Software Engineering Workshop Block 4: Review of Stochastic Models and Fitting Gaussian Mixture Models

June 2025

Jungsik Noh

UTSouthwestern
Medical Center

Outline

1. Review for Stochastic Models
   - Random variables, PDF, likelihood function, etc
   - Goal: to have clear understanding of Autoregressive-Hidden Markov Models

2. Expectation-Maximization algorithm (EM)
   for Gaussian Mixture Models (GMM)

A simple example of a stochastic model

- Experiment
  - Suppose we inject two types (wild-type/mutant) of cancer cells into mice and measure tumor sizes after 4 weeks.
- Variables
  - x_i: an observed tumor size of the i-th mouse in the WT group (i = 1, 2, ..., 10)
  - y_j: an observed tumor size of the j-th mouse in the MT group (j = 1, 2, ..., 12)
- The t-test can determine whether the two group means of tumor sizes are the same or not.

- $\{x_1, \ldots, x_{10}\}$ is an "*Independent and Identically Distributed*" (***IID***) data.

- Let X_i denote a tumor size value of the i-th WT mouse that will be observed after experiment.
- [Def] A **random variable** is a quantity that can have different numerical values depending on the outcome of a random experiment. Or mathematically, it is a function from a set of possible outcomes to a measurable space like $\mathbb{R}^d$.
- [Stochastic Model] $X_1, \cdots, X_{10} \sim iid\ N(\mu_{WT}, \sigma_{WT}^2), Y_1, \cdots, Y_{12} \sim iid\ N(\mu_{MT}, \sigma_{MT}^2)$
- T-test for $H_0: \mu_{WT} = \mu_{MT}$

[Notation] X is a random variable, and x is a real number.

Probability and a probability density function (pdf)

- A **sample space** (S) is the set of all possible outcomes of a random experiment.
- A subset of a sample space is called an **event**.
  - Tossing a coin twice
  - S={HH, HT, TH, TT}
  - At least one head is observed = E = {HH, HT, TH}
  - Pr({HH, HT, TH}) = 0.75 → We assign a **probability, $\mathbf{Pr}(\cdot)$**, to an event.
  - 0 ≤ Pr(A) ≤ 1, for any $A \subset S$.
  - Pr(S) = 1. For example, Pr(“the first toss is head” or “the first toss is tail”) = $\Pr(A \cup A^c)$ = 1.

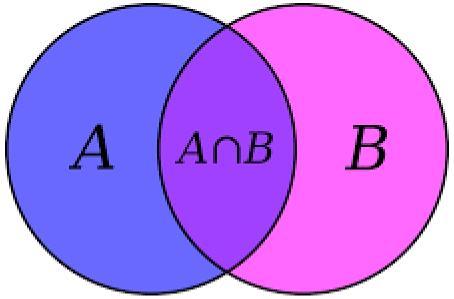
- Suppose X is a real-valued random variable. Then, “X belongs to a subset of $\mathbb{R}$” is an event.
  - Let X be the number of heads when tossing a coin twice.
  - (X ≥ 1) = “at least one head is observed”

- A **probability density function** of a random variable X, $f(x)$
  - (discrete r.v.’s) $f(x) = \Pr(X = x)$. Therefore, $0 \le f(x) \le 1$. (aka, P.M.F)
  - (continuous r.v.’s) $\Pr(a \le X \le b) = \int_a^b f(x)\, dx$, for any a, b and $f(x) \ge 0$.
  - (continuous r.v.’s) $\Pr(X \approx a) = \Pr(a - \epsilon \le X \le a + \epsilon) = \int_{a-\epsilon}^{a+\epsilon} f(x)dx = f(a) \cdot 2\epsilon$

Event independence and conditional probability

- Two events A and B are **independent** if and only if $Pr(A \cap B) = \Pr(A)\Pr(B)$.
  - Pr(“the 1st toss is head” and “the 2nd toss is tail”) = Pr({HT, HH} ∩ {HT, TT}) = Pr({HT}) = 0.25
  - Pr({HT, HH}) * Pr({HT, TT}) = 0.25

