STAT 403

4/21/25

Lecture 6

Thanks for sitting in the first 6 rows

Participation
⇓
attendance

Poll method 1: TBPosted

HW2 Monday TBPosted

MMP OH: PDL
Mon 4-5pm B321

Lecture II – Clustering – Part II: Non-parametric clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

CSE 547/STAT 548
Winter 2022

1. Paradigms for clustering ✓
2. Methods based on non-parametric density estimation ✓
3. Model-based: Dirichlet process mixture models ←

$x_1, x_2, \dots$ not iid

$f \in \{N(\mu,\Sigma), \dots\}$

$\{f(\,;\theta),\ \theta \in \Theta\}$

$$f(x) = \sum_{k=1}^{K} \pi_k f(x;\theta_k)$$

mixture proportions

$k = 1:K$ finite mixture (parametric)

K is r.v. in $\{1,2,\dots\}$

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \pi_k \geq 0 \text{ for all } k$$

The Dirichlet distribution

P on $S = \{1, .. r\}$

$\theta_j = P_r[j]$

$j = 1:r$

- $Z \in \{1 : r\}$ a discrete random variable, let $\theta_j = P_z(j)$, $j = 1, \dots r$.
- **Multinomial distribution** Probability of i.i.d. sample of size N from P_z

conjugate

$z_1, .. z_n \sim$ iid P

$$P(z^{1,\dots n}) = \prod_{j=1}^{r} \theta_j^{n_j}$$

$n_j = \#\{z_i = j\}$

where $n_j = \#$the value j is observed, $j = 1, \dots r$

likelihood $\ell(\mathcal{D})$

- $n_{1:r}$ are the **sufficient statistics** of the data.
- The **Dirichlet distribution** is defined over domain of $\theta_{1,\dots r}$, with real parameters $N'_{1,\dots r} > 0$ by

Bayesian

Prior : $$D(\theta_{1,\dots r}; n'_{1,\dots r}) = \frac{\Gamma(\sum_j n'_j)}{\prod_j \Gamma(n'_j)} \prod_j \theta_j^{n'_j - 1}$$

where $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt$.

Posterior : $P(\theta_{1:r} | \mathcal{D}) \propto \ell(\mathcal{D}) \cdot$ Prior

$\propto \prod_j \theta_j^{n_j} \theta_j^{n'_j - 1} = \prod_{j=1}^{r} \theta_j^{\boxed{n_j + n'_j} - 1}$

$\tilde{n}_j$

$r = 3$, $n = 5$

counts

$\mathcal{D}: z_{1:n} = 1\ 1\ 3\ 1\ 2 \rightarrow$ $n_1 = 3$, $n_2 = n_3 = 1$

$$D(\theta, n') = \frac{\Gamma(n')}{\prod_{j=1}^{r}\Gamma(n'_j)} \prod_{j=1}^{r} \theta_j^{n'_j - 1}$$

$\theta = (\theta_1, \cdots \theta_r)$

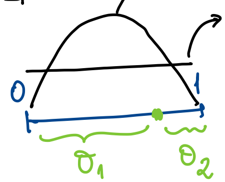
$n' = (n'_1, \cdots n'_r)$ ← hyperparameters

← distribution over Probability simplex

$$\underline{\underline{\Delta_r}} = \{\theta \,;\, \theta_j \geq 0 \text{ for all } j,\ \sum_{j=1}^{r} \theta_j = 1\}$$

$D(\,, n'_1, n'_2)$

$D(\,; 1, 1)$



$z \in \{1, 2\}$

$r = 2$

$\theta_1, \theta_2 \geq 0$

$\theta_1 + \theta_2 = 1$

Ex $r = 2$ $\quad n'_1 = 3$ $\quad \Rightarrow D(\ldots) = \dfrac{4!}{2!\,1!}\, \theta_1^{3-1}\, \theta_2^{2-1}$

$\quad n'_2 = 2$

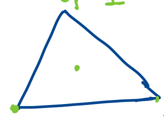
$\quad \overline{n' = 5}$

$$\boxed{\int_{\Delta_r} \prod_{j=1}^{r} \theta_j^{n'_j - 1}\, d\theta_1 d\theta_2 \cdots d\theta_r = \frac{\prod_{j=1}^{r}\Gamma(n'_j)}{\Gamma(n')}}$$

