CS6462

Probabilistic and Explainable AI

Lesson 16

Bayesian Nonparametric Models

*

Dirichlet Process Models for Clustering

by Emil Vassev

February 24, 2024

Clustering with Mixture Models

Clustering:

- grouping similar data into clusters
- data mining technique – data analytics for marketing, security, or sciences
- problem – we don’t know the number of clusters to be discovered
- Dirichlet Process Mixture (DPM) - automatically detects the number of clusters

Mixture Modeling in ML:

- machine learning algorithms used to classify data into different categories based on probability distribution

Clustering with Mixture Models:

- traditional mixture modeling approach to clustering - requires the number of clusters to be specified in advance of analyzing the data
- Bayesian nonparametric approach – estimates the number of needed clusters automatically and allows more clusters to be discovered if needed

Clustering with Mixture Models (cont.)

Bayesian nonparametric generalization of mixture models:

- estimates the number of components in a mixture model
- estimates the parameters of the individual mixture components

Finite Mixture Models:

- probability density function

$$P_X(x) = \sum_{k=1}^{K} \pi_k p(x|\Theta_k)$$

- X – random variable vector
- K – the number of components in the mixture model
- Θ_k – set of parameters associated with component k
- π_k – mixing proportion of the k component: the probability that observation x belongs to the k component of the mixture model

Clustering with Mixture Models (cont.)

Finite Mixture Models:

- probability density function: sum of the conditional probability that the observation x belongs to the k component considering Θ_k

$$P_X(x) = \int p(x|\Theta) * G(\Theta) * d\Theta$$

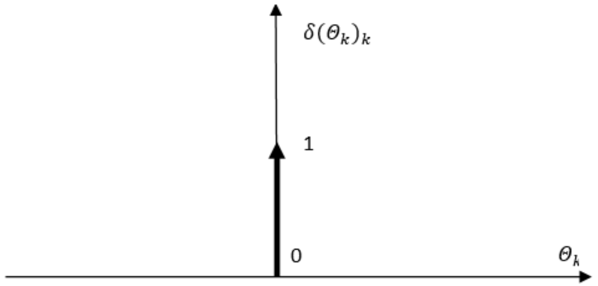
$$\boxed{P_X(x) = \sum_{k=1}^{K} \pi_k p(x|\Theta_k)}$$

$$G(\Theta) = \sum_{k=1}^{K} \pi_k \delta_{\Theta_k} \Longrightarrow G(\Theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\Theta_k}$$

Bayesian nonparametric mixtures use mixing distributions consisting of a countably infinite number of atoms.

- $G(\Theta)$ a discrete mixing distribution encapsulating all the parameters Θ of the mixture model
- δ_{Θ_k} – is *a dirac delta distribution** (atom) centered at Θ_k (set of parameters associated with component k)

$$\int_{-\infty}^{\infty} \delta(\Theta_k) * d\Theta_k = 1, \delta(\Theta_k) = 0 \text{ if } \Theta_k \neq 0$$



* Dirac Distribution at http://nlab-pages.s3.us-east-2.amazonaws.com/nlab/show/Dirac+distribution

Dirichlet Process Mixture Models

Dynamism of Mixture Models:

- applied to a finite training set - only a finite (but varying) number of components will be used to model the data
- one component k is associated with many data items X (random variables), but a data item x is associated with one component k only
- inference in the model automatically provides the number of components to use and the parameters of the components

$$\boxed{G(\Theta) = \sum_{k=1}^{K} \pi_k \delta_{\Theta_k}}$$

Dirichlet Process Mixture:

- Dirichlet Process (DP) is a probability distribution over distributions

$G(\Theta) \sim DP(\alpha, G_0(\Theta))$ where α - positive scaling parameter, G_0- base distribution

- Bayesian approach requires a ***prior*** distribution over the mixing distribution $G(\Theta)$
- Dirichlet process (DP) - the most common ***prior*** distribution to use

Dirichlet Process Mixture Models (cont.)

Formal definition of DP:

- we observe a sample $X = \{X_1, X_2, \dots, X_\mathrm{n}\}$ from a mixture of distributions controlled by the DP considering the parameters Θ:

$G(\Theta) \sim DP(\alpha, G_0(\Theta))$

- $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$ – finite partition of the parameter space (K clusters)
- G_0 is a **prior** distribution over distributions (probability measures) $G(\Theta)$
- $G(\Theta) = \{G(\theta_1), G(\theta_2), \dots, G(\theta_k)\}$ – induced random vector of discrete mixing distributions
- $DP\big(\alpha * G_0(\theta_1), \alpha * G_0(\theta_2), \dots, \alpha * G_0(\theta_k)\big)$ Dirichlet process with parameters

$\{G(\theta_1), G(\theta_2), \dots, G(\theta_k)\} = DP\big(\alpha * G_0(\theta_1), \alpha * G_0(\theta_2), \dots, \alpha * G_0(\theta_k)\big)$

Chinese Restaurant Process:

- DP induces a distribution over partitions of integers that describes the **prior**

Summary

Bayesian Nonparametric Models – Dirichlet Process Models for Clustering

Clustering

Mixture Modeling in ML

Clustering with Mixture Models

- *Bayesian nonparametric generalization of mixture models*
- *Finite Mixture Models*

$$P_X(x) = \int p(x|\Theta) * G(\Theta) * d\Theta$$

Dirichlet Process Mixture

$$G(\Theta) \sim DP(\alpha, G_0(\Theta))$$

Next Lesson:

- Bayesian Networks - Theoretical Foundations of Bayesian Networks

Thank You!

Questions?