Hyperspherical Clustering

RAMÓN DANIEL REGUEIRO ESPIÑO
Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga)
Memorias das Bolsas EXPLORA
Titor/a: Jose Ameijeiras Alonso

Abstract

In this work, a review was made on how to extend clustering techniques to data with support in the hypersphere, a special emphasis was placed on the use of mixtures of von Mises-Fisher distributions. In addition, this procedure was applied to a real case of text mining in order to group in topics the summaries of the oral communications of the XV SGAPEIO Congress.

1 Introduction

Clustering techniques are a very useful and deeply studied tool for classifying data where belonging to each group is unknown. Probably the most studied case is that in which the data considered have their support in an interval of the real line or in a product of intervals of this line. Thus, a more unknown case is that in which we change the support, as it may be that this is the hypersphere, the extension of the sphere to larger dimensions. In this work we have studied the application of these techniques to this case.

As a new type of media is introduced, the measures used previously may not be appropriate for it. This statement will motivate new forms of parameterization and the use of the von Mises-Fisher distribution, which will play the role equivalent to the normal distribution on the real line. This will be summarized in Section 2. In Section 3, we will look at the EM algorithm, which attempts to adjust the parameters of a mixture in order to maximize the likelihood. In Section 4, we will apply the blending algorithm to a practical example of text mining where our initial data will be summaries of the oral communications of the XV SGAPEIO Congress. Finally, in Section 5, we will introduce brief conclusions about the overall work done.

2 Section 2: Modeling in the Sphere

We will begin this section by illustrating with an example why we cannot use the usual mean in the hypersphere. This will result in the use of other parameterizations, such as angular parameterization. Next, we will define the von Mises-Fisher distribution, which we will use later in Section 3.

Supposing that we have a data set in a unit sphere, if we have at least two different data, then the resulting average will be a norm vector strictly lower than unity, that is, it will not belong to the support of our data. Thus, we can consider its average length as its norm. From this average, and if the average length is not zero, applying a normalization we obtain a vector in the sphere that we will call average direction, and we will denote by μ.

A possible parameterization is the angular one:

$$x = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi),$$

where $\phi \in [0, 2\pi]$ e $\theta \in [0, \pi]$ with the aim to avoid possible different values, except for 0 and 2π, for the same point.

By fixing a direction vector μ, we can consider a new distribution, the von Mises-Fisher distribution, which will play a role equivalent to the normal one in the case of the real line. This distribution has as density function:

$$f(x; \mu, \kappa) = \frac{\kappa^{\frac{3}{2}-1}\, e^{\kappa\mu^T x}}{(2\pi)^{\frac{3}{2}} I_{\frac{3}{2}-1}(\kappa)}.$$

In this case, if we consider μ as our mean and κ as a *concentration parameter*, which will have an inverse role to variance in the case of the normal distribution. It should be noted that we can consider the case where $\kappa = 0$, which would result in the constant distribution along the hypersphere. In order to illustrate the effect of this parameter, in Figure 1 we can see how the distributions vary as the value of the concentration parameter changes.

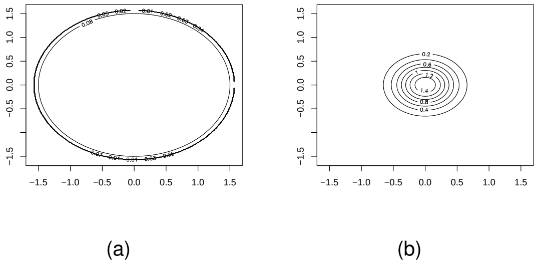

Figure 1: *Von Mises-Fisher distribution contour graph with κ taking as values 0.1 and 10 respectively.*

3 Section 3: EM Algorithm

Although there are different techniques for grouping in the hypersphere, we will focus on obtaining density mixtures by fixing the number of von Mises-Fisher distributions using the EM algorithm.

This algorithm consists of a first step of initialization, that usually is to assign to all the densities equal weights and the rest of parameters so that they are two to two different ones. Subsequently, two steps are applied, expectation (E) and maximization (M), those that give the algorithm its name, until a maximum of iterations is reached or until the improvement in probability is less than that considered as a convergence criterion.

Step E consists of estimating with what proportion each observation belongs to each of the different distributions considered in the mixture if we consider the parameters obtained in the previous iteration, or the initials of being the first.

Step M, from the one obtained in the previous step, estimates new values ??of the parameters that maximize the likelihood, equivalently its logarithm. In this case, the calculation of the concentration parameters requires the solution of an implicit equation, so approximations of this are used, as can be seen in Hornik and Grün (2014).

4 Section 4: Application

Let us now illustrate how we can apply the algorithm to text mining. In this case, based on the oral communications of the XV SGAPEIO Congress, we will try to make a separation into eight different groups, as was done in reality.

First, it is necessary to transform the texts into numerical data to which the algorithm can be applied. For this, we consider each text as an observation made up of a set of words, *tokens*. Next, we remove those words whose meaning is not relevant, such as articles. Finally, we count the frequency of each word and thus obtain a vector formed by the frequencies of each word of the set of words in the documents. If we normalize the vectors, we transform the initial texts into unit vectors of the dimension hypersphere the number of words with different meanings, having two vectors will be closer when they have similar frequencies of occurrence of each word.

Carrion	Armijos-Toro	Clarence-Safari	Ameijeiras-Alonso
Neira-Rueda	Barbera	Ghosh	Davila-Pena
Porcel-Mari	Iglesias-Patiño	Lopez-Cheda	Freijeiro-Gonzalez
.	Rodriguez-Barreiro	Panduro-Martin	Gonzalez-Maestro
.	.	Pelaez	Lopez-Perez
.	.	.	.
.	.	.	.
.	.	.	.

Alonso-deVelasco	Arias-Lopez	Gonzalez-Diaz	Lopez-Oriona
Fanjul-Hevia	Borrajo	Lopez-Vizcaino	Saavedra-Nieves
Ginzo-Villamayor	Conde-Amboage	Molina	.
Lado-Baleato	Diaz-Louzao	Novo-Perez2	.
Marques-deSousa	Garcia-Portugues	Saavedra-Nieves(A)	.
Santiago-Perez	Novo-Perez	.	.
Teijeiro-Campo	Teodoro	.	.
.	Vicente-Gonzalez	.	.

Table 1: *Resulting distribution obtained by von Mises-Fisher mixtures with $k = 8$.*

The different groups obtained in the case of oral communications can be seen in Table 1. It should be noted that only two groups coincide with those made in the congress. These groupes are the quality control group and the game theory and optimization groupe.

5 Section 5: Conclusions

The work done, and summarized in this paper, allows us to draw several conclusions about grouping into hyperspheres. In this case we will highlight two.

On the one hand, it is a field with different lines in which to continue researching, such as whether there is a way to harmonize those words without content or to develop another grouping technique that replaces it. On the other hand, its application is diverse and is a very useful tool for conducting, among others, text mining, such as the thematic classification of abstracts in a conference.

References

Banerjee, A., Dhillon, I., Ghosh, J. and Sra, S. (2005) Clustering on the Unit Hypersphere using von Mises-Fisher Distributions *Journal of Machine Learning Research* 46 (6).

Celeux, G. and Govaert, G. (1991) A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* 14 (3).

Hornik, K. e Grün, B. (2014) movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions. *Journal of Statistical Software* 58 (10).

Mardia, K.V. and Jupp, P.E. (2000) *Directional Statistics*. John Wiley & Sons.