$r = 3$

$\theta_{1,2,3} \geq 0$

$\theta_1 + \theta_2 + \theta_3 = 1$



↑

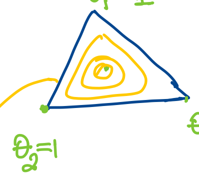
$D(\,; n'_{1,2,3})$

$$n' = \sum_{j=1}^{r} n'_j$$

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t}\, dt$$

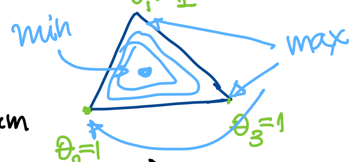
$\Gamma(p+1) = p!$ $\quad \Rightarrow \quad \Gamma(5) = 4!$

$p = 1, 2, 3, \cdots$



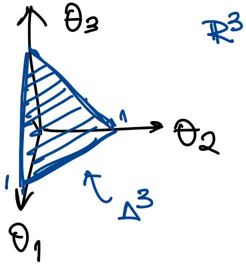
$n'_{1:r} = 1 \Rightarrow D = \text{uniform}$

$n'_1 = n'_2 = \cdots > 1$



↑

$\underline{n'_1 = n'_2 = \cdots \in (0, 1)}$



$$\begin{cases} \theta_1 + \theta_2 + \theta_3 = 1 \\ \theta_{1,2,3} \geq 0 \end{cases}$$

Dirichlet process mixtures

- Model-based
- generalization of mixture models to
  - infinite K
  - Bayesian framework
- denote θ_k = parameters for component f_k
- assume $f_k(x) \equiv f(x, \theta_k) \in \{f(x, \theta)\}$
- assume prior distributions for parameters $g_0(\theta)$
- prior with hyperparameter $\alpha > 0$ on the number of clusters
- very flexible model

A sampling model for the data

Dirichlet process

α = hyperparam

- **Example: Gaussian mixtures**, $d = 1$, $\sigma_k = \sigma$ fixed
- $\theta = \mu$
- prior for μ is $Normal(0, \sigma_0^2 I_d)$
- Sampling process
  - for $i = 1 : n$ sample $x_i, k(i)$ as follows

    denote $\{1 : K\}$ the clusters after step $i - 1$
    define n_k the size of cluster k after step $i - 1$

    1. $$k(i) = \begin{cases} k & \text{w.p } \frac{n_k}{i-1+\alpha}, \quad k = 1 : K \\ K+1 & \text{w.p } \frac{\alpha}{i-1+\alpha} \end{cases} \tag{1}$$

       size of cluster k

       $\frac{\alpha}{i+\alpha-1}$
    2. if $k(i) = K + 1$ sample $\mu_i \equiv \mu_{K+1}$ from $Normal(0, \sigma_0^2)$
    3. sample x_i from $Normal(\mu_{k(i)}, \sigma^2)$
- can be shown that the distribution of $x_{1:n}$ is **interchangeable** (does not depend on data permutation)

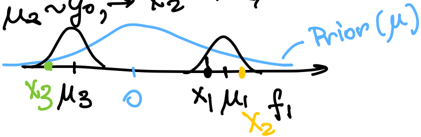
$n=1$: $\mu_1 \sim g_0$, $x_1 \sim N(\mu_1, \sigma^2)$

$\mu_2 = \mu_1$

$n=2$: $\nearrow x_2 \sim N(\mu_1, \sigma^2)$

$\searrow \mu_2 \sim g_0, \rightarrow x_2 \rightarrow N(\mu_2, \sigma^2)$

$n=3$:



$n=4$:

$f_k \in \{ N(\mu, \sigma^2) \}$

$f_k = N(\mu_k, \sigma^2)$

$\mu \sim N(0, \sigma_0^2) = g_0$

A sampling model for the data

Chinese restaurant

at step i:

C_1 μ_1, μ_2, $\cdots$ μ_{i-1} C_2

x_1 x_2 x_{i-1}

$\mu_1 = \mu_2$

$n(C_1) = n_1 = 2$

$n(C_2) = 1$

- **Example: Gaussian mixtures**, $d = 1$, $\sigma_k = \sigma$ fixed
- $\theta = \mu$
- prior for μ is $Normal(0, \sigma_0^2 I_d)$
- Sampling process
  - for $i = 1 : n$ sample x_i, $k(i)$ as follows

    denote $\{1 : K\}$ the clusters after step $i-1$
    define n_k the size of cluster k after step $i-1$

