Directed Reading

Supervisor: Prof.Oliver Schulte
Student: Yejia Liu

outline

- Gaussian Distribution
- Gaussian Mixture Model
- K-Means Clustering
- K-Means Clustering vs. GMM EM
- Bayesian Gaussian Mixture Model
- Exception Mining(Outlier Detection)
- Data Tables for AHL&NHL

Gaussian Distribution

- mainly used for continuous variables

(1) single-variable:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

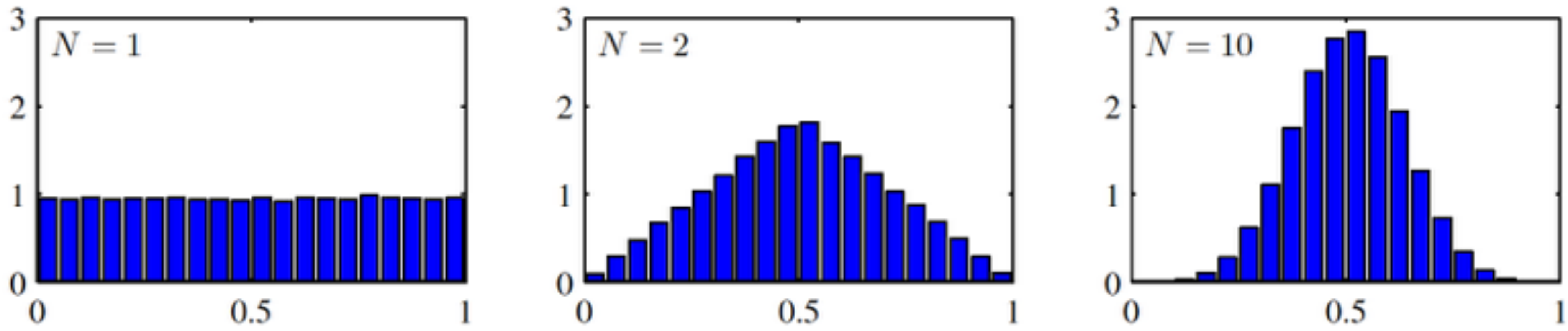
(2) multi-variate: for k-dimensional vector **x**, the distribution is:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}),\quad \boldsymbol{\mu} = \mathrm{E}[\mathbf{x}] = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]],$$

$$\boldsymbol{\Sigma} =: \mathrm{E}[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}] = [\mathrm{Cov}[X_i, X_j]; 1 \le i, j \le k].$$

Gaussian Distribution

- why gaussian?



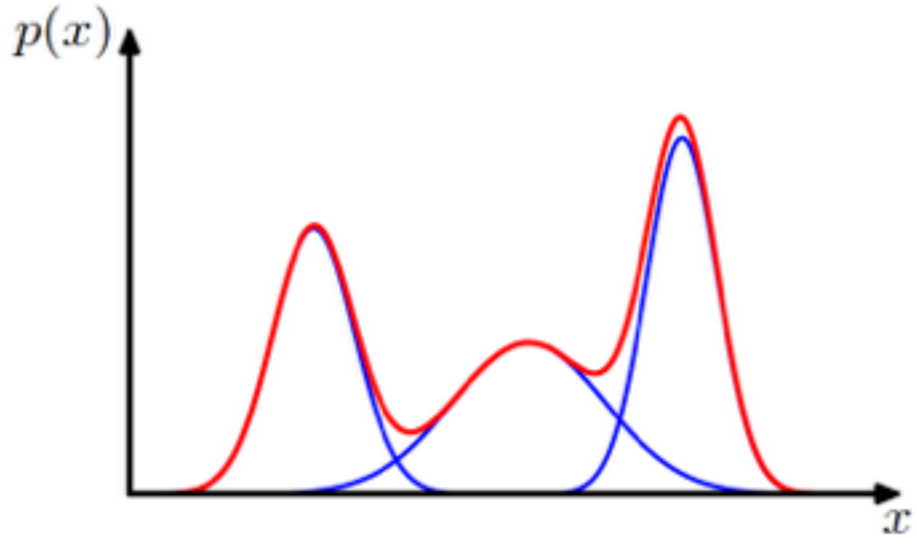
- limitations

  * too many free parameters

  * unimodal, not good at approximating multimodal distribution

Gaussian Mixture Model

- A linear superposition of Gaussian components



- Gaussian Mixture Distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Mixture Model

- Maximum likelihood for GMM

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

problems: (1) presence of singularities, when $\boldsymbol{\mu}_j = \mathbf{x}_n$,consider if $\sigma_j \to 0$,

term in the likelihood function $\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) = \frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma_j}$ will go to infinity.

