Machine Learning in Robotics
Lecture 5: Density Estimation - MLE

Prof. Dongheui Lee

Institute of Automatic Control Engineering
Technische Universität München

dhlee@tum.de

Summary of last lecture : Unsupervised Learning

- Iterative Optimization
- K-means, variations of K-means, Binary Splitting, LBG algorithm

1. Initialization: Choose K random vectors $\boldsymbol{Y} = \left\{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(K)}\right\}$
2. E-step: Label each data point.

$$\boldsymbol{X}_i = \left\{\boldsymbol{x}^{(j)} \mid d(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(i)}) \leq d(\boldsymbol{x}^{(j)}, \boldsymbol{y}^{(k)}),\ j = 1, \ldots, n;\ k = 1, \ldots, K\right\}$$

3. M-step: From the current clusters, their mean vectors are updated

$$\boldsymbol{y}^{(i)} = \frac{1}{N_i} \sum_{\boldsymbol{x} \in \omega_i} \boldsymbol{x} = \mathcal{C}(\boldsymbol{X}_i)$$

4. Calculate the total distortion

$$J = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in \omega_i} (\boldsymbol{x} - \boldsymbol{y}^{(i)})^2$$

5. Evaluate the convergence. If converged, stop. Else, go to step 2.

Today Lecture Outline

- Maximum Likelihood Estimation
- Gaussian Mixture Model
- GMM Learning
- General Expectation Maximization Algorithm

Density Estimation - Motivation

- In our previous lecture on decision theory, we saw that the optimal classifier could be expressed as a family of discriminant functions

$$g_i(\boldsymbol{x}) = p(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)p(\omega_i)}{p(\boldsymbol{x})}$$

- Decision rule was

$$\text{choose } \omega_i \text{ if } g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \quad \forall j \neq i$$

- We need to estimate both prior $p(\omega_i)$ and likelihood $p(\boldsymbol{x}|\omega_i)$
- During next lectures, techniques to estimate the likelihood density function $p(\boldsymbol{x}|\omega_i)$ will be introduced

Approaches for Density Estimation

Parametric Approach

- Assumes a particular form for the density (e.g., Gaussian) and estimates the parameters of the function (e.g., mean and covariance)
- called Parameter Estimation
  - Maximum Likelihood
  - Bayesian Estimation

Non-Parametric Approach

- No functional form for the density function is assumed. The density estimate is driven entirely by the data
- called Parameter Estimation
  - Kernel Density Estimation
  - Nearest Neighbor Rule

Maximum Likelihood Estimation (MLE)

- Consider a pdf $p(\boldsymbol{x}|\theta)$ which depends on a set of parameters $\boldsymbol{\theta} = [\theta_1 \dots \theta_p]^T$
- If examples in the data set are drawn independently from $p(\boldsymbol{x}|\boldsymbol{\theta})$, the joint probability density of the entire set is given by

$$L(\boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1} p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

- We seek the set of parameters $\boldsymbol{\theta}_{ML}$ that maximize the likelihood, and call it the Maximum Likelihood estimate of parameters $\boldsymbol{\theta}_{ML} = argmax_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta})$
- It is easier to work with the logarithm of the likelihood

$$l(\boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0} \Longrightarrow \boldsymbol{\theta}_{ML}$$

Maximum Likelihood Estimation (MLE)

- Has good convergence properties as the sample size increases
- Simpler than any other alternative techniques
- Why to use the logarithm of the likelihood?
- Summary
  - We define $l(\boldsymbol{\theta})$ as the log-likelihood function $l(\boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\theta})$
  - Solve

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \left[\frac{\partial l}{\partial \theta_1}, \cdots, \frac{\partial l}{\partial \theta_p}\right]^T = \mathbf{0}$$

MLE Example 1: Gaussian with unknown mean

- The estimation of the parameters of a Gaussian distribution $p(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Find the ML estimates
  - Parameters: $\boldsymbol{\theta} = \boldsymbol{\mu}$

$$p(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{m/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu})\right]$$

$$\ln p(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}) = -\frac{1}{2}\ln((2\pi)^m|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu})$$

