ICS663: Pattern Recognition

Department of Information and Computer Sciences
University of Hawai`i at Manoa

Kyungim Baek

Announcement

- Homework assignment #1
  - Due: **Monday September 21, by 5:00 PM**
- Project proposal
  - Due: **Wednesday Septmber 23, by 5:00 PM**
  - Presentation (~ 10 minutes)
    - Monday (9/28)
      - BJ, Thomas, Tyson, Sharif, Danny, Jeremy
    - Wednesday (9/30)
      - Tetsuya, Kelly, Nurit
- **Exam I: Wednesday, October 7**

Previously…

- **Parametric density estimation**
  - Estimating $P(\omega_i)$ is not very difficult in the context of supervised learning
  - Estimating $p(\mathbf{x} | \omega_i)$ is hard, especially in high-dimensional feature spaces and when the number of training samples available seems to be too small
    - Assumption: $p(\mathbf{x} | \omega_i) = p(\mathbf{x} | \omega_i, \theta)$ – *known* parametric form, $\exists$ a 'true' θ_0
  - Two approaches
    - Maximum-Likelihood estimation
    - Bayesian estimation

Lecture 6

- Parametric Density Estimation
  - Maximum-Likelihood Estimation

Previously…

- **Maximum-Likelihood (ML) estimation:** *parameters of probabilistic distributions are fixed* but unknown values
  - Parameters are unknown constants to be identified through training
  - Best estimate of parameter values is achieved by maximizing the probability of obtaining the samples observed

Maximum-Likelihood Estimation

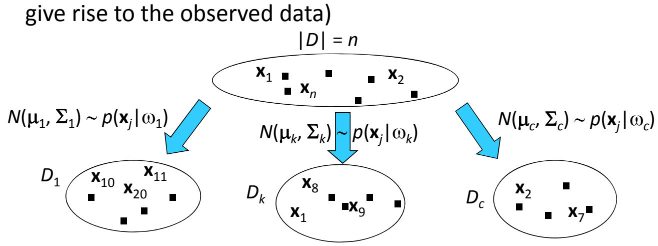
- General principle
  - Assume we have c classes and $p(\mathbf{x} \mid \omega_i) \equiv p(\mathbf{x} \mid \omega_i, \theta_i)$
    - For example, if $p(\mathbf{x} \mid \omega_i, \theta_i) \sim N(\mu_i, \Sigma_i)$ then

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (\mu_i^1, \mu_i^2, \ldots, \sigma_i^{11}, \sigma_i^{22}, \ldots, \mathrm{cov}(x_i^m, x_i^n), \ldots)$$

  - Use the information provided by the training samples to estimate $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, each θ_i ($i = 1, 2, \ldots, c$) is associated with each category
  - Assumption: independent and identically distributed (i.i.d.) samples

ML Problem Statement

- Let $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and divided into c data sets according to the class as $D_1, D_2, \ldots, D_c$
- Our goal is to determine θ_{ML} (value of θ that is most likely to give rise to the observed data)



- Find $\boldsymbol{\theta}_i^{ML}$ such that $\boldsymbol{\theta}_i^{ML} = \max_{\boldsymbol{\theta}_i} p(D_i \mid \boldsymbol{\theta}_i) = \max_{\boldsymbol{\theta}_i} \prod_{\mathbf{x}_k \in D_i} p(\mathbf{x}_k \mid \boldsymbol{\theta}_i)$

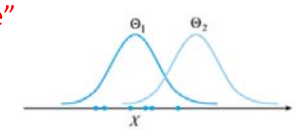
ML Estimation (cont'd)

- Suppose that D contains n samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The likelihood of θ with respect to D is:

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta})$$

- ML estimate of θ is, by definition the value θ_{ML} that maximizes $p(D \mid \theta)$

  "It is the value of θ that best agrees with the actually observed training sample"