- Conditional probability, $\Pr(A|B)$
  - The probability of an event A happening, given that another event B has already occurred.
  - $\Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$, if $\Pr(B) > 0$.
  - A = “the 2nd toss is tail”. B= “at least one head”.
  - $\Pr(\{HT, TT\} | \{HH, HT, TH\}) = \frac{\Pr(\{HT,TT\} \cap \{HH,HT,TH\})}{\Pr(\{HH,HT,TH\})} = \frac{0.25}{0.75} = 1/3$. (≠ Pr(A) = 0.5)
  - $\Pr(A|B) = \Pr(A)$ if and only if A, B are independent.
  - $P(A^c|B) = 1 - P(A|B)$



- $Pr(A_1 \cap A_2 \cap \cdots \cap A_n) = Pr(A_1) \frac{Pr(A_1 \cap A_2)}{Pr(A_1)} \frac{Pr(A_1 \cap A_2 \cap A_3)}{Pr(A_1 \cap A_2)} \cdots \frac{Pr(A_1 \cap A_2 \cap \cdots \cap A_n)}{Pr(A_1 \cap \cdots \cap A_{n-1})}$

$$= Pr(A_1)Pr(A_2|A_1)Pr(A_3|A_1, A_2) \cdots Pr(A_n|A_1, A_2, \dots, A_{n-1})$$

Bayes' theorem

- Suppose a blood test used to detect the presence of a particular banned sports drug is **99% sensitive** and **99% specific.** That is, the test will produce 99% **true positive** results for drug users and 99% **true negative** results for non-drug users. Suppose that **0.5%** of athletes are users of the drug. What is the *likelihood* that *a randomly selected athlete who **tests positive is a user***?

What we already know

What we want to know

$$
\begin{aligned}
P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\
&= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\
&\approx 33.2\%
\end{aligned}
$$

- **Even if an individual tests positive, it is more likely than not (1 - 33.2% = 66.8%) – they do *not* use the drug.** Why? Even though the test appears to be highly accurate, *the number of <u>non-users</u> is* very large *compared to the number of <u>users</u>*. Then, the count of *false positives* will be greater than the count of *true positives*.

*https://www.mun.ca/biology/scarr/

Joint PDF and likelihood function

- Tumor size experiment: $X_1, \cdots, X_{10} \sim iid\ N(\mu_{WT}, \sigma_{WT}^2)$
- A **joint PDF** of multiple r.v.'s, $X_1, \cdots, X_n$ is a PDF of a random vector (n-dim'l) defined on $\mathbb{R}^n$
  - (discrete case) $Pr\big(\ (X_1, \dots, X_n) = (x_1, \dots, x_n)\big)\ = f(x_1, x_2, \dots, x_n)$
  - (continuous case) $Pr(\ (X_1, \dots, X_n) \in D \subset \mathbb{R}^n)\ = \int_D\ f(x_1, x_2, \dots, x_n)\ dx_1 \cdots dx_n$
- $X_1, \cdots, X_n$ are **independent** if and only if

$$pdf_{X_1,\cdots,X_n}(x_1, \dots, x_n) = pdf_{X_1}(x_1)\, pdf_{X_2}(x_2) \cdots pdf_{X_n}(x_n)$$

- PDF of $N(\mu, \sigma^2)$, Gaussian distribution : $\phi(x \mid \mu, \sigma^2)\ = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- Joint PDF of $X_1, \cdots, X_n \sim iid\ N(\mu, \sigma^2)$:

$$pdf_{X_1,\cdots,X_n}(x_1, \dots, x_n \mid \mu, \sigma^2)\ = \ \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \qquad (\ \boldsymbol{x}\ \mapsto f(\boldsymbol{x} \mid \boldsymbol{\theta}),\ \ \boldsymbol{x} \in \mathbb{R}^n\ )$$

- **Likelihood function**
  - is the probability density function value at observed data as a function of parameters, θ.
  - $L(\mu, \sigma^2 \mid x_1, \dots, x_n)\ =\ \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \qquad (\ \boldsymbol{\theta}\ \mapsto f(\boldsymbol{x} \mid \boldsymbol{\theta}),\ \ \boldsymbol{\theta} \in \mathbb{R}^p\ )$
- Consider 1-dimensional Gaussian r.v. X.
  - The likelihood of mean 0 and variance 1, given that the observed (realized) X is 1 = $\phi(\ 1 \mid 0, 1)$
  - The likelihood of mean 1 and variance 1, given that the observed (realized) X is 1 = $\phi(\ 1 \mid 1, 1)$ is greater.

