Clustering

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

How can we sort objects?

- size,
- shape,
- colour,
- texture,
- ingredients.

Examples of Clustering

Examples of clustering problems:

- articles with similar content,
- search engines,
- suggestions from streaming sites,
- image segmentation,
- lossy data compression,
- bio informatics.
- Any suggestion?

K Means Clustering:

- Let $\mathbf{v}_1, \ldots, \mathbf{v}_N$ be the feature vectors of N data samples.
- Number of clusters fixed, K.
- *Hard clustering* assigns sample to the cluster with the nearest centre.
- Find cluster centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ such that the sum of the squared distances of each data sample to its assigned cluster centre is minimal.
- Minimize

$$J = \sum_{n=1}^{N} \min_k \|\mathbf{v}_n - \boldsymbol{\mu}_k\|^2.$$

- *NP hard*: no known algorithm to solve this in polynomial time, since as cluster centres move around, for each sample its nearest cluster centre can change.

- Separate interdependency.
- *Hidden (latent)* variables $\mathbf{z}_n$, one for each data sample $\mathbf{v}_n$.
- *1-of-K representation*: $\mathbf{z}_n \in \{0, 1\}^K$, one entry 1 and the others have to be 0.
- 1 in the k^{th} entry indicates that $\boldsymbol{\mu}_k$ is the nearest cluster centre to $\mathbf{v}_n$.
- Let z_{nk} be the k^{th} component of $\mathbf{z}_n$,

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{v}_n - \boldsymbol{\mu}_k\|^2.$$

K Means Clustering

- Quadratic in $\boldsymbol{\mu}_k$.
- Find minimum by differentiating with respect to $\boldsymbol{\mu}_k$ and setting this to zero,

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk}\mathbf{v}_n}{\sum_{n=1}^{N} z_{nk}}.$$

- $\sum_{n=1}^{N} z_{nk}$ number of data samples for which $\boldsymbol{\mu}_k$ is the closest centre.
- $\sum_{n=1}^{N} z_{nk}\mathbf{v}_n$ is the sum of those samples.
- $\boldsymbol{\mu}_k$ is the mean of the samples assigned to this particular cluster.
- Adjust indicator vectors $\mathbf{z_n}$.
- Alternate.
- Terminate, when after moving the centres, none of the indicator vectors changes.

- Local minimum.
- Highly dependent on the initialization of the cluster centres.
- At the start centres chosen randomly, or samples randomly assigned to clusters.
- Algorithm is run (possibly in parallel) with many different initializations. After convergence, the result with the lowest value of J is chosen.

K Medoids Algorithm:

- Use any dissimilarity measure.
- Minimization depends on the differentiability of the dissimilarity measure, and whether it is possible to find where the derivative vanishes.
- If this is not possible, we require each cluster centre to be one of the data samples.
- The minimization with respect to $\boldsymbol{\mu}_k$ is then a search among the data samples assigned to the k^{th} cluster.

K Means Clustering

4 clusters

16 clusters

Mixture Models:

- $\mathbf{z}_n$ indicates which process generated $\mathbf{v}_n$.
- Let $p_k(\mathbf{v})$ be the probability distribution of process k.
- Let π_k be the probability that process k generates a sample,

$$0 \leq \pi_k \leq 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1.$$

- Latent variables $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are drawn from from a probability distribution $p(\mathbf{z})$,

$$p(z_k = 1) = \pi_k.$$

Mixture Models

- Probability of $\mathbf{v}$ given $\mathbf{z}$ is the conditional probability distribution

$$p(\mathbf{v}|\mathbf{z}) = p(\mathbf{v}|z_k = 1) = p_k(\mathbf{v}).$$

- *Mixture distribution*:

$$\begin{aligned} p(\mathbf{v}) &= \sum_{\mathbf{z}} p(\mathbf{v}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{v}|\mathbf{z}) \\ &= \sum_{k=1}^{K} p(z_k = 1) p(\mathbf{v}|z_k = 1) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{v}). \end{aligned}$$

- π_k are known as *mixing coefficients*.