(2) identifiability, many components can give the same distribution

Finding Maximum likelihood: EM for Gaussian Mixtures

Goal: Given a gaussian mixture model, maximize the likelihood

Steps:

(1) Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k , and evaluate the initial value of log likelihood

(2) E step. Using current values of parameters to evaluate posterior probabilities

Gaussian Mixture Model

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

(3) M step. Using the current posterior to evaluate parameters

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}$$

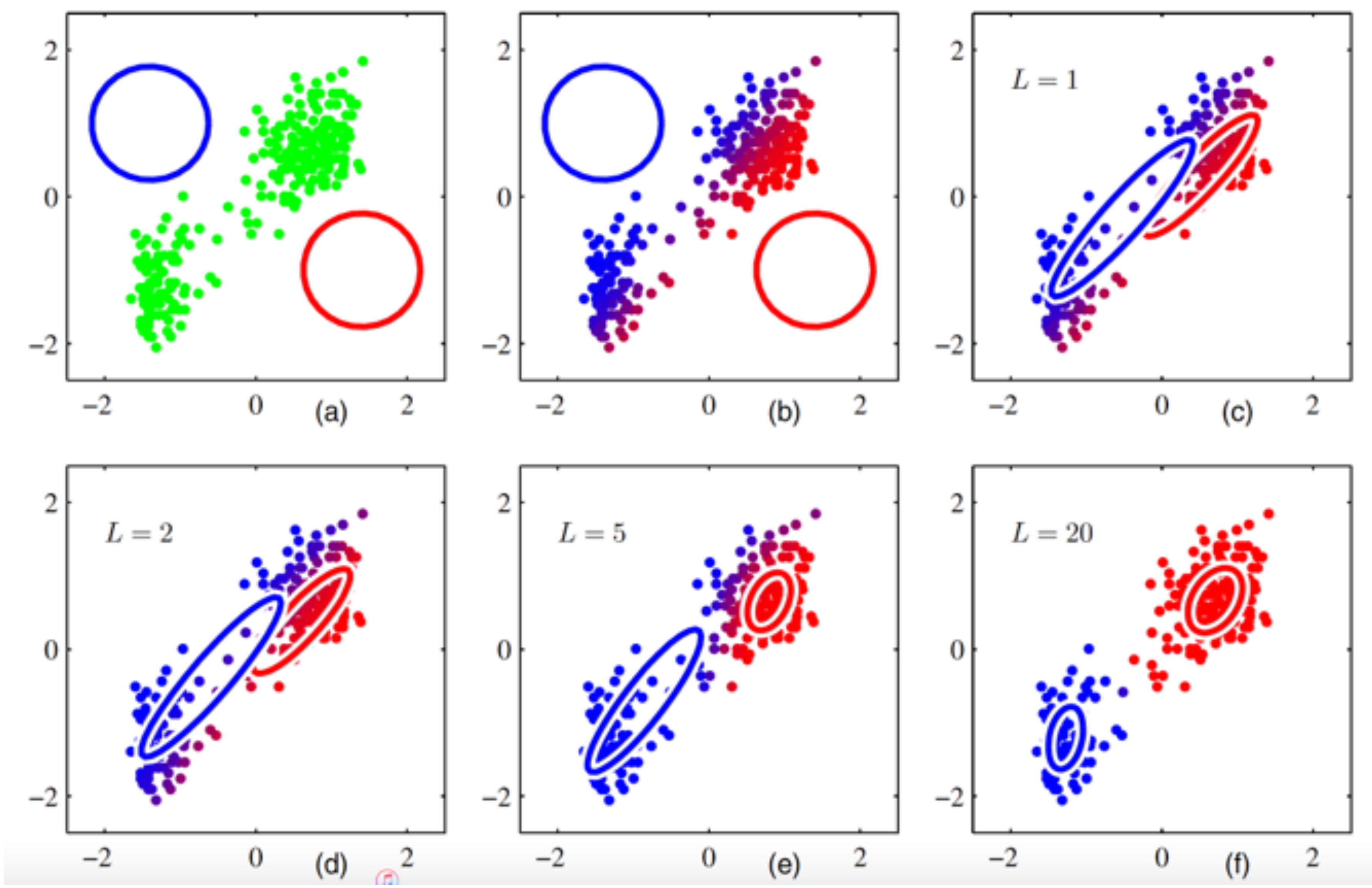
where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

