Module 1.

Bayes Decision Theory

$P(w_i|x) = \frac{P(x|w_i)P(w_i)}{P(x)}$

$g(w_i|x) \rightarrow$ omit $P(x)$

Ref. "chapter 2 of Duda"

Module 2

Parameter Estimation Methods

→ Discriminant function

→ Decision Surface

→ Gaussian for 1D & 2D (NDF)

→ Discriminant for NDF

1D $P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

20

$P(X) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)}$

Σ = covariance matrix.

✓ Supervised scenario → PARAMETER ESTIMATION: (μ, Σ) !!!

$$P(\omega_i | x) = \frac{P(\omega_i)\, P(x|\omega_i)}{P(x)}$$

$P(\omega_i)$? → estimate

↳ $P(x|\omega_i)$ → estimate → Gaussian Distrib^n in multivariable

↳ I don't know what is $P(x|\omega_i) = N(\mu, \Sigma)$

$$P(x|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

estimate what the $\hat{\mu}, \hat{\Sigma}$ (parameters) this distribution follows.

Parameter Estimation

- optimal classifier -> if we knew the prior $P(\omega i)$ and the class-conditional densities $p(x|\omega i)$.

- But is it always possible to have this information?

- What we have is a number of design samples or training data — particular representatives of the patterns we want to classify.

Parameter Estimation

- Can we use training data to design the classifier?

- Consider Supervised Learning scenario (labelled data, class labels are available)

Where the data samples and the class to which they belong is known. $p(x|\omega_i)$?

Then we can use the samples to estimate the unknown probabilities and probability densities. $P(x|\omega_i)$?

Parameter Estimation

- We assume that **p(x|ωi)** is a normal density with mean **μi** and covariance matrix **Σi**,

$$p(x|\omega_i) = N(\mu_i, \Sigma_i)$$

- Instead of estimating an unknown function **p(x|ωi)** -> now we have to estimating the parameters **μi** and **Σi**

Maximum Likelihood Estimation

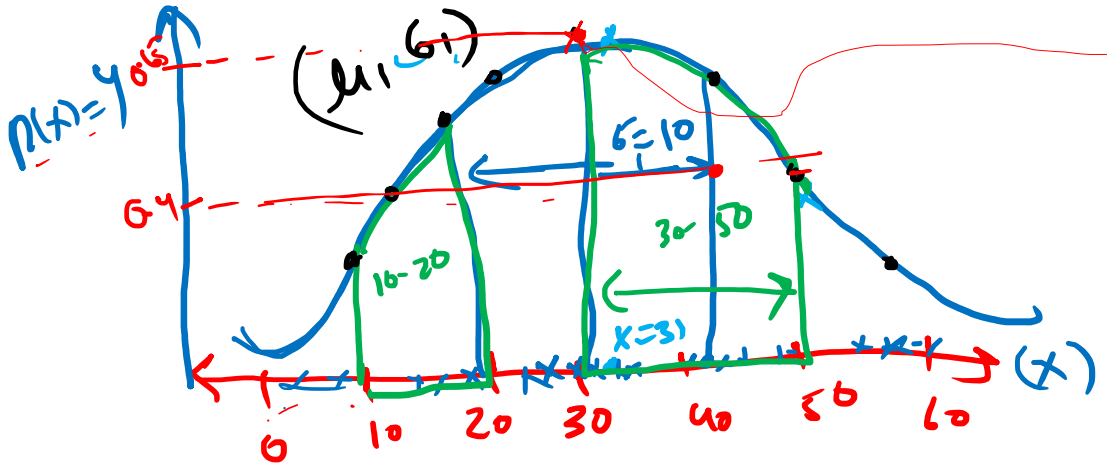
Probability

- Likelihood vs Probability

1D. NDF

$[\mu_1 = 30]\ [\sigma_1 = 10]$

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-30)^2}{10^2}}$$



$\mu_1 = 30$

$P(X=31) = 0.65$

$P(X=50) = 0.40$

$P(X = 30 \text{ to } 50 \mid \mu_1 = 30, \sigma_1 = 10)$

$P(X = 10 \text{ to } 20 \mid \mu_1 = 30, \sigma_1 = 10)$

prop.

(μ, Σ)

$\mu = 30, \sigma = 10$

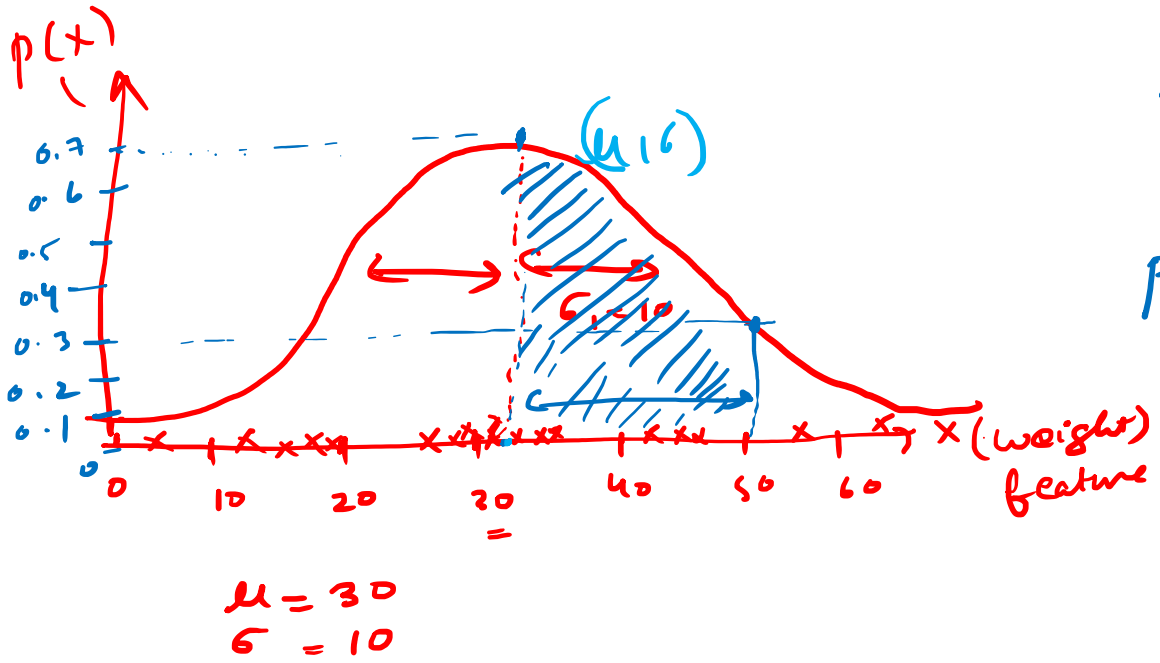
Maximum Likelihood Estimation

- Likelihood vs Probability

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\mu = 30, \sigma = 10$

Example of fishes



$P(X=30) = 0.7$ (given that $\mu=30, \sigma=10$)