Maximum Likelihood Estimator (MLE)

- Suppose $X_1, \cdots, X_n \sim iid\ N(\mu, \sigma^2)$.
- Log-likelihood function:

$$\log(L(\boldsymbol{\theta} \mid \boldsymbol{x})) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

- MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$ for given observations or data $\{x_1, x_2, \ldots, x_n\}$
  – $\widehat{\boldsymbol{\theta}}^{MLE} = argmax_{\theta}\ L(\boldsymbol{\theta} \mid \boldsymbol{x})$
  – A parameter value depending on data which maximizes the (log-) likelihood function.
  – When n data points are assumed to be independent realizations from a Normal distribution with a common mean and variance,
  [MLE with a closed-form]

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

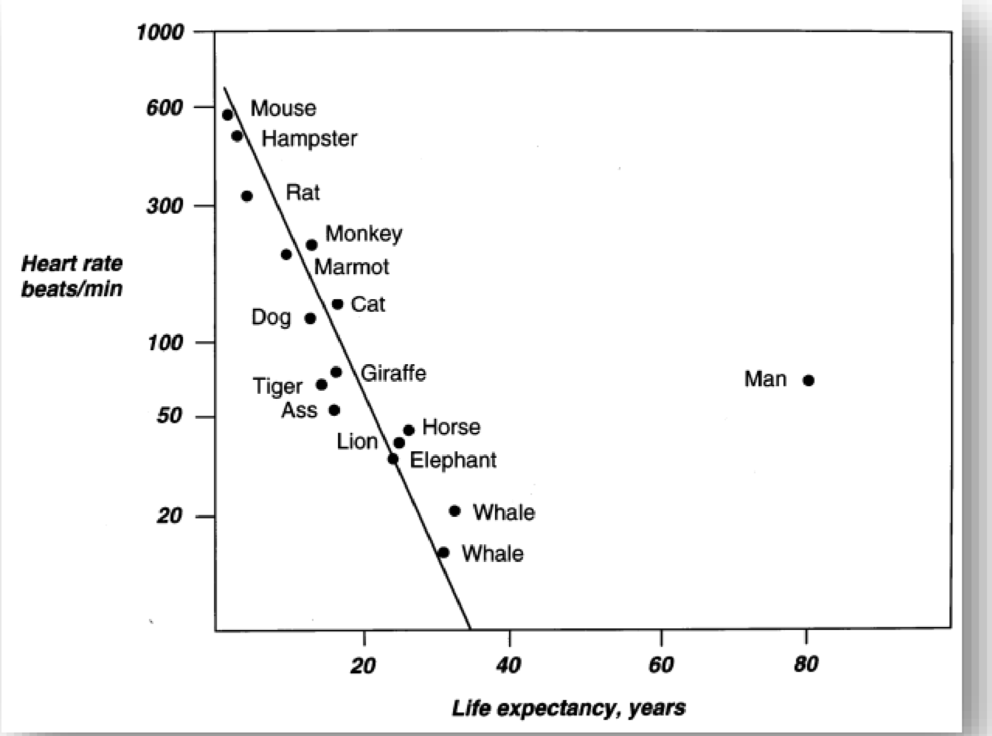
※ MLEs are known to be asymptotically optimal:

$$\text{As n} \to \infty, \quad \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{MLE} - \boldsymbol{\theta}_{true}\right) \Rightarrow^{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

$\boldsymbol{\Sigma}$ is the smallest among any other estimators.

Simple (Linear) Regression Model

Heart rate (HR) vs. Life expectancy (LE) in mammals



$$\text{Log(HR)} = \underbrace{}_{}\text{a} + \text{b*LE} + \text{Error}$$

Log(HR) = a + b*LE + Error

Response/Target variable

Explanatory/Predictor variable, covariate

*Levine (1997, J Am Coll Cardiol)

- Suppose we observe 2 features (x, y variables) for n sampling units.
- Suppose we want to explain y observations using the information about x observations.

