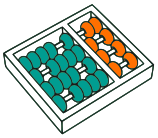

UNICAMP

Clustering

Lesson No. 13

Patrick de Carvalho Tavares Rezende Ferreira - 175480

1 Introduction

We can define clustering as the process of grouping similar objects, with two types of inputs that we can use. If we have a dissimilarity matrix of dimension NxN, or matrix of distances D, we will be in the case of similarity based clustering. We are in the case of feature based clustering if a design X matrix or NxD features matrix is the input of the algorithm. The advantages of each method are: similarity based allows the inclusion of domain-specific similarities or kernel functions, while feature based allows to treat raw data with noise. We also have flat and hierarchical clustering as possible types of output. The first where we classify objects into separate sets, and the second where we create a nested partition tree. Flat clustering has the advantage of being faster, while hierarchical clustering is more useful.

Most hierarchical algorithms are deterministic and do not require specifying the number of clusters, while flat algorithms require modeling of this quantity and are sensitive to the initial condition. Finally, these types of algorithms have approaches that can accelerate the inference in probalistic models.

We can define a dissimilarity matrix as a matrix with the main diagonal elements being zero and the rest being non-negative. It provides us with a measure of the distance between objects in each dimension of the matrix. Some algorithms require that the dissimilarity matrix (D) be a matrix of true distances. If we have a similarity matrix (S), we can obtain D by applying a monotonically decreasing function, as $D = max(S) - S$. Equation 1 shows the commom way of defining attribute based dissimilarity between objects.

$$\Delta(x_i, x_{i'}) = \sum_{j=1}^{D} \Delta_j(x_{ij}, x_{i'j}) \tag{1}$$

Purity of a cluster is defined as $p_i \stackrel{\Delta}{=} max_j p_{ij}$, where $p_{ij} = N_{ij}/N_i$, what is the empirical distribution over class labels for i-th cluster, and N_{ij} is the number of objects from class j in cluster i. We may calculate the total number of objects in cluster i by $N_i = \sum_{j=1}^{C} N_{ij}$.

2 Dirichlet process mixture models

Sometimes called model-based clustering, the simplest method for flat clustering is to use a finite mixture model, since we optimize a well-defined objective and define a probabilistic model for the data. Our challenge here consists of choosing the number of components K, but many times there is no well defined number of clusters.

We are going to a method that does not impose any a priori bound on K, a infinite mixture models, using non-parametric prior based on the Dirichlet process (DP). This way, the number of clusters is allowed to grow with the amount of data increase.

The usual form for a finite mixure model is shown in equations 2, 3 and 4.

$$p(x_i|z_i = k, \theta) = p(x_i|\theta_k) \tag{2}$$

$$p(z_i = k|\pi) = \pi_k \tag{3}$$

$$p(\pi|\alpha) = Dir(\pi|(\alpha/K)1_K) \tag{4}$$

Then we rewrite $p(x_i|\theta_k)$ as $x_i \sim F(\theta_{zi})$, being F the observation distribution and $p(\theta_k|\lambda)$ chosen as the conjugate to $p(x_i|\theta_k)$. In the same way, we may rewrite $\theta_k \sim H(\lambda)$, with H as the prior.

We have $\overline{\theta}_i$ as the parameter generating observation x_i, samples from G distribution, shown in equation 5.

$$G(\theta) = \sum_{k=1}^{K} \pi_k \delta_{\theta_k}(\theta) \tag{5}$$

In equation 5, $\pi \sim Dir(\frac{\alpha}{K}1)$, and $\theta_k \sim H$. Therefore, G is a finite mixture of delta functions with center on the cluster parameters θ_k. The prior probability for the cluster is shown in equation 6.

$$p(\overline{\theta}_i = \theta_k) = \pi_k \tag{6}$$

We will always get K cluster if sample from this model. with scattered around the clusters centers. It is desired a more flexible model for generating a variable number of clusters, as it becomes more likely to have a new visible cluster as we generate more data. We do this by replacing G distribution with a random probability measure.