$P(X=50) = 0.3$ ($\mu=30, \sigma=10$)

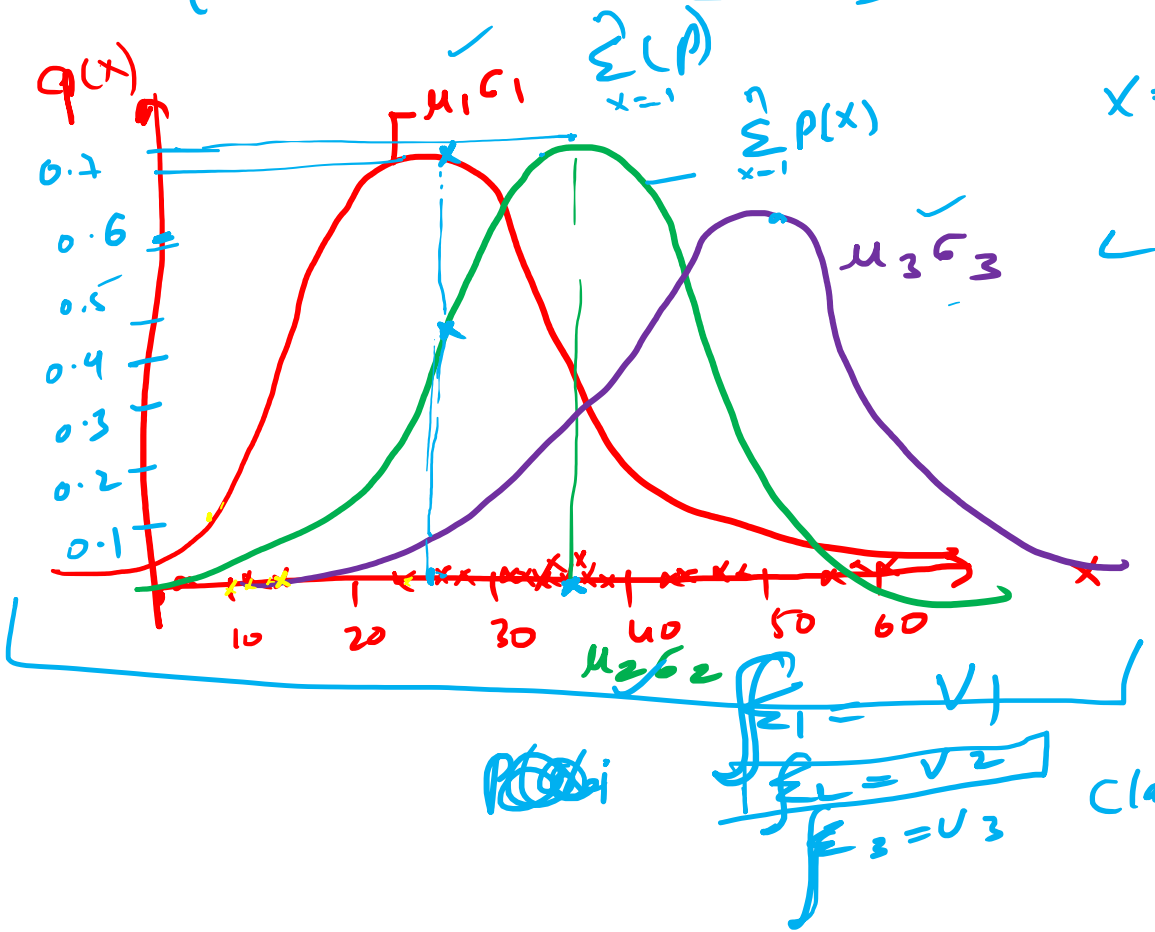
$P(X=30 \text{ to } 50 \mid \mu=30, \sigma=10)$

What is $P(x)$. given μ, σ.

Max. Likelihood Estimation.

Likelihood $(\mu_i \Sigma_i)$

$\sum_{i=1}^{?} p(\mu_j \Sigma_j | x_i)$ (D1) →



I only have sampler points (X) → Train data.

I don't know what is the $\mu_i \Sigma_i$?

$x = 25$

$p(x=25)$?

$p(\mu_1 \Sigma_1 | x=25) = 0.7$ (✓)

$p(\mu_2 \Sigma_2 | x=25) = 0.4$ → ✓

$p(\mu_3 \Sigma_3 | x=25) = 0.3$

$p(x=35)$

$p(\mu_1 \Sigma_1 | 35) = 0.6$

$p(\mu_2 \Sigma_2 | 35) = 0.75$ → ✓ $(\mu_2 \Sigma_2)$

$p(\mu_3 \Sigma_3 | 35) = 0.5$

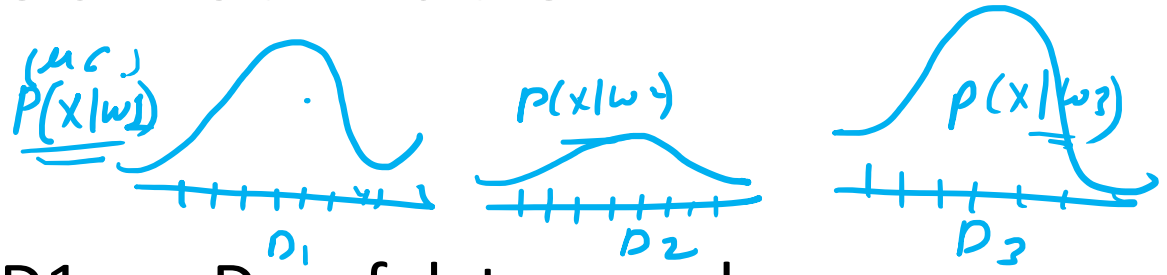
$\Sigma_1 = V_1$

$\Sigma_2 = V_2$

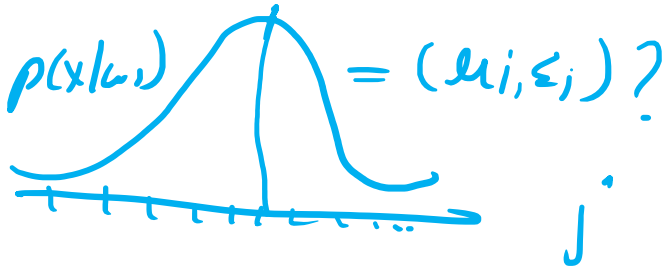
$\Sigma_3 = V_3$

Class ω_j

Maximum Likelihood Estimation



- Let say we have 'c' classes.
- We also have have 'c' sets, D1, ..., Dc, of data samples
- Samples in **Dj** having been drawn independently according to i.i.d. the probability law **p(x|ωj).**
- The distribution (x|ωj) has a known parametric form -> determined uniquely by parameter vector **θj.**
- **example, we might have p(x|ωj) ~ N(μj,Σj), where θj consists of the components of μj and Σj.**



Maximum Likelihood Estimation

- Objective : information provided by the training samples (D1, ..., Dc,) to estimate unknown parameter vectors **θ1, ..., θc.**
- samples in Di give no information about θj , if i != j
- that is, we shall assume that the parameters for the different classes are functionally independent.
- This permits us to work with each class separately, and to simplify our notation by deleting indications of class distinctions.
- **p(x|ω, θ) -> p(x|θ)**

Maximum Likelihood Estimation

- Use a set D of training samples drawn independently from the probability density p(x|θ) to estimate unknown parameter vector θ.