$\Rightarrow$ Find a functional (linear) relation between x and y.

(A **sampling unit** is

- 'physical entity or object from which data is collected.'
- what you are actually picking when you take a sample.
- In the example, a species is the sampling unit. Two features are HR and LE.)

Model eqn:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim iid\ N(0, \sigma^2), \quad i = 1, 2, \ldots, n$$

- Log-likelihood function conditional on $\boldsymbol{X} = \boldsymbol{x}$

$$\log(L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{x})) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \alpha - \beta x_i)^2$$

- MLE (=Least-Squares Estimator, LSE)

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x - \bar{x})(y - \bar{y})}{\sum_i (x - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_i \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2$$

Conditional probability density function

- Suppose a random vector (X, Y) follows a Bivariate normal distribution with a mean vector and covariance matrix.
  - $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} E[(X-\mu_X)^2] & E[(X-\mu_X)(Y-\mu_Y)] \\ E[(X-\mu_X)(Y-\mu_Y)] & E[(Y-\mu_Y)^2] \end{pmatrix}$

- $pdf_X(x) = \phi(x \mid \mu_X, \sigma_X^2) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}$
- Conditional PDF of Y conditional on $X = x$

(discrete case) $\quad pdf_{Y|X}(y \mid x) = Pr(Y = y \mid X = x)$

(continuous case) $\quad pdf_{Y|X}(y \mid x) = \frac{pdf_{X,Y}(x,y)}{pdf_X(x)}$

- Auto-regressive model of order 1 for a time series $\{X_t\}$

$$X_t = \alpha + \beta X_{t-1} + \epsilon_t, \qquad \epsilon_t \sim (0, \sigma^2)$$

  - $|\beta| < 1$ is necessary for $\{X_t\}$ to be stationary.
- AR(1)-Hidden Markov Model

A_t : a hidden state $\in \{1, \dots, M\}$

$$X_t = \alpha_{A_t} + \beta_{A_t} X_{t-1} + \epsilon_t, \ \epsilon_t \sim N(0, \sigma_{A_t}^2)$$