Bayes Rule

Bayes Rule: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

- A and B are events.
- $P(A)$ and $P(B)$ are the probabilities of A and B without regard to each other.
- $P(A|B)$ is the *conditional probability* of A given that B is true. $P(B|A)$ is the conditional probability of B given that A is true.
- Often expressed as $P(A|B) \propto P(B|A)P(A)$ where $\propto$ means that the two sides are proportional to each other.

Mixture Models

- *Responsibility* process k takes for explaining the sample $\mathbf{v}$ is the probability that it was generated by process k:

$$p(z_k = 1|\mathbf{v}) = \frac{p(z_k = 1)p(\mathbf{v}|z_k = 1)}{p(\mathbf{v})} = \pi_k \frac{p_k(\mathbf{v})}{p(\mathbf{v})}$$

  by Bayes rule.
- Responsibilities sum to 1.
- *Soft clustering* makes cluster assignments according to the values $p(z_{nk} = 1|\mathbf{v}_n)$ for $k = 1, \ldots, K$.
- Possible that for a particular sample the probabilities are the same for two (or even more) values of k. These are samples which lie between clusters.

Mixture Models

- Maximize joint likelihood of the data samples $\mathbf{v}_1, \dots, \mathbf{v}_n$:

$$\prod_{n=1}^{N} p(\mathbf{v}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n).$$

- Or alternatively its logarithm

$$\mathcal{L} = \sum_{n=1}^{N} \log p(\mathbf{v}_n) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right).$$

- Subject to $\sum_{k=1}^{K} \pi_k = 1$.
- Using a Lagrange multiplier λ, we maximize

$$\sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Mixture Models

- Differentiating with respect to π_k and setting to zero gives

$$\sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n)} p_k(\mathbf{v}_n) + \lambda = 0.$$

- Multiplying through by π_k and summing over all k, gives $\lambda = -N$.
- Inserting this and again multiplying by π_k, results in

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} p(z_{nk} = 1 | \mathbf{v}_n).$$

- The mixing coefficient π_k is the average responsibility that all data samples are generated by process k.
- Note: $p(z_{nk} = 1|\mathbf{v}_n)$ depend on π_k itself $\Rightarrow$ iterative procedure.

Mixture Models

Caution:

- Assume that $K = 2$ and that all data samples are roughly grouped together apart from one outlier.
- π_1 tends to $(N-1)/N$ while π_2 tends to $1/N$.
- $p_1(\mathbf{v})$ roughly describes the distribution of $N-1$ samples.
- Likelihood can be increased again and again by concentrating the probability mass of $p_2(\mathbf{v})$ more and more tightly around the outlier.
- Evaluation of $p_2(\mathbf{v})$ at the outlier tends to infinity.
- Area where $p_2(\mathbf{v})$ is zero or close to zero tends to zero.
- $K!$ equivalent solutions.

- $p_k(\mathbf{v})$, $k = 1, \ldots, K$, are normal distributions, $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,

$$p_k(\mathbf{v}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{v}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k)\right).$$

- Derivative of $p_k(\mathbf{v})$ with respect to $\boldsymbol{\mu}_k$:

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}p_k(\mathbf{v}) = p_k(\mathbf{v})\boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k).$$

- Derivative of $p_k(\mathbf{v})$ with respect to $\boldsymbol{\Sigma}_k$:

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_k}p_k(\mathbf{v}) = -\frac{1}{2}p_k(\mathbf{v})\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}-\boldsymbol{\mu}_k)(\mathbf{v}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right].$$

Gaussian Mixture Models

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L} &= \sum_{n=1}^{N} \frac{1}{p(\mathbf{v}_n)} \pi_k p_k(\mathbf{v}_n) \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k) \\ &= \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n) \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n - \boldsymbol{\mu}_k).\end{aligned}$$

- Expected number of samples in cluster k: $N_k = \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n)$.
- $\boldsymbol{\mu}_k$ is a weighted average of all samples in the data set where the weights are the responsibilities that the sample was generated by process k,

