University of Pittsburgh

ECE 2195: Special Topics – Computers Machine Learning

Parameter Estimation

Mai Abdelhakim, PhD

Assistant Professor of ECE

Swanson School of Engineering

University of Pittsburgh

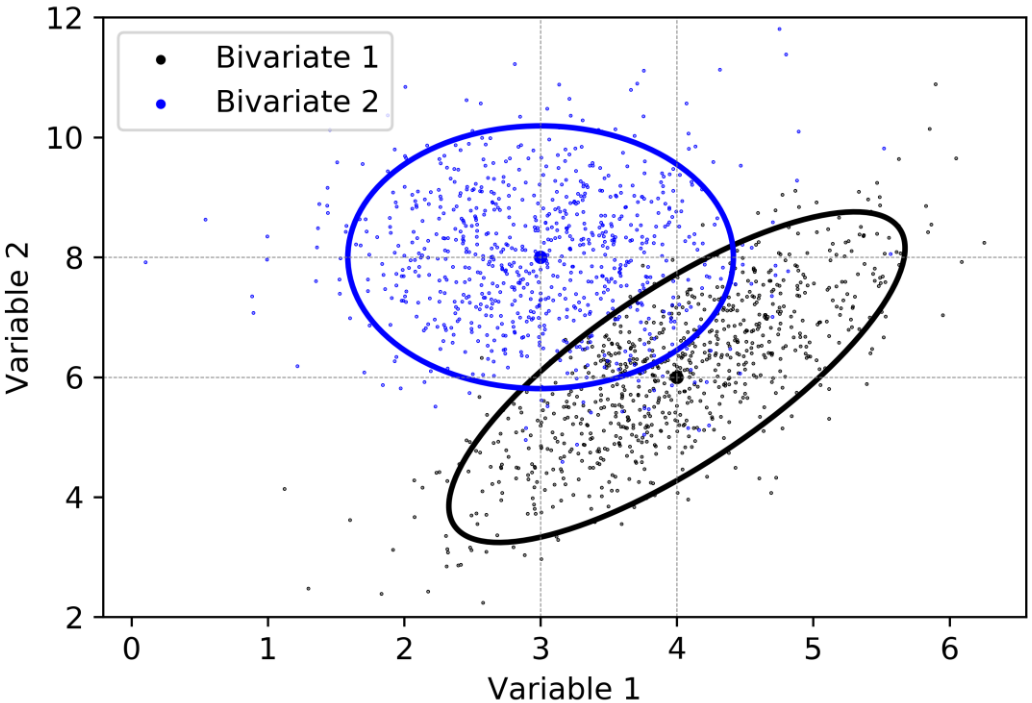
maia@pitt.edu

Recommended reading

- Parameter estimation
  - http://ciml.info/dl/v0_99/ciml-v0_99-ch09.pdf

Looking into the features, we can find distribution

- From data -> get parameters of the density function
- E.g. Multivariant normal density - > find parameters

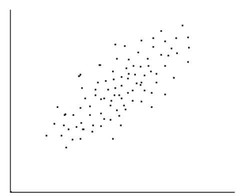


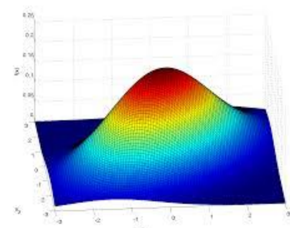
Ref: https://geostatisticslessons.com/lessons/errorellipses

Density (Parameter) Estimation

Data (X)
from a class

Estimate P(X)
in the class





- Learn the model representing the features in the data
- Given the data $D = \{D_1, D_2, \dots, D_n\}$, $\quad D_i = X_i$
- P features of data sample i, $X_i = \{x_{i1}, x_{i2}, \dots., x_{ip}\}$
- Density estimation – Learn the probability distribution P(X) from the data

Assumptions

- Data points are independent and identically distributed (iid)
  - Samples are independent of each other, and they are drawn from identical distribution
- Parametric model:
  - Model the probability distribution based on set of parameters Θ
  - Find the parameters that describe the data $\rightarrow$ Parameter estimation

Simple example of a density estimation – Tossing a biased coin example - the ML estimate

- Tossing a coin - let the P(head) = β

- Experiment conducted several times we got
  - H H T H T …… (H = head, T = tail)
  - P(D|β) = β β (1-β) β (1-β) ….. = $(\beta^{Nh}(1-\beta)^{Nt})$

- The parameter is Θ= β

  ➔ estimate $\boxed{\hat{\beta}\text{=arg max P(D| }\beta\text{) = }\frac{Nh}{Nh+Nt}}$

Maximum Likelihood Parameter Estimation

- Find the parameters Θ using the Data $D = \{D_1, D_2, \ldots, D_n\}$

- Maximum Likelihood (ML) : Find Θ that maximizes $P(D)$ given Θ

  - $\Theta_{ML} = \arg max_{\Theta}\, P(D)$

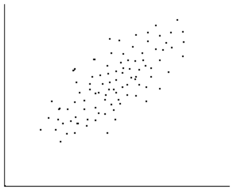
    - $P(D) = \prod_{i=1}^{n} P(D_i)$ Key assumption : data is drawn independently