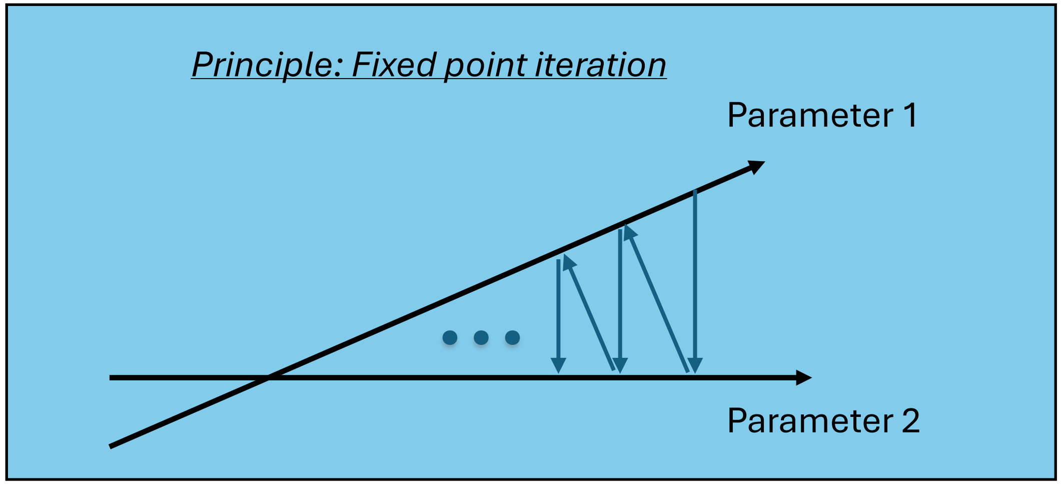
Outline

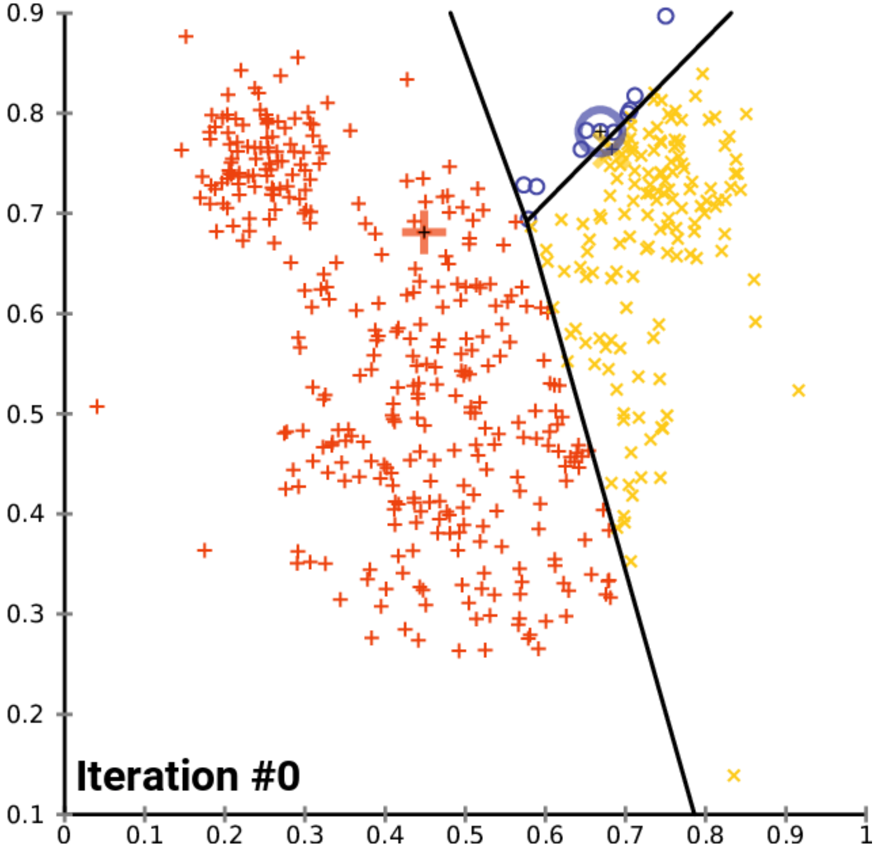
1. Review for Stochastic Models
   - Random variables, PDF, likelihood function, etc
   - Goal: to have clear understanding of Autoregressive-Hidden Markov Models

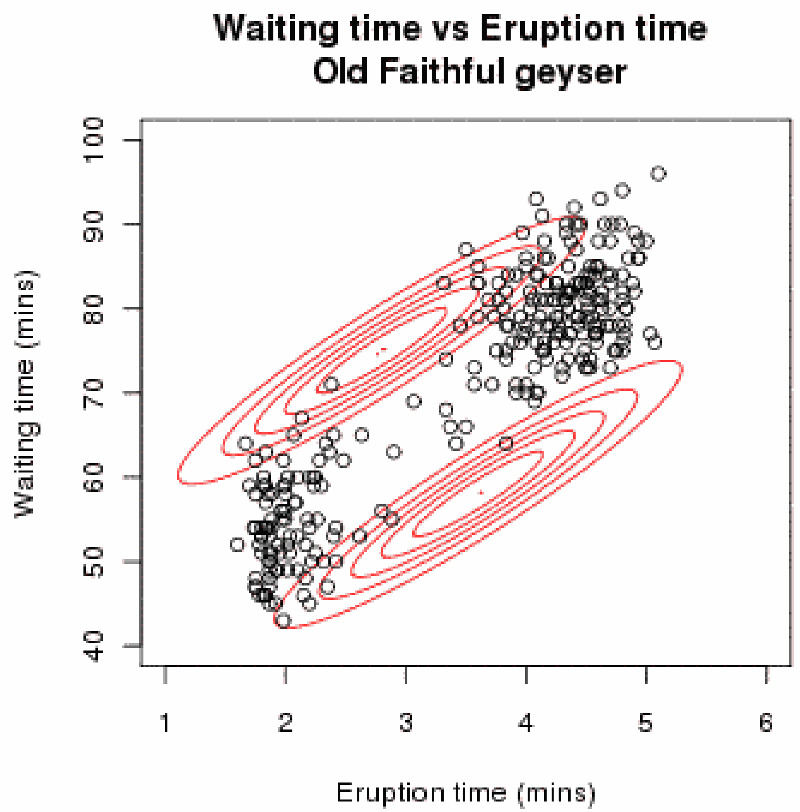
2. Expectation-Maximization algorithm (EM)
   for Gaussian Mixture Models (GMM)