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n)\mathbf{v}_n.$$

Gaussian Mixture Models

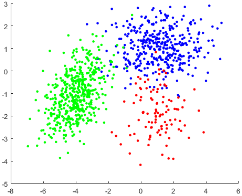
$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{L} = \frac{1}{2}\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{v}_n)\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{v}_n-\boldsymbol{\mu}_k)(\mathbf{v}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right].$$

- Set to zero and multiply through with 2 and $\boldsymbol{\Sigma}_k$ from both sides.
- Similar to sample covariance,

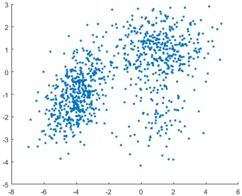
$$\boldsymbol{\Sigma}_{\boldsymbol{k}} = \frac{1}{N_k}\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{v}_n)(\mathbf{v}_n-\boldsymbol{\mu}_k)(\mathbf{v}_n-\boldsymbol{\mu}_k)^T.$$

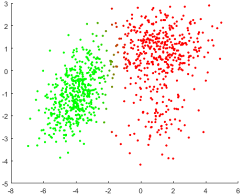
Gaussian Mixture Models

1. Choose K and convergence threshold.
2. Initialize means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k for $k = 1, \ldots, K$, (e.g from K means) and calculate the initial value of the logarithm of the likelihood.
3. For $n = 1, \ldots, N$ and $k = 1, \ldots, K$, calculate all the responsibilities $p(z_{nk} = 1|\mathbf{v}_n)$.
4. Use these responsibilities to update means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$, and mixing coefficients π_k for $k = 1, \ldots, K$.
5. Evaluate the change in the logarithm of the likelihood and terminate if this is below the convergence threshold (or if the change in parameters is below the convergence threshold). Otherwise return to step 3.

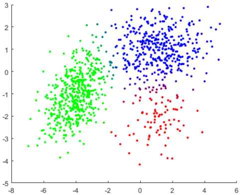


Data samples from three processes.





Two clusters fitted.



Three clusters fitted.

Expectation-Maximization

- How to generally maximize

$$\mathcal{L} = \sum_{n=1}^{N} \log p(\mathbf{v}_n) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k p_k(\mathbf{v}_n) \right)?$$

- Note: When normal distributions are used, maximizing

$$\begin{aligned} \widehat{\mathcal{L}} &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1 | \mathbf{v}_n) \left[\log \pi_k + \log p_k(\mathbf{v}_n) \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1 | \mathbf{v}_n) \log \Big(p(z_{nk} = 1) p(\mathbf{v}_n | z_{nk} = 1) \Big) \end{aligned}$$

leads to the same update formulae.

Expectation-Maximization

- $\widehat{\mathcal{L}}$ is the expectation of the logarithm of the complete data likelihood

$$\sum_{n=1}^{N} \log p(\mathbf{v}_n, \mathbf{z}_n),$$

  where the expectation is taken with respect to the responsibilites, that is the posterior probabilities of the latent variables.
- How are $\mathcal{L}$ and $\widehat{\mathcal{L}}$ related and why do the parameters where their derivatives vanish coincide?

Expectation-Maximization

- Using the product rule for probabilities,

$$\sum_{n=1}^{N} \log p(\mathbf{v}_n, \mathbf{z}_n) = \sum_{n=1}^{N} \log p(\mathbf{v}_n) + \log p(\mathbf{z}_n|\mathbf{v}_n).$$

- Both sides are as functions of the random variables $\mathbf{z}_n$, and the expectation with respect to any distribution $q(\mathbf{z}_n)$ can be taken.
- Since $\mathbf{z}_n$ is a 1-of-K representation, the expectation is calculated by summing over all possible values for $\mathbf{z}_n$.

Expectation-Maximization

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(\mathbf{v}_n, z_{nk}=1) \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(\mathbf{v}_n) + \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(z_{nk}=1|\mathbf{v}_n) \\
&= \underbrace{\sum_{n=1}^{N}\log p(\mathbf{v}_n)}_{\mathcal{L}} + \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log p(z_{nk}=1|\mathbf{v}_n),
\end{aligned}
$$

because of $\sum_{k=1}^{K} q(z_{nk}=1) = 1$.