- Suppose that D contains n number of samples: x1, ...,xn.
- Since the samples were drawn independently, we have

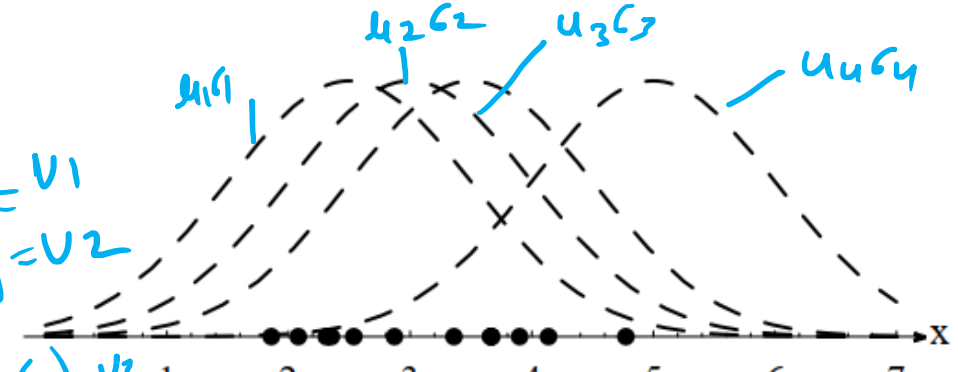
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$$

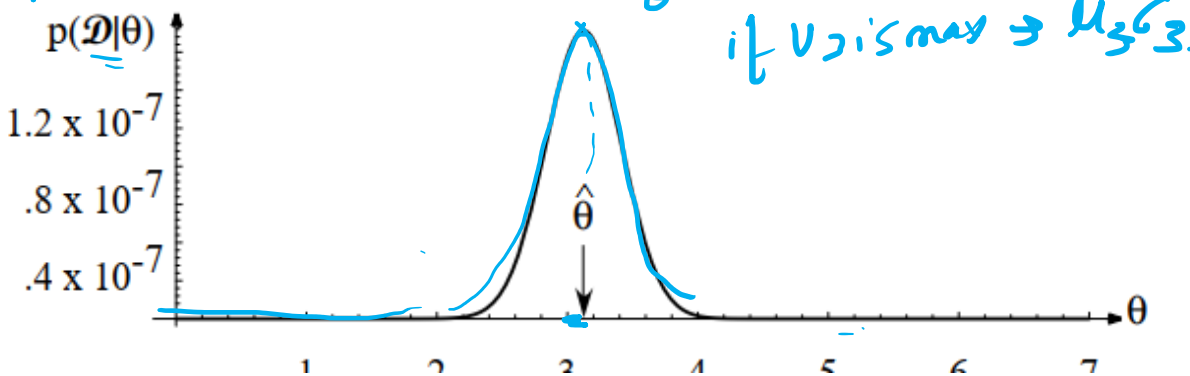
Maximum Likelihood Estimation

- **p(D|θ)** is called the likelihood of **θ** with respect to the set of samples\

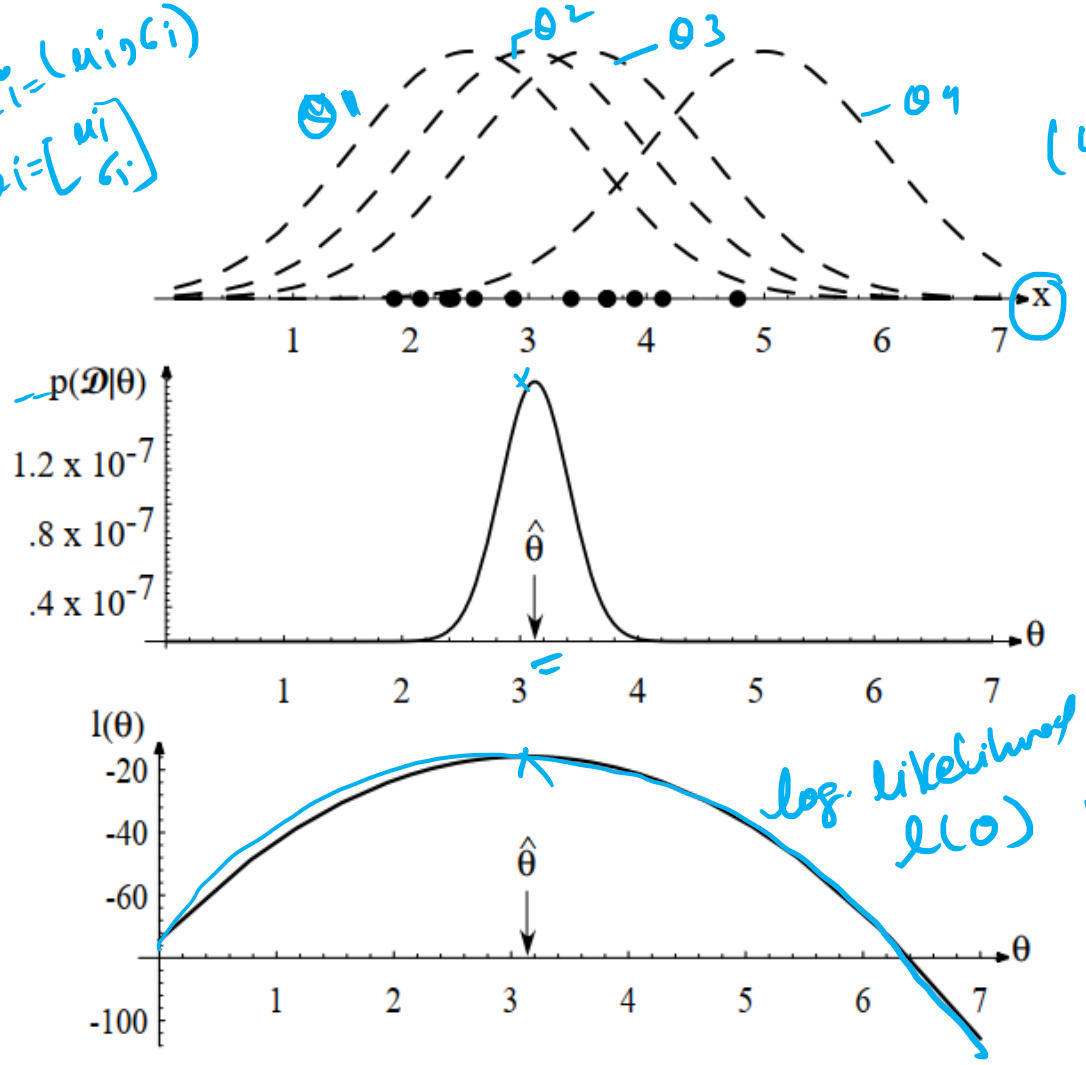
- The maximum likelihood estimate of θ is, value $\hat{\theta}$ that maximizes p(D|θ).