It is well known tha a gaussian process is a distribution over functions with form $f : \chi \to \mathbb{R}$. Also we have defined that $p(f(x_1), ..., f(x_N))$ be jointly gaussian, within any set of points $x_i \in \chi$. This Gaussian can have its parameters computed by a mean function $\mu()$ and covariance/kernel function K(). A Dirichlet process consists of a distribution over probabilitymeasures $G : \theta \to \mathbb{R}^+$, where we require $G(\theta) \geq 0$ and $\int_\theta G(\theta) d\theta = 1$. Dp is implicitly defined by the requirement that $(G(T_1), ..., G(T_k))$ has a joint Dirichlet distribution such as equation 7.

$$Dir(\alpha H(T_1), ..., \alpha H(T_k)), \tag{7}$$

where $(T_1, ..., T_k)$ is a finite partition of θ. We may write $G \sim DP(\alpha, H)$, where α is called concentration parameter and H is the base measure. The marginals in each cell has probability beta distributed, as equation 8.

$$Beta(\alpha H(t_i), \alpha \sum_{j \neq i} H(T_j)) \tag{8}$$

If $\pi \sim Dir(\alpha)$, and $z|\pi \sim Cat(\pi)$, then we can integrate π to obtain hte predictive distribution for dirichlet-multinoulli model, as equation 9.

$$z \sim Cat(\alpha_1/\alpha_0, ..., \alpha_K/\alpha_0), \tag{9}$$

for $\alpha_0 = \sum_k \alpha_k$. We can also calculate the posterior updated for π given one observation given by equation 10.

$$\pi|z \sim Dir(\alpha_1 + \mathbb{I}(z = 1), ..., \alpha_k + \mathbb{I}(z = K)) \tag{10}$$

Dirichlet process use this as arbitrary partitions. If $G \sim DP(\alpha, H)$, then $p(\theta \in T_i) = H(T_i)$ and the posterior is given by equation 11.

$$p(G(T_1), ..., G(T_k)|\theta, \alpha, H) = Dir(\alpha H(T_1) + \mathbb{I}(\theta \in T_1), ..., \alpha H(T_k) + (\theta \in T_\theta)) \tag{11}$$

Equation 11 is valid for any set of partitions. If observing multiple samples $\overline{\theta}_i \sim G$, the new posterior is expressed in equation 12.

$$G\overline{\theta}_1, ..., \overline{\theta}_N, \alpha, H \sim DP(\alpha + N, \frac{1}{\alpha + N}(\alpha H + \sum_{i=1}^{N} \delta_{\theta_i})) \tag{12}$$

Therefore, a conjugate prior is defined by DP for arbitrary measurable spaces, and the concentration parameter α is like the effective sample size for H base measure.

2.1 Stick breaking

The definition for DP shown bellow is known as stick breaking construction. Given $\pi = \pi_{k_{k=1}}^{\infty}$ as an infinite sequence of mixture weights from equations 13 and 14.

$$B_k \sim Beta(1, \alpha) \tag{13}$$

$$\pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_1) = \beta_k (1 - \sum_{l=1}^{k-1} \pi_l), \tag{14}$$

and equation 15 is sometimes an expression for 14.

$$\pi \sim GEM(\alpha) \tag{15}$$

This process will terminate with probability 1, even with the number of elements incresing with α. Moreover, π_k components' size decreases on average. Then we define equation 16.

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \tag{16}$$

where $\theta_k \sim H$. Hence we see that samples from DP are discrete with probability one, if we keep sampling, we will see repeated. Data samples from $\overline{\theta}_i$ will cluster around the θ_k. It is an indicator that DP might be useful for this purpose.

2.2 Chinese restaurant

We can avoid problems of working with sticks if exploiting the clustering propoerty to draw sample from a GP. The result is that, if $\overline{\theta}_i \sim G$ are N observations from $G \sim DP(\alpha, H)$, taking on K values of θ_k, our predictive distribution comes to equation 17.

$$p(\overline{\theta}_{N+1} = \theta | \overline{\theta}_{1:N}, \alpha, H) = \frac{1}{\alpha + N}(\alpha H(\theta) + \sum_{k=1}^{K} N_k \delta_{\theta_k}(\theta)), \tag{17}$$

with the number of previous observations, $N_k = \theta_k$. As it is more convenient to work with discret variables z_i, which specify a value of θ_k to use. Then, we get to equation 18.

$$p(z_{N+1} = \theta | z_{1:N}, \alpha) = \frac{1}{\alpha + N}(\alpha \mathbb{I}(z = k^*) + \sum_{k=1}^{K} N_k \mathbb{I}(z = k)), \tag{18}$$

where k^* representing a new cluster not used yet. This processed is known as chinese restaurant [1] process. The fact that currently occupied tables are more likely to get customers is known as rich get richer and one can derive an expression for the distribution of cluster sizes by prior process. The number os occupied tables K approaches $\alpha log(N)$ as $N \to \infty$, indicating that the model complexity grows logarithmically.