Expectation-Maximization

- Subtracting and adding the term $\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1)\log q(z_{nk}=1)$ gives

$$\begin{aligned}\mathcal{L} &= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\frac{p(\mathbf{v}_n, z_{nk}=1)}{q(z_{nk}=1)} \\ &\quad+\sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}=1)\log\frac{q(z_{nk}=1)}{p(z_{nk}=1|\mathbf{v}_n)},\end{aligned}$$

- In the last line, each sum over $k = 1, \ldots, K$ is the *Kullback–Leibler divergence (KL divergence)* from the discrete distribution $p(\mathbf{z}_n|\mathbf{v}_n)$ to the discrete distribution $q(\mathbf{z}_n)$.

$$\mathcal{L} = \widetilde{\mathcal{L}} + D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n)).$$

Kullback–Leibler divergence

- Given two discrete probability distributions P and Q, the Kullback–Leibler divergence from Q to P is defined as

$$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

- Only defined if $Q(i) = 0$ implies $P(i) = 0$ to avoid a division by zero.
- If both distributions are the same, then the Kullback–Leibler divergence is zero.
- It is also non-negative, since $\log x \leq x - 1$ and therefore

$$\begin{aligned} D_{KL}(P\|Q) &= -\sum_i P(i) \log \frac{Q(i)}{P(i)} \geq -\sum_i P(i) \left(\frac{Q(i)}{P(i)} - 1 \right) \\ &= -\sum_i Q(i) + \sum_i P(i) = 0. \end{aligned}$$

Expectation-Maximization

- $\widetilde{\mathcal{L}}$ is a lower bound for $\mathcal{L}$, since $D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n))$ is non-negative.

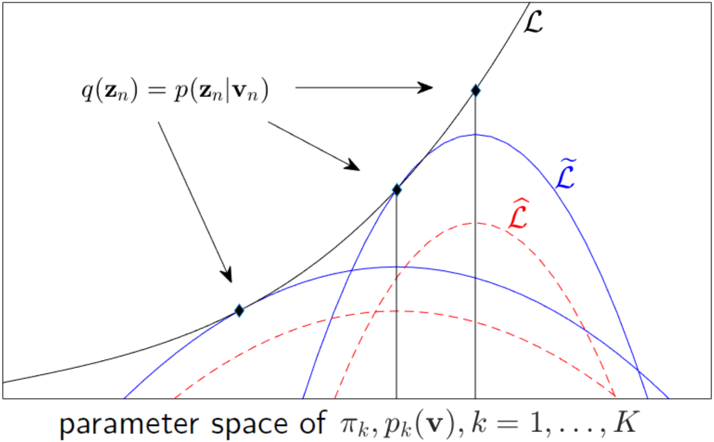
$$\begin{aligned}
\widetilde{\mathcal{L}} &= \sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\frac{p(\mathbf{v}_n, z_{nk}=1)}{q(z_{nk}=1)} \\
&= \overbrace{\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log\Big(p(z_{nk}=1)p(\mathbf{v}_n|z_{nk}=1)\Big)}^{\widehat{\mathcal{L}}} \\
&\quad -\sum_{k=1}^{K}\sum_{n=1}^{N} q(z_{nk}=1)\log q(z_{nk}=1).
\end{aligned}$$

- It has the same value as $\mathcal{L}$, if $q(\mathbf{z}_n)=p(\mathbf{z}_n|\mathbf{v}_n)$, because then $D_{KL}(q(\mathbf{z}_n)\|p(\mathbf{z}_n|\mathbf{v}_n))=0$.

Maximize by alternating between

- maximizing with respect to $q(\mathbf{z}_n)$ which means setting $q(\mathbf{z}_n)$ to $p(\mathbf{z}_n|\mathbf{v}_n)$, where the responsibilities are evaluated using the current parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and and mixing coefficients $\pi_k = p(z_{nk} = 1)$,
- maximizing with respect to the parameters of $p_k(\mathbf{v})$, $k = 1, \ldots, K$ and mixing coefficients π_k.