  - ML estimates must satisfy

$$\nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}) = \mathbf{0}$$

$$\boldsymbol{\theta}_{ML} = \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)}$$

Example 2: Univariate Gaussian with unknown μ, σ

- The estimation of the parameters of a Gaussian distribution
- Find the ML estimates for the univariate case
  - Parameters: $\theta_1 = \mu$, $\theta_2 = \sigma$
  - Log-likelihood

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

  - Gradient

$$\nabla_{\boldsymbol{\theta}} \ln p(x^{(i)}|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2^2}(x^{(i)} - \theta_1) \\ -\frac{1}{\theta_2} + \frac{(x^{(i)} - \theta_1)^2}{\theta_2^3} \end{bmatrix}$$

  - ML estimates

$$\theta_{1,ML} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \quad , \quad \theta_{2,ML}^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \hat{\mu})^2$$

theta_1: 对i从1到n累加得到, x^(i)累加为1.
总共n个theta都相等, 因此乘以系数n.

Example 3: Multivariate Gaussian with unknown μ and Σ

多变量：mu,Sigma是多维向量和矩阵.

Find the ML estimates for the multivariate Gaussian

- Parameters: $\theta_1 = \mu$, $\theta_2 = \Sigma$
- ML (Maximum-likelihood) estimates for μ and σ

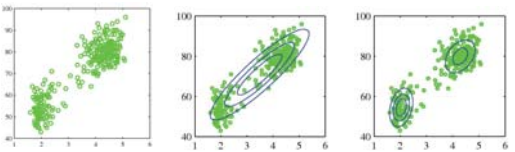
$$\theta_{1,ML} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\theta_{2,ML} = \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T$$

Mixture models

Consider the problem of modeling a pdf given a dataset of examples $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}\}$

- If the form of the underlying pdf is known (e.g. Single Gaussian distribution), the problem could be solved using the Maximum Likelihood Estimation method



Old Faithful data from Bishop2006

- Now we will consider an alternative density estimation method which is modeling the pdf with a mixture of parametric densities. In particular, we will focus on mixture models of Gaussian densities

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} p(\boldsymbol{x}|\boldsymbol{\theta}_j)p(\omega_j) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

p(omega_j) = pi_j

Gaussian Mixture Model (GMM)

- Mixture of Gaussians
  - A superposition of K Gaussian densities $p(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$
  - Parameters $\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$
  - Properties: Asymmetry, multi-modality $0 \leq \pi_j \leq 1, \quad \sum_{j=1}^{K} \pi_j = 1$
- Previously, we estimated parameters for a single Gaussian distribution by MLE
- Log-likelihood function

$$l(\boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \left[ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \right]$$

Gaussian Mixture Model (GMM)

Log-likelihood function

$$l(\boldsymbol{\theta}) = \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \left[ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \right]$$

Find the maximum of this function by differentiation
for $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$

$$\frac{\partial l}{\partial \boldsymbol{\mu}_j} = 0 \rightarrow \hat{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{n} p(\omega_j|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \boldsymbol{x}^{(i)}}{\sum_{i=1}^{n} p(\omega_j|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})}$$

$$\frac{\partial l}{\partial \sigma_j} = 0 \rightarrow \hat{\sigma_j}^2 = \frac{\sum_{i=1}^{n} p(\omega_j|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})(\boldsymbol{x^{(i)}} - \boldsymbol{\mu}_j)^2}{\sum_{i=1}^{n} p(\omega_j|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})}$$

$$\frac{\partial l}{\partial \pi_j} = 0 \rightarrow \hat{\pi}_j = \frac{\sum_{i=1}^{n} p(\omega_j|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})}{\sum_{k=1}^{K} \sum_{i=1}^{n} p(\omega_k|\boldsymbol{x}^{(i)}, \boldsymbol{\theta})}$$

Gaussian Mixture Model (GMM)