- Intuitively, this estimate corresponds to the value of θ that in some sense best agrees with or supports the actually Observed training samples





Maximum Likelihood Estimation



$P(D|\theta)$ = is max for a particular $\hat{\theta}$

$P(D|\theta) = \prod_{k=1}^{n} P(x_k|\theta_i) =$ max. likelihood.

$\log P(D|\theta)$ = log of max. likelihood

$\log p(D|\theta)$.

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 figures 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 — i.e., the log-likelihood $l(\theta)$, shown at the bottom. Note especially that the likelihood lies in a different space from $p(x|\hat{\theta})$, and the two can have different functional forms.

Maximum Likelihood Estimation

It is usually easier to work with the logarithm of the likelihood than with the likelihood itself

- Lets define l(θ) as the log-likelihood function
-

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

since

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

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

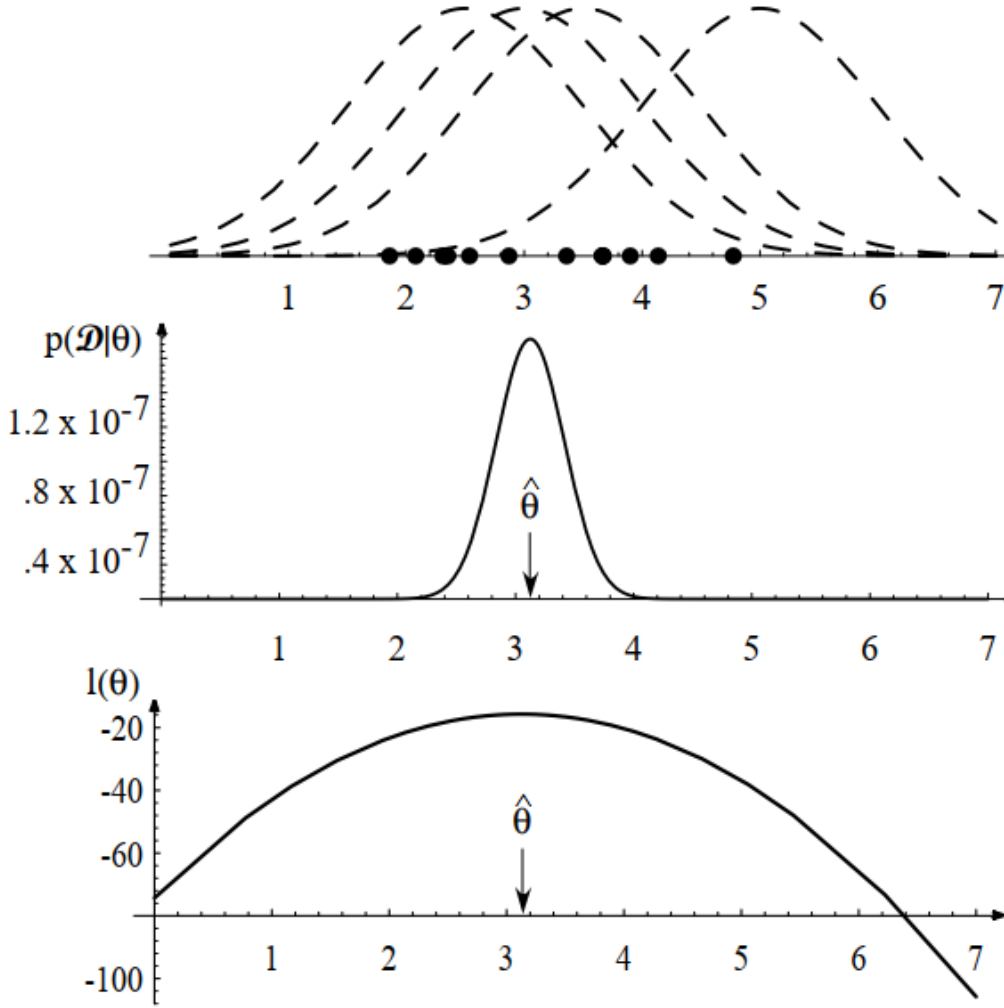
Maximum Likelihood Estimation

we don't know init what is θ!

- We can then write our solution formally as the argument θ that maximizes the loglikelihood, i.e.,

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

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



Maximum Likelihood Estimation

- We can then write our solution formally as the argument θ that maximizes the loglikelihood, i.e.,

