.
- However, estimation of **class conditional densities** is quite a difficult thing to do. Samples are often too small for class-conditional estimation (and large dimension of feature space!)
- If we know the number of parameters of the distribution in advance and our general knowledge about the problem permits us to parameterize the conditional densities, then the severity of these problems can be reduced significantly.

# Example

- Assume that $p(\mathbf{x}|w_i)$ is a normal density with mean $\boldsymbol{\mu_i}$ and covariance $\boldsymbol{\Sigma_i}$, although we do not know the exact values of these quantities.
- This knowledge simplifies the problem from one of estimating an unknown function $p(\mathbf{x}|w_i)$ to one of estimating the parameters $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$.

# Problem of parameter estimation

- Two common ways of approaching this problem:
  - ❖ Maximum likelihood estimation
  - ❖Bayesian estimation

**MLE** : Views the parameters as quantities whose values are fixed but unknown.

**Bayesian Methods :** View the parameters as random variables having some known prior distribution.

In either case, posterior densities are used for the classification rule.

# Maximum likelihood estimation

- It has good convergence properties as the number of training samples increases.
- **General Principle :**
  - Assume we have ***c*** classes, so that we have c datasets, $D_1, \dots, D_c$ and the samples have been drawn independently according to the probability law $p(\mathbf{x}|w_i)$.
  - Such samples are i.i.d (independent and identically distributed) random variables.
  - Assume that $p(\mathbf{x}|w_i)$ has a known parametric form, and is therefore determined uniquely by the value of a parameter vector $\theta_j$.
  - For example, $p(\mathbf{x}\,|\omega_j) \sim N(\ \ _j, \Sigma_j)$
  - In general $p(\mathbf{x}\,|\omega_j) \equiv p(\mathbf{x}\,|\omega_j, \theta_j)$
  - **Goal**: to obtain good estimates for the unknown parameter vectors $\theta_1, \dots, \theta_c$.

- For simplicity, assume that the parameters for the different classes are functionally independent.
- Use a set $D$ of training samples drawn independently from the probability density $p(\mathbf{x}|\theta)$ to estimate the unknown parameter vector $\theta$.
- Suppose that $D$ contains n samples $\mathbf{x_1}, \ldots, \mathbf{x_n}$,we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$$

- The maximum likelihood estimate of $\theta$ is, by definition, the value $\hat{\theta}$ that maximizes $p(D|\theta)$ (which is the likelihood of $\theta$ w.r.t the set of samples) .

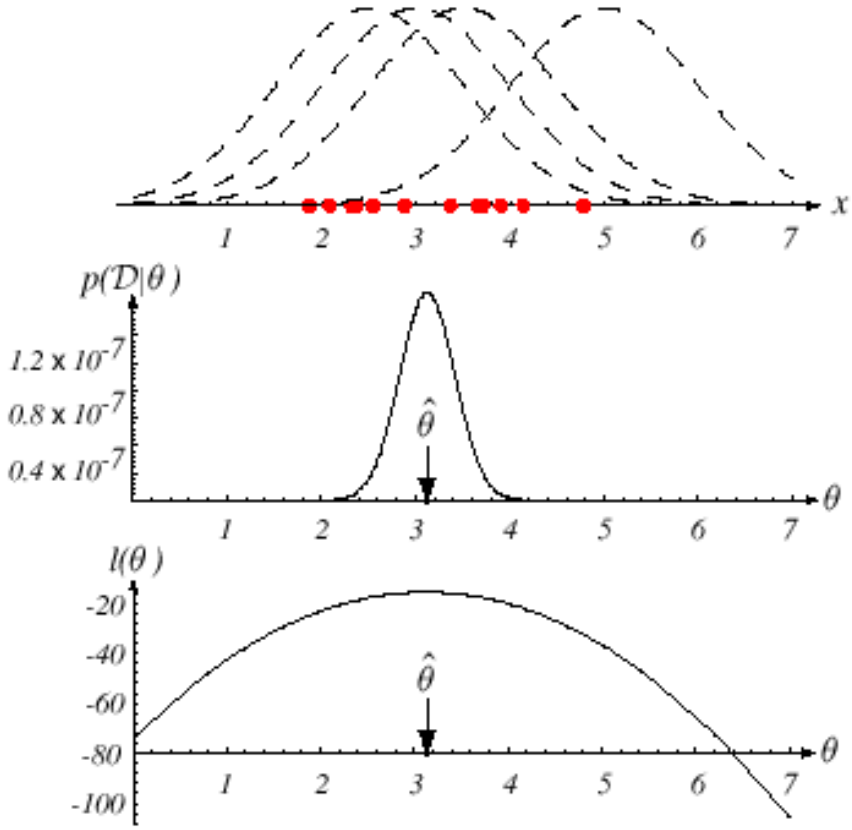


**FIGURE 3.1.** The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of $\theta$ whereas the conditional density $p(x|\theta)$ is shown as a function of $x$. Furthermore, as a function of $\theta$, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. 