K-means and EM for GMM

- Both are techniques to find clusters in data points.
- K-means is a hard-thresholding.
- EM for GMM is a soft-thresholding.



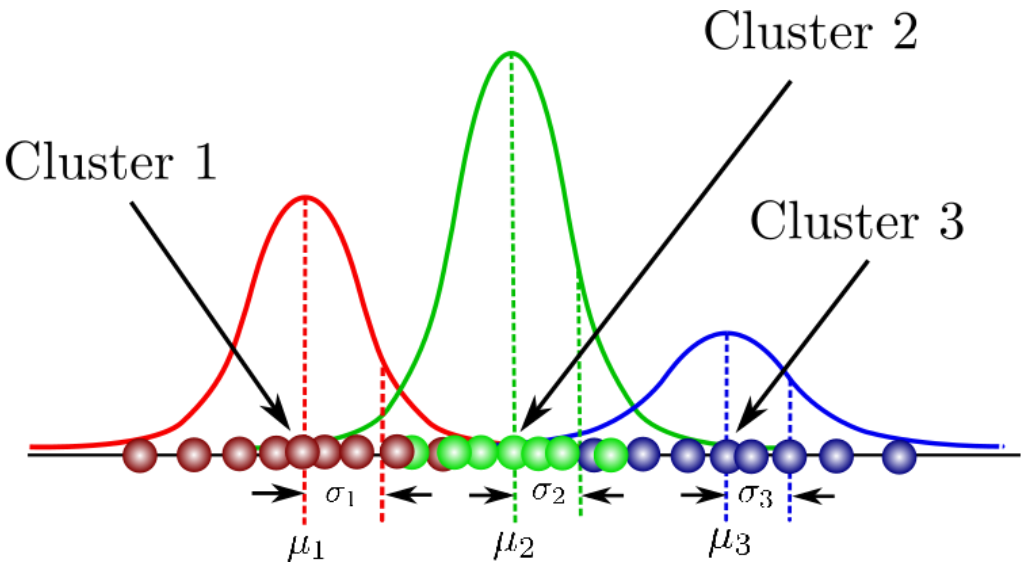




*By Chire, https://commons.wikimedia.org/w/index.php?curid=6319946

Gaussian Mixture Models (GMMs)

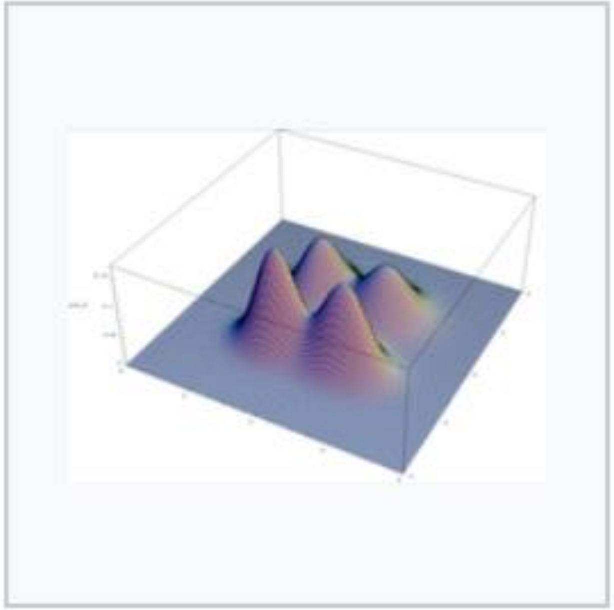
1-dimensional data points



$$pdf(x) = \sum_{k=1}^{K} p_k \phi_{\mu_k,\sigma_k}(x)\,, \qquad \phi_{\mu,\sigma}(x) \sim Nornal(\mu, \sigma^2)$$