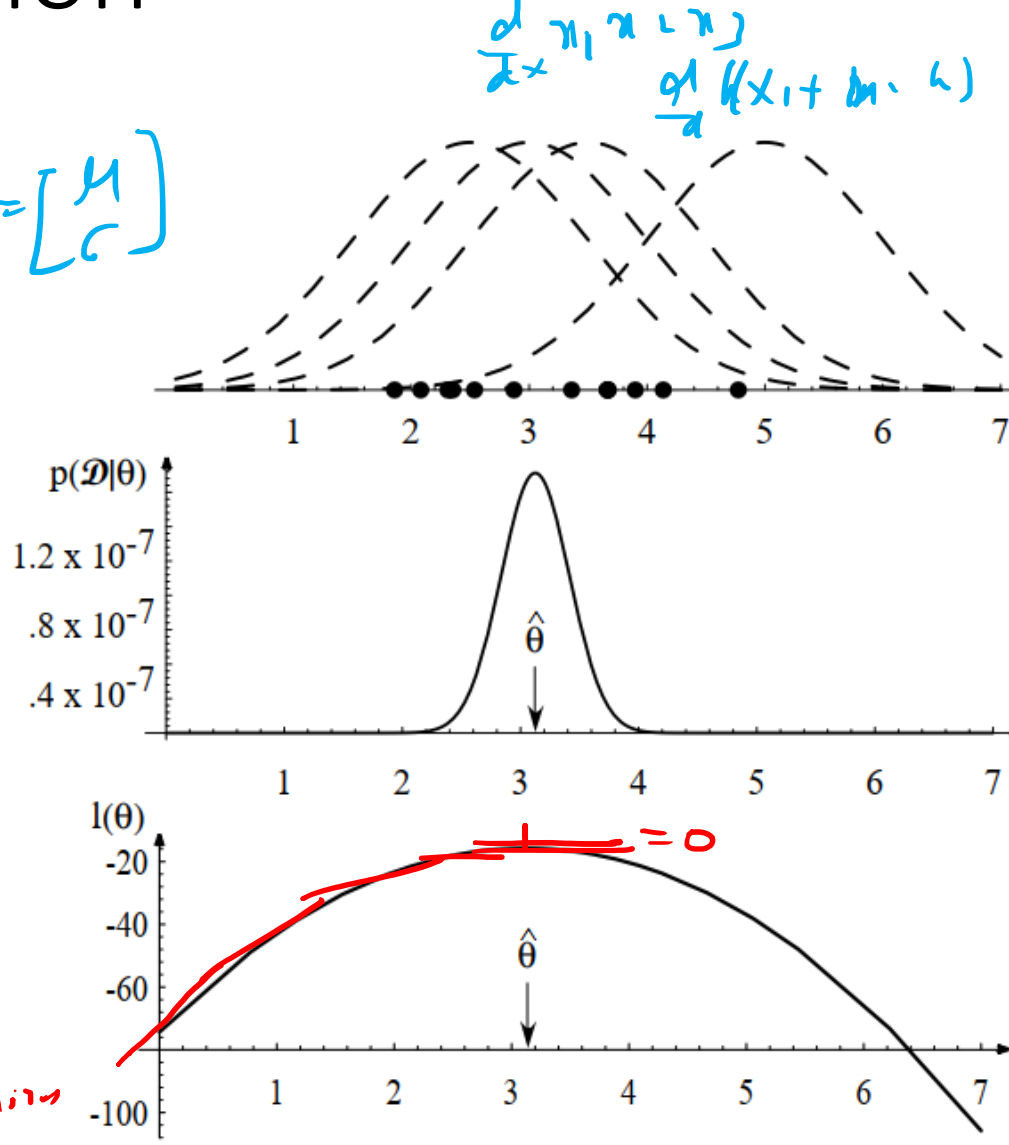
$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})$$

Take first derivative:

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

Equate to zero to get a maxima:

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



Maximum Likelihood Estimation

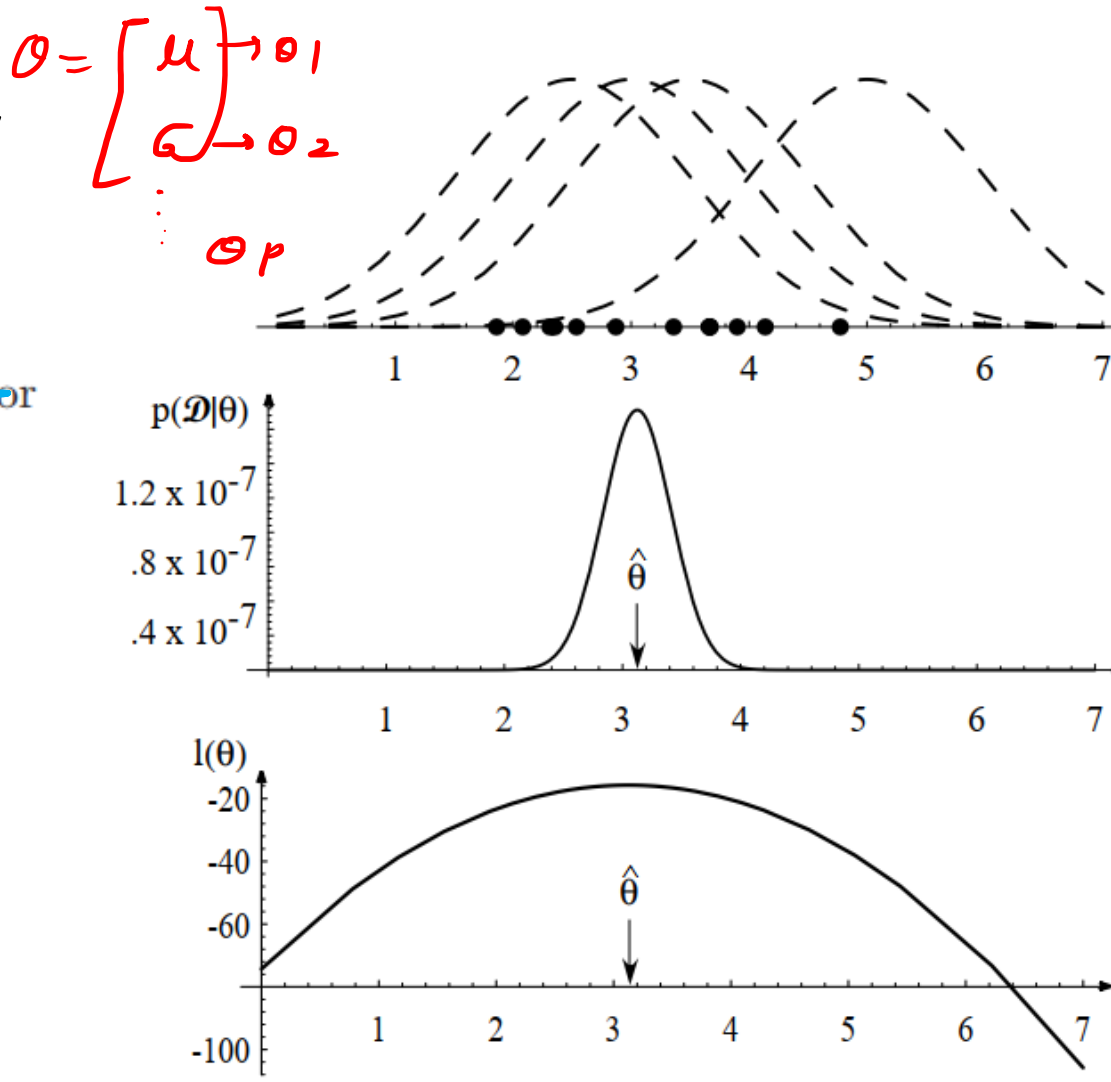
- We can then write our solution formally as the argument θ that maximizes the loglikelihood, i.e.,

p-component vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$, and $\nabla_{\boldsymbol{\theta}}$ be the gradient operator

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

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

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



The Gaussian Case: Unknown μ

- For simplicity, consider first the case where only the mean is unknown.

$\theta = \begin{bmatrix} \theta_1 = \mu \\ \vdots \\ \theta_r \end{bmatrix}$

"θ_1"

The Gaussian Case: Unknown μ

- For simplicity, consider first the case where only the mean is unknown.
- Under this condition, we consider a sample point x_k and find

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

The Gaussian Case: Unknown μ

- For simplicity, consider first the case where only the mean is unknown.
- Under this condition, we consider a sample point x_k and find

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

- and

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

The Gaussian Case: Unknown μ

- that the maximum likelihood estimate for μ must satisfy

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

Equating the derivative to zero

- Multiplying by Σ and rearranging, we obtain

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

= (1) proved over here with the help of the equation.