Expectation-Maximization



Expectation-Maximization

1. Choose K and convergence threshold.
2. Initialize all parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and mixing coefficients $\pi_k = p(z_k = 1)$.
3. *E-step*: Evaluate the responsibilities $p(z_{nk} = 1|\mathbf{v}_n)$ for $n = 1, \ldots, N$ and $k = 1, \ldots, K$.
4. *M-step*: Maximize

$$\sum_{k=1}^{K} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{v}_n) \log \Big(p(z_{nk} = 1) p(\mathbf{v}_n | z_{nk} = 1) \Big)$$

   with respect to the parameters of $p_k(\mathbf{v}) = p(\mathbf{v}|z_k = 1)$ and mixing coefficients $\pi_k = p(z_k = 1)$.
5. Terminate, if all changes are below the convergence threshold. Otherwise return to step 3

Distributions of parameters

Prior assumptions:

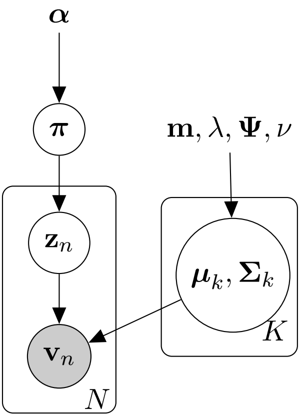
- K remains fixed.
- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ follows a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha/K, \ldots, \alpha/K)^T$.
- Probability distributions p_k are drawn themselves from a probability distribution over distributions, known as *base distribution* G_0.

Distributions of parameters

If each process is a normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, these can be drawn from the *normal inverse Wishart distribution*, $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathrm{NIW}(\mathbf{m}, \lambda, \boldsymbol{\Psi}, \nu)$, with four parameters:

- the *location vector* $\mathbf{m}$ lying in the feature space,
- the *mean fraction* λ,
- the *inverse scale matrix* $\boldsymbol{\Psi}$, which has to be symmetric and positive definite,
- and ν, which has to be at least the number of dimensions d of the feature space and regulates the degrees of freedom.

Distributions of parameters



Indefinite Number of Clusters

- Fixing the number of clusters is undesirable.
- A new sample can be generated from existing processes or a completely new process.
- *Base distribution*: A distribution from which the distributions of the processes are generated.
- *Dispersion, concentration, scaling parameter or strength*: α.
- Probability that the second sample is generated by the first process: $1/(1+\alpha)$.
- Probability that it is generated by a new process: $\alpha/(1+\alpha)$.
- If $\alpha = 1$, both are one half.
- If $\alpha > 1$, then a new process is favoured.
- If $\alpha < 1$, the existing process is more likely to generate it.

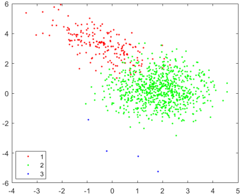
Indefinite Number of Clusters

- The n^{th} sample is generated by

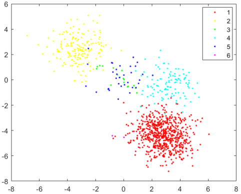
$$\begin{cases} \text{process } k \text{ with probability} & \dfrac{n_k}{n-1+\alpha} \\ \text{a new process with probability} & \dfrac{\alpha}{n-1+\alpha} \end{cases},$$

  where n_k is the number of samples generated by process k so far.
- When summing the samples generated by each process over all processes, the result is $n-1$. Therefore the probabilities sum to 1.
- Note that as more and more samples are generated by a particular process, it gets more likely that this process will generate further samples, since $n_k/(n-1+\alpha)$ increases. This is known as *rich-get-richer*.

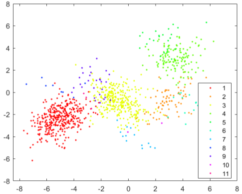
Indefinite Number of Clusters



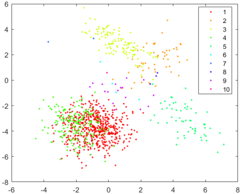
$\alpha = 0.5.$



$\alpha = 1.$



$\alpha = 2.$



$\alpha = 2.$

Chinese Restaurant Process (CRP)