Log-Likelihood

- The value θ of maximizing the likelihood function also maximizes its logarithm, known as the **log-likelihood** function
$$l(\boldsymbol{\theta}) = \ln[p(D \mid \boldsymbol{\theta})] = \sum_{k=1}^{n} \ln[p(\mathbf{x}_k \mid \boldsymbol{\theta})]$$
- ML estimate
  - Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_{θ} be the gradient operator
$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p}\right]^t$$
  - New problem statement: determine θ that maximizes the log-likelihood $\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$
  - ML condition: $$\nabla_{\boldsymbol{\theta}}[l(\boldsymbol{\theta})] = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}}[\ln p(\mathbf{x}_k \mid \boldsymbol{\theta})] = 0$$

ML Estimation Summary

$$P(\omega_i \mid \mathbf{x}) = P(\omega_i \mid \mathbf{x}, \boldsymbol{\theta}_i) = \frac{\boxed{p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)} P(\omega_i \mid \boldsymbol{\theta}_i)}{p(\mathbf{x} \mid \boldsymbol{\theta}_i)}$$

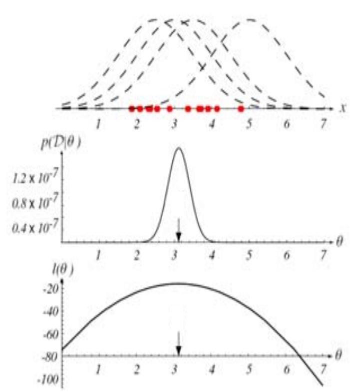
$$\boxed{p(\mathbf{x} \mid \boxed{\hat{\boldsymbol{\theta}}})}$$

maximize the likelihood: $\boxed{\arg\max_{\boldsymbol{\theta}} \boxed{p(D \mid \boldsymbol{\theta})}}$

$$\boxed{\prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})} \Rightarrow \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$: log-likelihood

Graphical View of ML Estimation

- 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 $p(D \mid \theta)$ 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 by θ_{ML}; it also maximizes the logarithm of the likelihood shown at the bottom.
- Note: even though they look similar, the likelihood $p(D \mid \theta)$ is shown as a function of θ whereas the conditional density $p(x \mid \theta)$ is shown as a function of x.



The Gaussian Case: Unknown μ

- Samples are drawn from a multivariate normal population: $p(\mathbf{x}_i \mid \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \Sigma)$
$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$
$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$
$$\boxed{\begin{aligned}&\frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^t\mathbf{x}] = \mathbf{y}\\ &\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}] = (\mathbf{M} + \mathbf{M}^t)\mathbf{x}\end{aligned}}$$
- The ML estimate for μ must satisfy:
$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{\mathrm{ML}}) = \mathbf{0}$$
- Multiplying by Σ and rearranging, we obtain: $\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$

sample mean

Just the arithmetic average of the training samples!

Unknown μ and σ

- Univariate case: $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$\ln p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

$$\nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \mu} \ln p(x_k \mid \boldsymbol{\theta}) \\ \frac{\partial}{\partial \sigma^2} \ln p(x_k \mid \boldsymbol{\theta}) \end{pmatrix}$$

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

Unknown μ and σ (cont'd)

- The ML estimate for μ and σ^2 must satisfy:

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\sigma_{\mathrm{ML}}^2}(x_k - \mu_{\mathrm{ML}}) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\sigma_{\mathrm{ML}}^2} + \sum_{k=1}^{n} \frac{(x_k - \mu_{\mathrm{ML}})^2}{(\sigma_{\mathrm{ML}}^2)^2} = 0 & (2) \end{cases}$$

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

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad ; \quad \sigma_{\mathrm{ML}}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_{\mathrm{ML}})^2$$

MLE: Multivariate Gaussians

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

The log-likelihood function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is:

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

$$= -\frac{nd}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$= -\frac{nd}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k - 2\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k) + n\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

MLE: Multivariate Gaussians (cont'd)