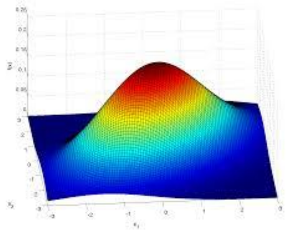
    - Log $P(D) = \sum_{i=1}^{n} \log[\,P(D_i)]$

Gaussian Distribution

- Suppose samples are drawn from Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$
- **Find ML estimate of the parameters**

Data (X) from a class → Estimate P(X) in the class





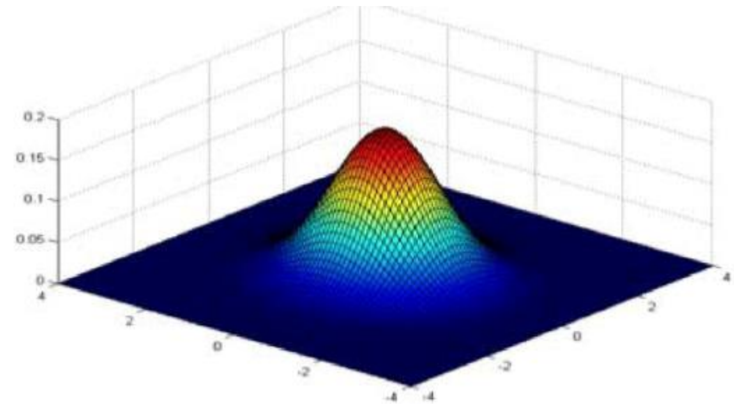
Recall - Gaussian Distribution

- 1-Dimensional Gaussian

$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- 2-Dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



- d-Dimensional Gaussian

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Gaussian Distribution

- Suppose samples are drawn from Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$
- **The ML estimate of the parameters are**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \longrightarrow$$ Biased estimate (its expected value over n is (n-1) $\boldsymbol{\sigma^2}/\boldsymbol{n}$)

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \longrightarrow$$ Unbiased estimate (its expected value over n is $\boldsymbol{\sigma^2}$)

Samples from multiple classes – each has Gaussian distribution

- Suppose K classes and each has a prior Π_k -
  - Sample i has $\boldsymbol{y_i = k}$, $\boldsymbol{k = 1, 2, 3 \dots K}$
  - Π_k =P(sample is from class k) = P($y_i = k$),
  - Features from class k are independent, and have Gaussian distribution with:

    Mean μ_k = $\{\mu_{k,1}, \mu_{k,2}, \dots.., \mu_{k,p}\}$,
    and variance of feature σ_k^2= $\{\sigma_{k,1}^2, \sigma_{k,2}^2, \dots.., \sigma_{k,p}^2\}$

Samples from multiple classes – each has Gaussian distribution

- Suppose K classes and each has a prior Π_k -
  - Sample i has $\boldsymbol{y_i = k}$, $\quad \boldsymbol{k = 1, 2, 3 \ldots K}$
  - Π_k =P(sample is from class k) = P($y_i = k$),
  - Features from class k are independent, and have Gaussian distribution with:

  Mean $\mu_k = \{\mu_{k,1}, \mu_{k,2}, \ldots, \mu_{k,p}\}$,
  and variance of feature $\sigma_k^2 = \{\sigma_{k,1}^2, \sigma_{k,2}^2, \ldots, \sigma_{k,p}^2\}$

$$p(D) = \overbrace{\prod_i \underbrace{\Pi_{yi}}_{\text{choose label}} \underbrace{\prod_d \underbrace{\frac{1}{\sqrt{2\pi\sigma_{y_i,d}^2}} \exp\left[-\frac{1}{2\,\sigma_{y_i,d}^2}(x_{i,d} - \mu_{y_i,d})^2\right]}_{\text{choose feature value}}}_{\text{for each feature}}}^{\text{for each example}}$$

$$\sigma_{k,d}^2 = \sigma_{y_i,d}^2$$

Formulating using Lagrange to minimize objective under constraint

- $J(\Theta)$ = - ln P(D) + $\underbrace{\lambda(\sum_{j=1}^{K} \Pi_j - 1)}_{\text{Constraint that sum probability is to 1}}$, λ = Lagrange multiplier

- $\frac{\partial J(\Theta)}{\partial \Pi_k} = 0$, $\frac{\partial J(\Theta)}{\partial \mu_{k,f}} = 0$, $\frac{\partial J(\Theta)}{\partial \sigma_{k,f}^2} = 0$

ML estimates of parameters of classes can be obtained from data

- Taking the $\log_e$ and the derivative w.r.t (with respect to) each of the parameters (priors, means, variances)

- We get
  - $\Pi_k = \frac{n_k}{\sum_{k=1}^{K} n_k} = \frac{n_k}{n}$

    The total no. of samples: n= $\sum_{k=1}^{K} n_k$
    - n_k is the number of samples from class k
  - $\mu_{k,f} = \frac{\sum_{i:y_i=k} x_{i,f}}{n_k}$, mean of class k feature f
  - $\sigma_{k,f}^2 = \frac{\sum_{i:y_i=k}(x_{i,f} - \mu_{k,d})^2}{n_k}$, variance of class k feature f