This is a very satisfying result. It says that the maximum likelihood estimate for the unknown population mean is just the arithmetic average of the training samples — the *sample mean*

The Gaussian Case: Unknown μ and Σ

- Here neither the mean **μ** nor the covariance matrix **Σ** is known.
- Thus, these unknown parameters constitute the components of the parameter vector **θ**.
- Consider first the univariate case with θ1 = μ and θ2 = σ2.
- Here the log-likelihood of a single point is

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

The Gaussian Case: Unknown μ and Σ

- Here neither the mean **μ** nor the covariance matrix **Σ** is known.
- Thus, these unknown parameters constitute the components of the parameter vector **θ**.
- Consider first the univariate case with θ1 = μ and θ2 = σ2.
- Here 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

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

The Gaussian Case: Unknown μ and Σ

follow book.

- Taking derivative and equating to zero we have:

we have proved by taking MLE & equating derivative to zero.

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \quad \text{//w.r.t } \theta 1$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0, \quad \text{//w.r.t } \theta 2$$

Result of MLE.

We get

sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\theta}_1$$

sample variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2. \qquad \hat{\theta}_2$$

sample mean → mean of the given population ($\hat{\mu}$)

& sample var → variance of the given population ($\hat{\sigma}^2$)

The Gaussian Case: Unknown μ and Σ

Today's class

- the maximum likelihood estimates for μ and Σ are given by

(1) Prob (2) Likelihood
(3) log likelihood.

Read your ebook (MLE)

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

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

Max likelihood $P(D|\theta)$

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

$\theta = (\theta_1, \theta_2)$

$\theta_1 = \mu_1 =$ sample mean

$\theta_2 = \sigma_1^2 =$ sample variance

Thus, once again we find that the maximum likelihood estimate for the mean vector is the sample mean. The maximum likelihood estimate for the covariance matrix is the arithmetic average of the n matrices $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$. Since the true covariance matrix is the expected value of the matrix $(\mathbf{x} - \hat{\boldsymbol{\mu}})\ (\mathbf{x} - \hat{\boldsymbol{\mu}})^t$, this is also a very satisfying result.

Bayesian estimation

- Maximum likelihood method considers the true parameter vector, θ, to be a fixed (constant) value,
- Bayesian learning considers θ to be a random variable,

- We don’t know the exact value of parameter, but we know that the parameter falls in a certain range of values.

- We capture our lack of knowledge about the value of parameter -> through a probability density over the parameter space.

Bayesian estimation

- We have a **prior density p(θ) the parameter** (prior meaning without having any look at the training data).
- prior density p(θ) captures information about the parameters.

- We can then have access to our data **D** (the training samples).

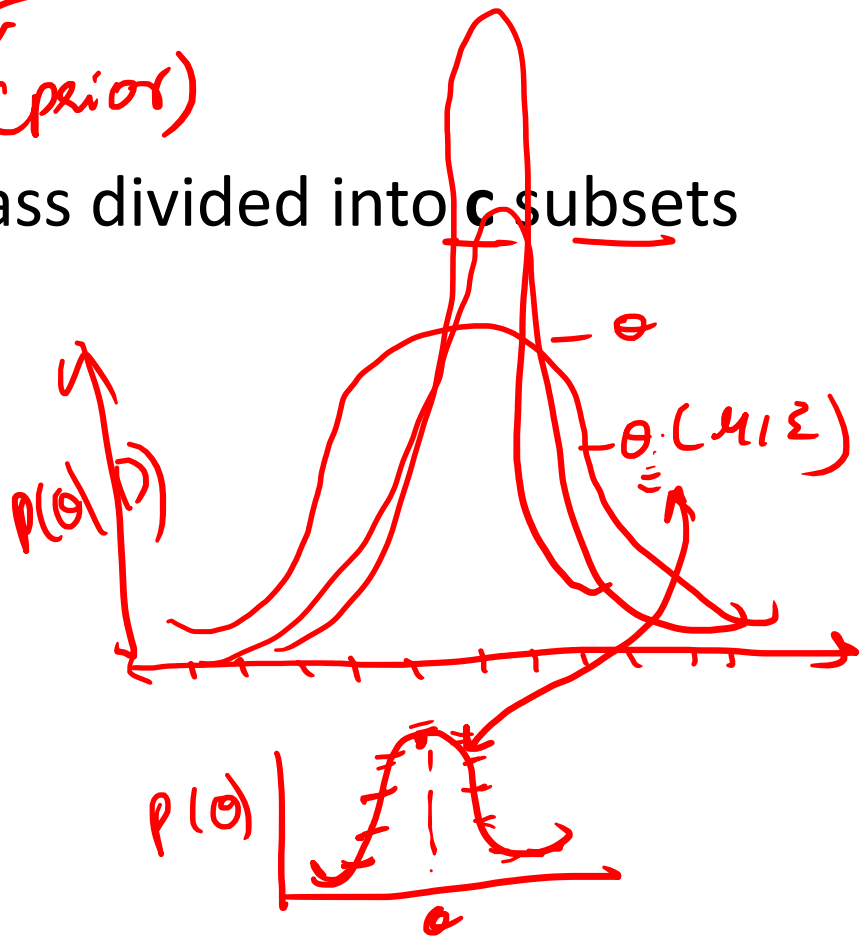
- We then see how our data **D** is useful in mapping from **prior density p(θ) into posterior density p(θ|D) for the parameter θ.**

Notations

- The parameter is denoted by **θ**
- The prior density of parameter -> **p(θ)**
- We have the training samples for ever class divided into **c** subsets
  - D1, …, Dc,
  - with the samples in D_i belonging to class ω_i
- We have posterior density **p(θ|D)**

Notations

- The parameter is denoted by **θ**
- The prior density of parameter -> **p(θ)**
- We have the training samples for ever class divided into **c** subsets
  - D1, …, Dc,
  - with the samples in D_i belonging to class ω_i
- We have posterior density **p(θ|D)**



Recap Bayesian Rule

- Let D denote the set of samples, then we can emphasize the role of the samples by saying that our
- goal is to compute the posterior probabilities P(ωi|x, D)

$$P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})}.$$

- Given ‘c’ classes we have c separate parameter estimation problems p(x|D) corresponding to each class ωi.

Recap Bayesian Rule

- Let D denote the set of samples, then we can emphasize the role of the samples by saying that our
- goal is to compute the posterior probabilities P(ωi|x, D)

$$P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})}.$$

- Given 'c' classes we have c separate parameter estimation problems p(x|D) corresponding to each class ωi.