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{nd}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k - 2\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k) + n\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

$$= -\frac{1}{2}\left(-2\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathbf{x}_k + 2n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$$

$$= \mathbf{0}$$

$$\boxed{\begin{aligned} &\frac{\partial}{\partial \mathbf{x}}[\mathbf{y}^t\mathbf{x}] = \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{y}] = \mathbf{y} \\ &\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{M}\mathbf{x}] = (\mathbf{M} + \mathbf{M}^t)\mathbf{x} \end{aligned}}$$

Therefore, $\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}\mathbf{x}_k = n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$

This gives the ML solution,

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

MLE: Multivariate Gaussians (cont'd)

- For Σ, we need *trace* operator: $tr(\mathbf{A}) = \Sigma_{i=1\ldots n} a_{ii}$
- Trace is invariant under cyclical permutations of matrix product: $tr(\mathbf{ABC}) = tr(\mathbf{CAB}) = tr(\mathbf{BCA})$
- Derivatives of quadratic forms:
  - $\mathbf{x}^t\mathbf{A}\mathbf{x} = tr(\mathbf{x}^t\mathbf{A}\mathbf{x}) = tr(\mathbf{x}\mathbf{x}^t\mathbf{A})$ $\boxed{\because \mathbf{x}^t\mathbf{A}\mathbf{x} \text{ is a scalar}}$
  - Derivative of $tr(\mathbf{BA}) = \mathbf{B}^t$

$$\frac{\partial}{\partial a_{ij}} tr(\mathbf{BA}) = \frac{\partial}{\partial a_{ij}} \sum_k \sum_l b_{kl} a_{lk} = b_{ji}$$

  - Therefore

$$\frac{\partial}{\partial \mathbf{A}} \mathbf{x}^t\mathbf{A}\mathbf{x} = \frac{\partial}{\partial \mathbf{A}} tr(\mathbf{x}\mathbf{x}^t\mathbf{A}) = (\mathbf{x}\mathbf{x}^t)^t = \mathbf{x}\mathbf{x}^t$$

MLE: Multivariate Gaussians (cont'd)

$$l(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{nd}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \qquad \boxed{|\boldsymbol{\Sigma}|^{-1} = |\boldsymbol{\Sigma}^{-1}|}$$

$$= -\frac{nd}{2}\ln 2\pi + \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{k=1}^{n}tr[(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})]$$

$$= -\frac{nd}{2}\ln 2\pi + \frac{n}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\sum_{k=1}^{n}tr[(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}]$$

$$\frac{\partial l(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{n}{2}\boldsymbol{\Sigma} - \frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^t = \mathbf{0} \qquad \boxed{\frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = \mathbf{A}^{-t}}$$

This give the ML solution,

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

ML Analysis

- **Geometrical view**: samples are a cloud of points whose center is the sample mean (employed as ML estimator)
- **ML Bias**: ML estimator σ^2_{ML} for σ^2 is ***biased*** over a small population of samples, but it is **asymptotically unbiased** as the number of training samples becomes very large

$$E\left[\sigma^2_{\mathrm{ML}}\right] = E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu_{\mathrm{ML}})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \qquad \boxed{\text{See the notes.}}$$

- The variance over a population of n training samples is:

$$\tilde{\sigma}^2 = \frac{n}{n-1}\sigma^2_{\mathrm{ML}} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\mu_{\mathrm{ML}})^2 \qquad \boxed{\text{unbiased estimator for } \sigma^2}$$

- An elementary unbiased estimator for Σ is the ***sample covariance matrix***:

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

MLE Example Problem

- Suppose we toss a coin n times and observe the outcomes. We model the coin by a single parameter θ that represents the probability of tossing heads. Given n independent observed tosses $D = \{x_1, x_2, \ldots, x_n\}$, where $x_i = 0$ if tail and $x_i = 1$ if head, let n_H be the number of times the coin turned up heads. Then, the likelihood function is:

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

Show that the maximum likelihood estimate of θ is $\hat{\theta}_{ML} = \frac{n_H}{n}$.