- For analytical purposes, it is easier to work with the logarithm of the likelihood than with the likelihood itself (as logarithm is monotonically increasing).
- If the number of parameters to be estimated is $p$, then we let $\theta$ denote the $p$-component vector $\theta = \left(\theta_1, \ldots, \theta_p\right)^t$, and we let $\nabla_\theta$ be the gradient operator

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}.$$

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

$$l(\boldsymbol{\theta}) \equiv \ln\, p(\mathcal{D}|\boldsymbol{\theta}).$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}),$$

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}}\, \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}).$$

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

$$\boxed{\nabla_{\boldsymbol{\theta}} l = \mathbf{0}.}$$

# The Gaussian Case: Unknown $\mu$

- Suppose that the samples are drawn from a multivariate normal population with mean $\mu$ and covariance matrix $\Sigma$.

**CASE 1**: For simplicity, consider the case where only the mean is unknown.

- Under this condition, we consider a sample point $x_k$ and find

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

$$\nabla_{\boldsymbol{\theta}}\ln\, p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}).$$

- The MLE for $\mu$ must satisfy $\sum_{k=1}^{n}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0},$

- Multiplying by $\Sigma$ and rearranging, we obtain,

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

- It says that the MLE for the unknown population mean is just the arithmetic average of the training samples – the sample mean.

# The Gaussian Case: Unknown $\mu$ and $\Sigma$

- In general multivariate normal case, neither the mean $\mu$ nor the covariance matrix $\Sigma$ is known.
- These unknown parameters constitute the components of the parameter vector $\theta$.
- Consider first the univariate case with $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.
- The log-likelihood of a single point is

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

- And its derivative is $\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln\, p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}.$

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

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

- Where $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the maximum likelihood estimates for $\theta_1$ and $\theta_2$, respectively. By substituting $\hat{\mu} = \widehat{\theta_1}$ and $\widehat{\sigma^2} = \widehat{\theta_2}$ and doing a little rearranging, we obtain the following MLE for $\mu$ and $\sigma^2$.

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

- For multivariate case,

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

- Thus once again we find that the MLE for the mean vector is the sample mean.
- The MLE for the covariance matrix is the arithmetic average of the n matrices $(x_k - \hat{\mu})(x_k - \hat{\mu})^t$

# Bias

- The MLE for the variance $\sigma^2$ is biased; that is, the expected value over all data sets of size n of the sample variance is not equal to the true variance.

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

- The MLE of the covariance matrix is similarly biased.
- An unbiased estimate is given by:

$$\mathbf{C}=\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t$$

# Entropy

- The entropy is the average uncertainty of a single random variable.
- Let $\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{x})$;
- $\boldsymbol{H}(\boldsymbol{p}) = \boldsymbol{H}(\boldsymbol{X}) = -\sum_{\boldsymbol{x} \in \boldsymbol{X}} \boldsymbol{p}(\boldsymbol{x}) \log_{\boldsymbol{2}} \boldsymbol{p}(\boldsymbol{x})$
- In other words, entropy measures the amount of information in a random variable. It is normally measured in bits.

https://view.officeapps.live.com/op/view.aspx?src=http://www.site.uottawa.ca/~nat/Courses/NLP-Course/NLP-Lecture-5.ppt

# Joint Entropy and Conditional Entropy

- The ***joint entropy*** of a pair of discrete random variables $X, Y \sim p(x, y)$ is the amount of information needed on average to specify both their values.
- $\boldsymbol{H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(X, Y)}$
- The ***conditional entropy*** of a discrete random variable $Y$ given another $X$, for $X, Y \sim p(x, y)$, expresses how much extra information you still need to supply on average to communicate $Y$ given that the other party knows $X$.
- $\boldsymbol{H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y|x)}$
- ***Chain Rule*** for Entropy: $\boldsymbol{H(X, Y) = H(X) + H(Y|X)}$

# Relative Entropy or Kullback-Leibler Divergence

- For 2 pmfs, $p(x)$ and $q(x)$, their ***relative entropy*** is:

$$D(p||q) = \Sigma_{\boldsymbol{x}\in\boldsymbol{X}}\, \boldsymbol{p}(\boldsymbol{x})\log(\boldsymbol{p}(\boldsymbol{x})/\boldsymbol{q}(\boldsymbol{x}))$$

- The relative entropy (also known as the ***Kullback-Leibler divergence***) is a measure of how different two probability distributions (over the same event space) are.
- The KL divergence between $p$ and $q$ can also be seen as the average number of bits that are wasted by encoding events from a distribution $p$ with a code based on a not-quite-right distribution $q$.