The Parameter Distribution

- Although desired pdf p(x) is unknown, we assume that it has a known parametric form.
- Lets say parametric form is a gaussian distribution.
- The only thing assumed unknown is the value of a parameter vector θ.

The Parameter Distribution

- our basic goal is to compute p(x|D)
- We do this by integrating the joint density **p(x, θ|D)** over **θ**.
- That is

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta},$$

- we can write **p(x, θ|D**) as the product **p(x|θ, D). p(θ|D)**

- Since the selection of x and that of the training samples in D is done independently, the first factor **p(x|θ, D)** is merely **p(x|θ)**

The Parameter Distribution

- That is, the distribution of x is known completely once we know the
- value of the parameter vector **θ**

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta},$$

$$\boxed{p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}.}$$

prior θ - $p(\theta)$

Data seen D -

$p(\theta|D)$ = posterior prob.

- This key equation links the desired class-conditional density **p(x|D)** to the **posterior density p(θ|D)** for the unknown parameter vector **θ**

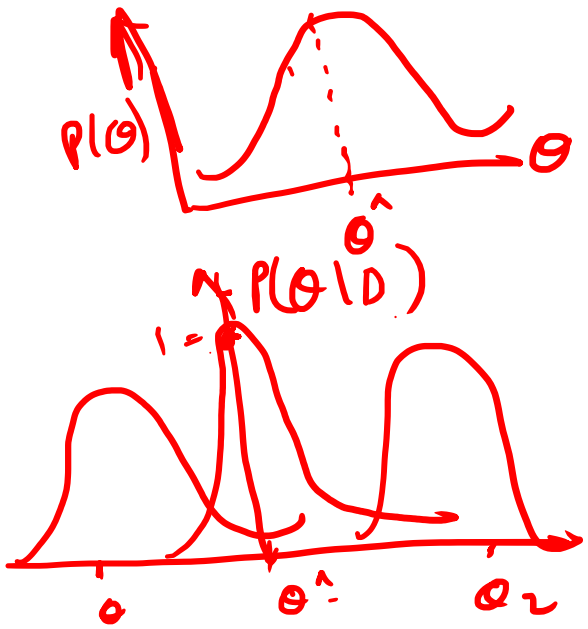
The Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}.$$

- Lets say we have a given range of parameter θ .
- If p(θ|D) peaks very sharply at some value θˆ, we obtain p(x|D)=p(x|θˆ),
- i.e., the result we would obtain by substituting the estimate θˆ for the true parameter vector.
- when the unknown densities have a known parametric form, the samples exert their influence on **p(x|D)** through the posterior density **p(θ|D).**

The Parameter Distribution

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}.$$



- Lets say we have a given range of parameter θ .
- If p(θ|D) peaks very sharply at some value θˆ, we obtain p(x|D)≅p(x|θˆ),
- i.e., the result we would obtain by substituting the estimate θˆ for the true parameter vector.
- when the unknown densities have a known parametric form, the samples exert their influence on **p(x|D)** through the posterior density **p(θ|D).**

Bayesian Parameter Estimation: Gaussian Case

3.2

- Bayesian estimation techniques to calculate the a **posteriori density $p(\theta|D)$** and the desired **probability density $p(x|D)$** for the case where
- $p(x|\mu) \sim N(\mu,\Sigma)$.

$p(x|\theta) = p(x|\mu) \sim N(\mu, \Sigma)$

The Univariate Case: p(μ|D)

- the case where μ is the only unknown parameter.

$$p(x|\mu) \sim N(\mu, \sigma^2),$$

- where the only unknown quantity is the mean μ.
- We assume that whatever prior knowledge we might have about μ can be expressed by a known **prior density p(μ).**

$$p(\mu) \sim N(\mu_0, \sigma_0^2),$$

- where both μ_0 and σ_0^2 are known
- And μ_0 represents our best a priori guess for μ, and σ_0^2 measures our uncertainty about this guess.

The Univariate Case: p(μ|D)

- the case where μ is the only unknown parameter.

$$p(x|\mu) \sim N(\mu, \sigma^2),$$

- where the only unknown quantity is the mean μ.
- We assume that whatever prior knowledge we might have about μ can be expressed by a known **prior density p(μ).**

$$p(\mu) \sim N(\mu_0, \sigma_0^2),$$

- where both μ_0 and σ_0^2 are known
- And μ_0 represents our best a priori guess for μ, and σ_0^2 measures our uncertainty about this guess.

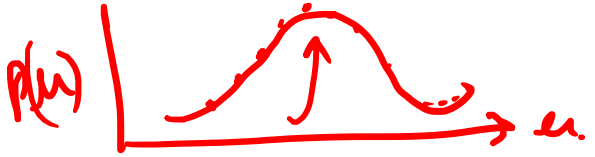
The Univariate Case: p(μ|D)

- Having selected the a priori density for μ, we can view the situation as follows.

- Imagine that a value is drawn for μ from a population governed by the probability law p(μ).

The Univariate Case: p(μ|D)

- Having selected the a priori density for μ, we can view the situation as follows.



- Imagine that a value is drawn for μ from a population governed by the probability law p(μ).
- Once this value is drawn, it becomes the true value of μ and completely determines the density for x.
- Suppose now that n samples $x_1, \ldots, x_n$ are independently drawn from the resulting population.

The Univariate Case: p(μ|D)

- Letting D = {x_1, ..., x_n} and we use Bayes' formula

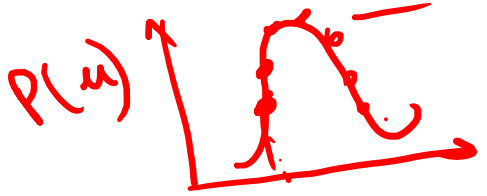
$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\ d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

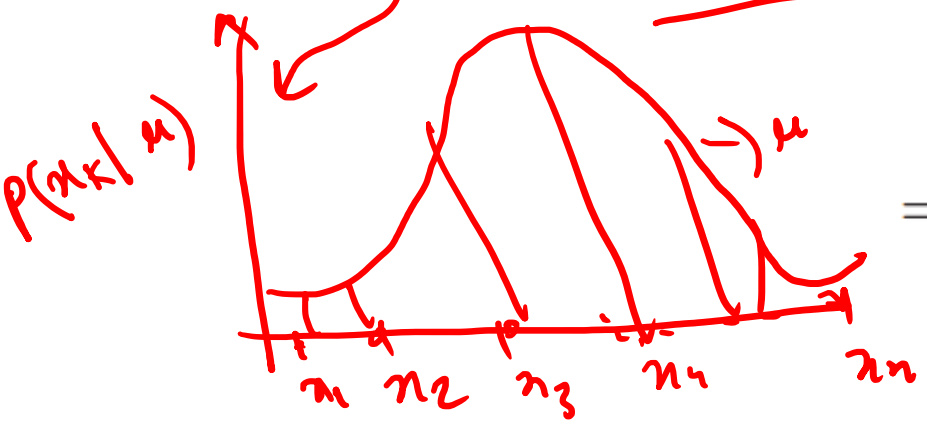
- where α is a normalization factor that depends on D but is independent of μ
- it relates the prior density $p(\mu)$ to an a posteriori density $p(\mu|D)$