- K : number of clusters
- p_k: mixing probabilities

2-dimensional data points



Multivariate mixture distribution, showing four modes

EM: Maximum Likelihood Estimation (MLE) for GMMs

- Suppose data points in $\mathbb{R}^d$ are $\{x_1, x_2, \cdots, x_n\}$.
- K-component GMM (K is known):
  - $\Pr(X{\approx}x) = \sum_{k=1}^{K} \Pr(Z = k) \Pr(X{\approx}x \mid Z = k)$, $\quad$ Z is a membership (latent) variable
  - $pdf(x) = \sum_{k=1}^{K} p_k \phi_{\mu_k,\sigma_k}(x)$
  - $pdf(x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \phi_{\mu_k,\sigma_k}(x_i)$

- Law of total probability

$S = \cup_i B_i$, $B_i \cap B_j = \emptyset$ for $\mathrm{i} \neq j$

(The sample space has a partition. Then)

Because $A = A \cap S = A \cup (\cup_i B_i) = \cup_i (A \cap B_i)$,

$$Pr(A) = Pr\big(\cup_i (A \cap B_i)\big) = \sum_i Pr(A \cap B_i)$$
$$= \sum_i Pr(B_i)\ Pr(A|B_i)$$

(divide and conquer)

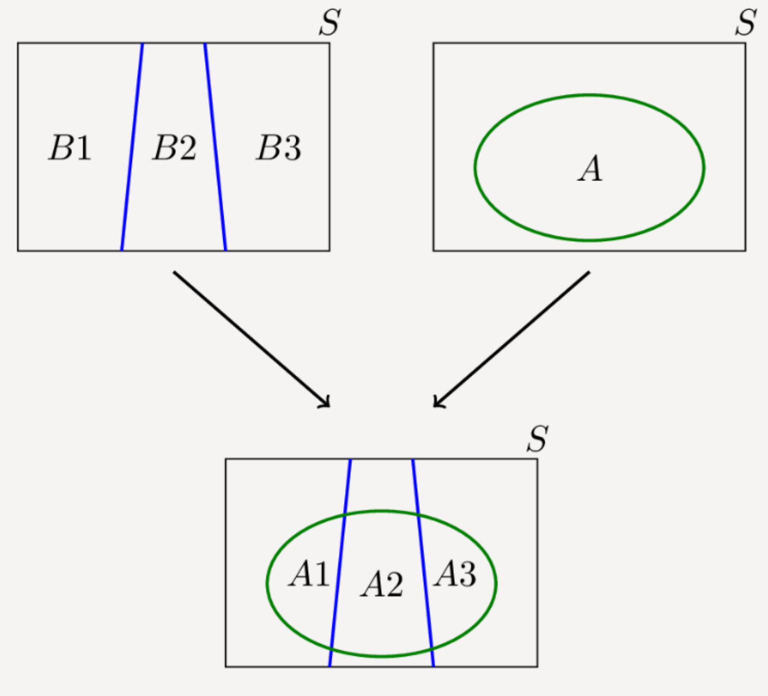


Fig.1.24 - Law of total probability.

EM: Maximum Likelihood Estimation (MLE) for GMMs

- Suppose data points in $\mathbb{R}^d$ are $\{x_1, x_2, \cdots, x_n\}$.
- K-component GMM (K is known):
  - $\Pr(X \approx x) = \sum_{k=1}^{K} \Pr(Z = k) \Pr(X \approx x \mid Z = k)$, Z is a membership (latent) variable
  - $pdf(x) = \sum_{k=1}^{K} p_k \phi_{\mu_k, \sigma_k}(x)$
  - $pdf(x_1, \dots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \phi_{\mu_k, \sigma_k}(x_i)$

- How to optimize the joint p.d.f to obtain MLE of $\theta = (p_1, p_2, \dots, p_K; \mu_1, \dots, \mu_K; \sigma_1, \dots, \sigma_K)$?
- **Idea**:
  - Consider the membership variables, $z = (z_1, z_2, \dots, z_n)$.
  - Instead of directly optimizing θ, optimize θ when z is given.
  - Then, optimize z when θ is given.
  - Repeat until convergence.