(4) Evaluate the log likelihood

Gaussian Mixture Model

Illustration:



K-means Clustering

- Algorithm

Let $X = \{x_1, x_2, x_3, \ldots\ldots.., x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

K-means Clustering vs. GMM EM

- K-means

(1) **Hard assign** a data point to one particular cluster on convergence

(2) It makes use of the L2 norm when optimizing

- GMM EM

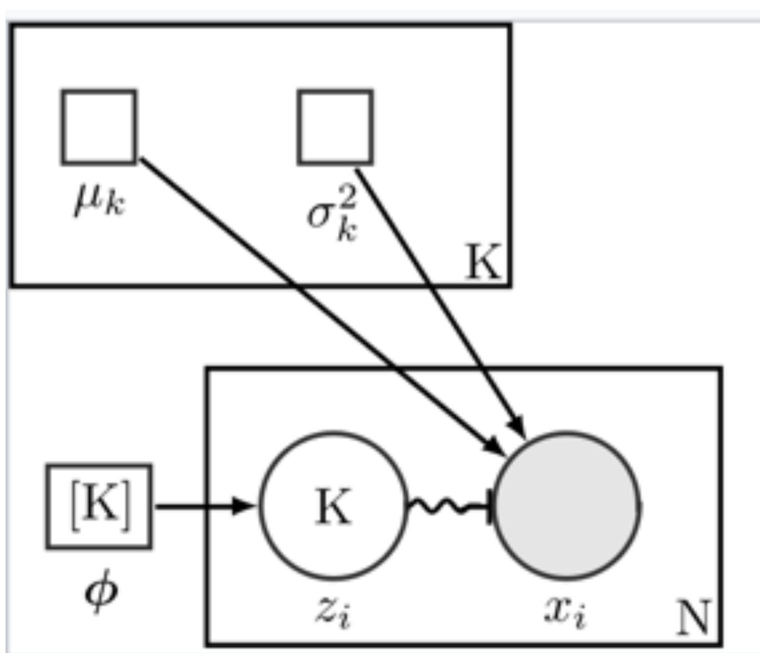
(1) **Soft assign** a point to clusters (so it give a probability of any point belonging to any centroid)

(2) It doesn't depend on the L2 norm, but is based on the Expectation, i.e., the probability of the point belonging to a particular cluster. This makes K-means biased towards spherical clusters

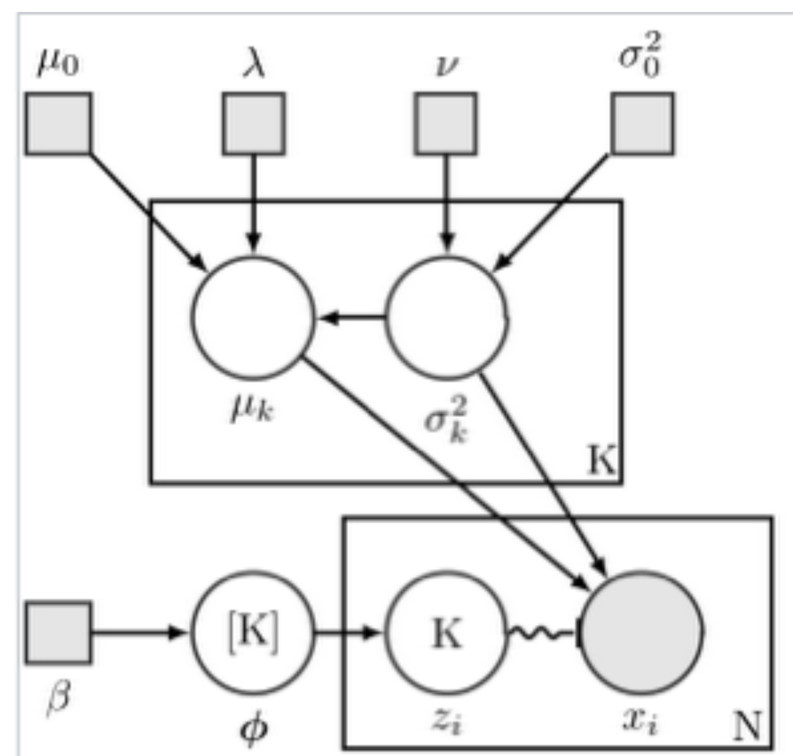
It's very common to run k-means in order to find a suitable initialization for a Gaussian Mixture model that subsequently adopts using EM.