The Univariate Case: p(μ|D)

- Letting D = {x_1, ..., x_n} and we use Bayes' formula



$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\, d\mu}$$



$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

- where α is a normalization factor that depends on D but is independent of μ
- it relates the prior density $p(\mu)$ to an a posteriori density $p(\mu|D)$

The Univariate Case: p(μ|D)

- Since

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

- And $p(x_k|\mu) \sim N(\mu, \sigma^2)$ and $p(\mu) \sim N(\mu_0, \sigma_0^2)$, we have

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

The Univariate Case: p(μ|D)

- Since $$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu),$$

- And $p(x_k|\mu) \sim N(\mu, \sigma^2)$ and $p(\mu) \sim N(\mu_0, \sigma_0^2)$, we have

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$p(\mu|D) = \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

The Univariate Case: p(μ|D)

- On simplifications

$$p(\mu|\mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

- Thus, p(μ|D) is an exponential function of a quadratic function of μ, i.e., is again a normal density.
- Since this is true for any number of training samples,
- p(μ|D) remains normal as the number n of samples is increased.

- p(μ|D) is said to be a **reproducing density**

The Univariate Case: p(μ|D)

- If we write $p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$, then μ_n and σ_n^2 can be found by equating coefficients

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

$$p(\mu|\mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

The Univariate Case: p(μ|D)

- If we write $p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$, then μ_n and σ_n^2 can be found by equating coefficients

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

$$p(\mu|\mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

- Which gives

- $$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$ and $$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma_2}\bar{x}_n + \frac{\mu_0}{\sigma_0^2},$$ and $$\bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k.$$

where $\bar{x}_n$ is the sample mean

The Univariate Case: p(μ|D)

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

We solve explicitly for μ_n and σ_n^2 and obtain

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

sample

estimated μ0 of p(μ)

and

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}.$$

p(μ)

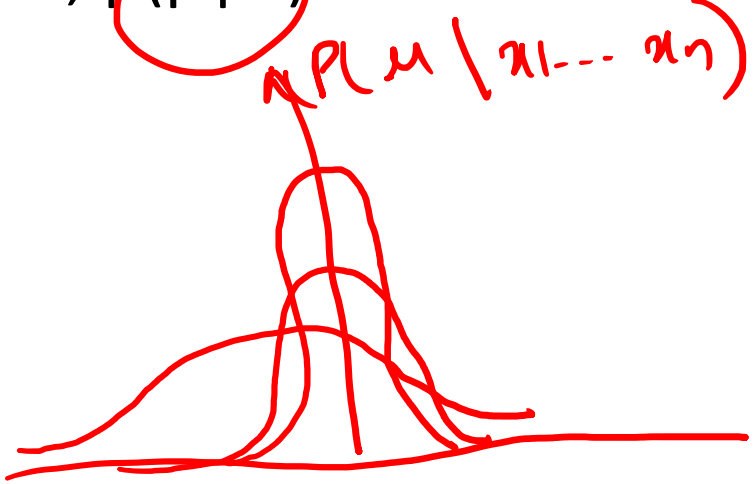
- These equations show how the prior information is combined with the empirical information in the samples to obtain the a posteriori density p(μ|D)

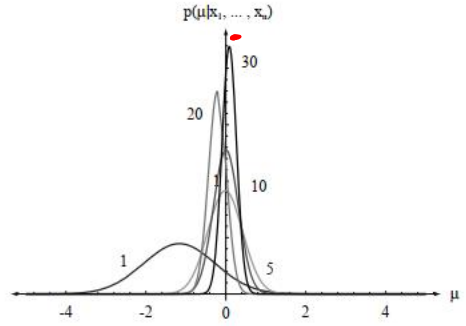
p(μ|D)

D

The Univariate Case: p(μ|D)

- Roughly speaking, μ_n represents our best guess for μ after observing n samples, and σ_n^2 measures our uncertainty about this guess.
- Since σ_n^2 decreases monotonically with n —— each additional observation decreases our uncertainty about the true value of μ.
- As n increases, p(μ|D) becomes more and more sharply peaked,





Computing p(x|D)

- p(μ | D) is computed
- p(x | D) remains

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n), \qquad (37)
\end{aligned}
$$

$$
p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2).
$$

(Desired class-conditional density p(x | Dj, ωj))

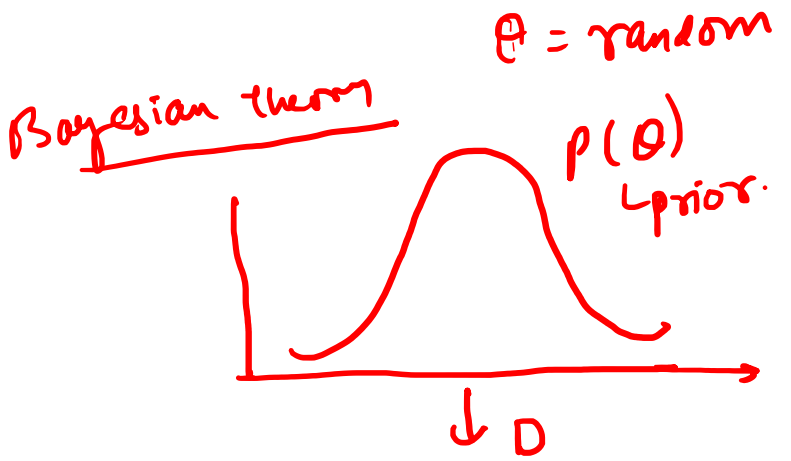
The Multivariate Case

- Univariate case : $p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$.

- Multivariate case: $p(\mathbf{x}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$

The Multivariate Case

- Univariate case : $p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2).$
- Multivariate case: $p(\mathbf{x}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$



$p(\theta|\mathcal{D})$ = posterior dens.

Please go & read chapter 3!!!

Bayesian Parameter Estimation: General Theory

- The form of the density p(x|θ) is assumed to be known, but the value of the parameter vector θ is not known exactly.
- Our initial knowledge about θ is assumed to be contained in a known a priori density p(θ).
- •The rest of our knowledge about θ is contained in a set D of n samples x1, ...,xn drawn independently according to the unknown probability density p(x).

Bayesian Parameter Estimation: General Theory

- Compute posterior density p(θ | D)

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}},$$

By independence assumption:

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

- then p(x |D) using

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}.$$

Bayesian Parameter Estimation: General Theory

- Compute posterior density p(θ | D)

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}},$$

By independence assumption:

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

- then p(x |D) using

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}.$$