$z = (z_1, z_2, \dots, z_n)$
(Membership observations)

$\theta = (p_1, p_2, \dots, p_K; \mu_1, \dots, \mu_K; \sigma_1, \dots, \sigma_K)$
(Mixing probabilities, distribution parameters)

EM iterations for GMM

- Expectation step:
  - Based on a current estimate of $\theta^{(t)}$,
    compute membership probabilities (expectation), $\Pr\left(Z_i = k \mid \theta^{(t)}, X_i = x_i\right)$ for $i = 1, 2, \ldots, n$.

$$E\left(I(Z_i = k) | \theta^{(t)}, X_i = x_i\right) = Pr\left(Z_i = k | \theta^{(t)}, X_i = x_i\right) = \frac{\Pr(Z_i = k) \Pr\left(X_i = x_i | Z_i = k, \theta^{(t)}\right)}{\Pr(X_i = x_i \mid \theta^{(t)})}$$

Bayes' Theorem

$$Pr^{(t)}\left(Z_i = k | \theta^{(t)}, X_i = x_i\right) = \frac{p_k^{(t)} \phi_{\mu_k^{(t)}, \sigma_k^{(t)}}(x_i)}{\sum_{k=1}^{K} p_k^{(t)} \phi_{\mu_k^{(t)}, \sigma_k^{(t)}}(x_i)}$$

Law of total prob.

EM iterations for GMM

- Expectation step at iteration t:
  - Based on a current estimate of $\theta^{(t)}$, compute membership probabilities (expectation), $\Pr\left(Z_i = k \mid \theta^{(t)}, X_i = x_i\right)$ for $i = 1, 2, \ldots, n$.

$$E\left(I(Z_i = k) \mid \theta^{(t)}, X_i = x_i\right) = Pr\left(Z_i = k \mid \theta^{(t)}, X_i = x_i\right) = \frac{\Pr(Z_i = k)\Pr\left(X_i = x_i \mid Z_i = k, \theta^{(t)}\right)}{\Pr(X_i = x_i \mid \theta^{(t)})}$$

Bayes' Theorem

$$Pr^{(t)}\left(Z_i = k \mid \theta^{(t)}, X_i = x_i\right) = \frac{p_k^{(t)} \phi_{\mu_k^{(t)}, \sigma_k^{(t)}}(x_i)}{\sum_{k=1}^{K} p_k^{(t)} \phi_{\mu_k^{(t)}, \sigma_k^{(t)}}(x_i)}$$

Law of total prob.

- Maximization step at iteration $(t+1)$:
  - When the membership probability of each data point ($Pr^{(t)}$ $(Z_i = k)$) is given, the maximization of the joint p.d.f results in the MLE having "weighted" average forms.
  - $p_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} Pr^{(t)}\,(Z_i = k)$, $\quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} Pr^{(t)}\,(Z_i = k)\, x_i}{\sum_{i=1}^{n} Pr^{(t)}\,(Z_i = k)}$, $\quad \sigma_k^{2\,(t+1)} = \frac{\sum_{i=1}^{n} Pr^{(t)}\,(Z_i = k)\left(x_i - \mu_k^{(t)}\right)^2}{\sum_{i=1}^{n} Pr^{(t)}\,(Z_i = k)}$

$\Pr(Z_i = k)$
(membership probability)

Maximization →

← Expectation

$\theta = (p_1, p_2, \ldots, p_K; \mu_1, \ldots, \mu_K; \sigma_1, \ldots, \sigma_K)$
(Mixing probabilities, distribution parameters)

Quiz

- Suppose a 1-dim'l random variable X follows a 2-component GM distribution, 0.25*N(0, 1) + 0.75*N(2,1). N(0,1) is the 1st component and N(2,1) is the 2nd.

1. When X=1 is observed, compute the probability that it is from the 1st and 2nd normal distribution. That is, compute $P(Z = 1 \mid X = 1)$ and $P(Z = 2 \mid X = 1)$.
2. Do the same for X=-5.
3. Do the same for X=+5.

(Don't need to compute the real number, but do not write just formula.)