2.3 Fitting a DP

Here we state how to fit a DP mixture model. First, we have equation 19.

$$p(z_i = k | z_{-i}, x, \alpha, \lambda) \propto p(z_i = k | z_{-i}, \alpha) p(x_i | x_{-i}, z_i = k, z_{-i}, \lambda) \tag{19}$$

With 19, it is possible to assume that z_i is the last customer entering the restaurant. Equation 20 gives us the fit term.

$$p(z_i | z_{-i}, \alpha) = \frac{1}{\alpha + N - 1}(\alpha \mathbb{I}(z_i = k^*) + \sum_{k=1}^{K} N_{k_j - i} \mathbb{I}(z_i = k)), \tag{20}$$

as k^* being a new cluster and K being the number of clusters by z_{-i}. Rewriting this, we have 21.

$$p(z_i = k | z_{-i}, \alpha) = \begin{cases} \frac{N_{k_j - i}}{\alpha + N - 1} & \text{if k has been seen before} \\ \frac{\alpha}{\alpha + N - 1} & \text{if k is a new cluster} \end{cases} \tag{21}$$

In order to compute the second term of 19, we are going to partition data x_{-i} into clusters based on z_{-i}. Since equation 22 represents the data assigned to cluster c, x_i is conditionally independent of data points not assigned to cluster k. We then arrive to equation 22.

$$x_{-i,c} = \{x_j : z_j = c, j \neq i\} \tag{22}$$

$$p(x_i | x_{-i}, z_{-i}, z_i = k, \lambda) = p(x_i | x_{-i,k}, \lambda) = \frac{p(x_i, x_{-i,k}|, \lambda)}{p(x_{-i,k} | \lambda)}) \tag{23}$$

With 22 representing the posterior predictive distribution for cluster k evaluated at x_i. We then arrive to equation 24, for $z_i = k^*$ corresponding to a new cluster. This represents the prior predictive distribution for a new cluster evaluated at x_i.

$$p(x_i|x_{-i}, z_{-i}, z_i = k^*, \lambda) = p(x_i|\lambda) = \int p(x_i|\theta) H(\theta|\lambda) d\theta \tag{24}$$

2.4 Spectral clustering

Spectral clustering is a technique based on roots using graph theory, since our approach used to identify clusters of nodes is in a graph based on the edges connecting them. It is flexible, allowing us to cluster non graph data. Spectral clustering uses eigenvalues of special matrices from the data set. For finding partitions A_k in K cluster, we have equation 25 as minimize criteria.

$$cut(A_1, ..., A_k) \triangleq \frac{1}{2} \sum_{k=1}^{K} W(A_k, \overline{A}_k) \tag{25}$$

The optimal solution usually only partitions a single data point from the rest, and we are going to ensure a minimum large partition defining a normalized cut by equation 26.

$$Ncut(A_1, ..., A_k) \triangleq \frac{1}{2} \sum_{k=1}^{K} \frac{cut(A_k, \overline{A}_k)}{vol(A_k)} \tag{26}$$

We then have the graph splitted into K clusters in a way nodes within each cluster are similar to each other and different to other clusters. We can formulate this problem in terms of searching for binary vectors, although this is NP problem.

2.5 Hierarchical clustering

Hierarchical cluster analysis is a clustering algorithm that creates clusters that have predominant ordering top to bottom. It groups similar objects into clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Mixture models produce a flat clustering, while hierarchical clustering can be nested in each other.

Main approaches to hierarchical clustering are bottom up and top down. In the first, most similar groups are merged at each step, whereas top-down approach splits groups using different criteria. Both methods are heuristics, dont optimizing any specific function. It makes hard to stablish the quality of their clustering, besides they will always produce a clustering of the input data, even if the data has no structure at all

References

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.