    1.
$$k(i) = \begin{cases} k & \text{w.p } \frac{n_k}{i-1+\alpha}, \quad k = 1 : K \\ K+1 & \text{w.p } \frac{\alpha}{i-1+\alpha} \end{cases} \tag{1}$$

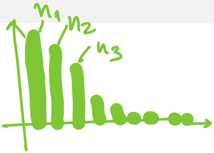
$$\sum_{k=1}^{K+1} \cdots = 1$$

    2. if $k(i) = K+1$ sample $\mu_i \equiv \mu_{K+1}$ from $Normal(0, \sigma_0^2)$
    3. sample x_i from $Normal(\mu_{k(i)}, \sigma^2)$
- can be shown that the distribution of $x_{1:n}$ is **interchangeable** (does not depend on data permutation)

$\sum n_k = i-1$

- $K \nearrow$ with n (slowly)
- $\frac{\alpha}{n-1+\alpha} \to$ with n
- K at n $\nearrow$ with α

The hyperparameters



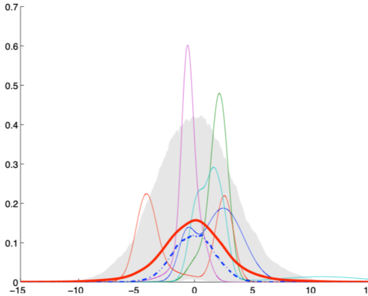
- σ_0 controls spread of centers
  - should be large

  g_0
- α controls number of cluster centers
  - α large $\Rightarrow$ many clusters

  smoothness param.
- cluster sizes non-uniform (larger clusters attract more new points)
- many single point clusters possible

General Dirichlet mixture model

- cluster densities $\{f(x, \theta)\}$
- parameters θ sampled from prior $g_0(\theta, \beta)$
- cluster membership $k(i)$ sampled as in (1)
- x_i sampled from $f(x, \theta_{k(i)})$
- **Model Hyperparameters** α, β



Clustering with Dirichlet mixtures ← fitting model to data

The clustering problem

- $\alpha, g_0, \beta, \{f\}$ given model class
- $\mathcal{D}$ given
- wanted $\theta_{1:n}$ (not all distinct!) $\mu_{1:K}$
- note:
  - $\theta_{1:n}$ determines a hard clustering Δ $k(i)$ = cluster assignment for x_i
  - the posterior of $\theta_{1:n}$ given the data determines a soft clustering via $P(x_i \mid k) \propto \int f(x_i|\theta_k)g_k(\theta_k)d\theta_k$

Estimating $\theta_{1:n}$ cannot be solved in closed form
Usually solved by MCMC (Markov Chain Monte Carlo) sampling

Clustering with Dirichlet mixtures via MCMC (alternate

Gibbs sampling
$k(i) \sim$
$\theta_k \sim$)

MCMC estimation for Dirichlet mixture

Input $\alpha, g_0, \beta, \{f\}, \mathcal{D}$

State cluster assignments $k(i)$, $i = 1 : n$, ← random
parameters θ_k for all distinct k ← $(\mu_i = x_i)$

Iterate

1. for $i = 1 : n$ (reassign data to clusters)
   1. remove i from its cluster (hence $\sum_k n_k = n - 1$)
   2. resample $k(i)$ by
   $$k(i) = \begin{cases} \text{existing } k & \text{w.p} \propto \frac{n_k}{n-1+\alpha} f(x_i, \theta_k) \\ \text{new cluster} & \text{w.p } \frac{\alpha}{n-1+\alpha} \int f(x_i, \theta) g_0(\theta) d\theta \end{cases} \quad (2)$$
   3. if $k(i)$ is new label, sample a new $\theta_{k(i)} \propto g_0 f(x_i, \theta)$
2. for $k \in \{k(1:n)\}$ (resample cluster parameters)
   1. sample θ_k from posterior $g_k(\theta) \propto g_0(\theta, \beta) \prod_{i \in C_k} f(x_i, \theta)$

   g_k can be computed in closed form if g_0 is conjugate prior

Output a state with high posterior