Bayesian Gaussian Mixture Model

- In a Bayesian setting, the mixture weights and parameters will themselves be random variables, and **prior distributions** will be placed over the variables.
- A Bayesian Gaussian mixture model is commonly extended to fit a vector of unknown parameters, or multivariate normal distributions



Non-Bayesian Gaussian mixture model using plate notation. Smaller squares indicate fixed parameters; larger circles indicate random variables. Filled-in shapes indicate known values. The indication [K] means a vector of size *K*.



Bayesian Gaussian mixture model using plate notation. Smaller squares indicate fixed parameters; larger circles indicate random variables. Filled-in shapes indicate known values. The indication [K] means a vector of size *K*.

Exception Mining(Outlier Detection)

- Using BGMM EM to fit means of datasets, the fit each player, get exception by calculating L2 distance (i.e. package *sklearn.mixture.GaussianMixture*)
- Using k-means to find outlier(code example: https://github.com/liuyejia/South-African-Heart-Disease-data-analysis-using-python/tree/master/part3)

row	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12	5.73	23.11	Present	49	25.3	97.2	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.6	3.5	27.78	Present	60	25.99	57.34	49	1
6	132	6.2	6.47	36.21	Present	62	30.77	14.14	45	0
7	142	4.05	3.38	16.2	Absent	59	20.81	2.62	38	0
8	114	4.08	4.59	14.6	Present	62	23.11	6.72	58	1
9	114	0	3.83	19.4	Present	49	24.86	2.49	29	0
10	132	0	5.8	30.96	Present	69	30.11	0	53	1
11	206	6	2.95	32.27	Absent	72	26.81	56.06	60	1

Data Tables for AHL&NHL

AHL_NHL_skater_demograhic:

- metrics:

Field	Type	Null	Key	Default	Extra
PlayerId	int(11)	YES		NULL	
PlayerName	text	YES		NULL	
Height	text	YES		NULL	
Weight	int(11)	YES		NULL	
Position	text	YES		NULL	
Country	text	YES		NULL	

- sample:

PlayerId	PlayerName	Height	Weight	Position	Country
8444894	Greg Adams	6'4"	196	L	Canada
8444919	Tommy Albelin	6'2"	195	D	Sweden
8445000	Dave Andreychuk	6'4"	225	L	Canada
8445176	Donald Audette	5'8"	191	R	Canada
8445266	Murray Baron	6'3"	236	D	Canada

Data Tables for AHL&NHL

AHL_NHL_skater: (performance)

- metrics:

Field	Type	Null	Key	Default	Extra
SkaterId	int(11)	YES		NULL	
Season	mediumtext	YES		NULL	
SeasonType	mediumtext	YES		NULL	
Team	mediumtext	YES		NULL	
GP	int(11)	YES		NULL	
CareerG	int(11)	YES		NULL	
CareerA	int(11)	YES		NULL	
CareerP	int(11)	YES		NULL	
CareerPlusMinus	int(11)	YES		NULL	
CareerPIM	int(11)	YES		NULL	
PPG	int(11)	YES		NULL	
SHG	int(11)	YES		NULL	
GWG	int(11)	YES		NULL	
CareerS	int(11)	YES		NULL	
ShotPercentage	double	YES		NULL	

- sample:

SkaterId	Season	SeasonType	Team	GP	CareerG	CareerA	CareerP	CareerPlusMinus	CareerPIM	PPG	SHG	GWG	CareerS	ShotPercentage
8444894	1984-1985	Regular Season	MAINE MARINERS-AHL	41	15	20	35	0	12	0	0	0	0	0
8444894	1984-1985	Playoffs	MAINE MARINERS-AHL	11	3	4	7	0	0	0	0	0	0	0
8444894	1984-1985	Regular Season	DEVILS	36	12	9	21	-14	14	5	0	0	65	18.5
8444894	1985-1986	Regular Season	DEVILS	78	35	42	77	-7	30	10	0	2	202	17.3