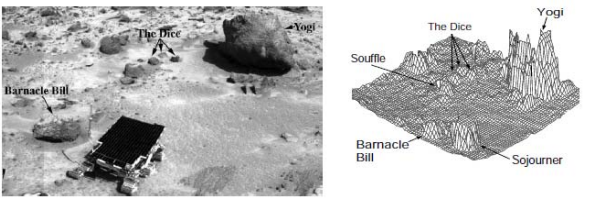
- NOT a closed form analytical solution for GMM parameters
- Due to responsibility depends on the GMM parameters
- Highly non-linear coupled system of equations

$\Rightarrow$ Iterative Numerical Optimization Technique is necessary. EM algorithm

Maximum Likelihood technique for a rover localization

漫游者

Task: To perform rover localization by matching range maps.
Motivation: For greater autonomy in Mars rovers.



(Left) Annotated image, (Right) Terrain map generated from stereo image

全景的

- Global Map: panoramic imagery generated at the center of the area from lander.
- Local Map: Occupancy map of local terrain is generated using stereo vision on Sojourner.

Olson, Clark F., and Larry H. Matthies. *Maximum likelihood rover localization by matching range maps*. IEEE International Conference on Robotics and Automation. 1998.

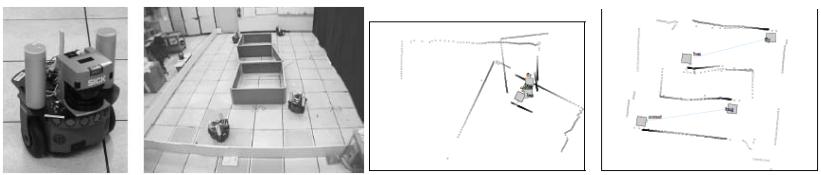
Maximum Likelihood technique for a rover localization

Task: To perform rover localization by matching range maps

- Rover position (x and y coordinates) denoted as X.
- Distances of nearby voxels (n occupied voxels) denoted as $D_1, D_2, \ldots, D_n$. 立体像素
- The position X yielding the maximum likelihood value i.e. $\ln L(X) = l(X) = \sum_{i=1}^{n} \ln p(D_i|X)$ is chosen to be the position of the rover.
- $p(D_i|X)$ is a normal distribution with a constant additive term. Normal distribution models difference occupied voxels in btw global map and local map
- Results: Comparison of rover position determined by a human operator and by the proposed localization method. Similar results.

Localization for Mobile Robot Teams

Task: to infer the relative pose of every robot in the team without the use of GPS, external landmarks or instrumentation of the environment

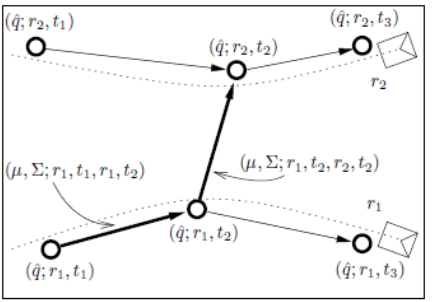


A. Howard, M. Mataric, G. Sukhatme. *Localization for Mobile Robot Teams Using Maximum Likelihood Estimation*. IROS 2002.

Localization for Mobile Robot Teams

Task: to infer the relative pose of every robot in the team

- r_1, r_2 : robots 1 and 2.
- Nodes: robot pose estimates.
- Arcs: observations (motion observation and robot observation).



Aim: Maximize $P(O|H)$ using Maximum Likelihood Estimation , O: observation set (motion sensor and robot sensor), H: robot pose estimation set.

Localization for Mobile Robot Teams

Task: to infer the relative pose of every robot in the team

- experimental snapshots
- At t=1, the relative robot pose is completely unknown
- By time t=12 sec, both robots following the outer wall have observed both robots following inner wall. The two robots on the outer wall can correctly determine each other pose even though they have never seen each other.
- errors: At t=0, the localization error is high. By the time t=20, the robot performs stable localization, The average range error is about 5.5cm.