Determining the number of clusters in a data set

From Wikipedia, the free encyclopedia

Determining the number of clusters in a data set, a quantity often labeled k as in the k-means algorithm, is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem.

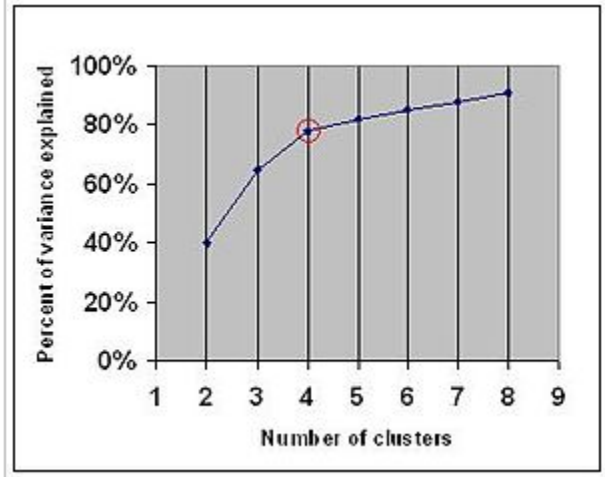
For a certain class of clustering algorithms (in particular k-means, k-medoids and expectation–maximization algorithm), there is a parameter commonly referred to as k that specifies the number of clusters to detect. Other algorithms such as DBSCAN and OPTICS algorithm do not require the specification of this parameter; hierarchical clustering avoids the problem altogether.

The correct choice of k is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing k without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered its own cluster (i.e., when k equals the number of data points, n). Intuitively then, *the optimal choice of* k *will strike a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster*. If an appropriate value of k is not apparent from prior knowledge of the properties of the data set, it must be chosen somehow. There are several categories of methods for making this decision.

Contents

The elbow method

The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.[1] Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance.[2]



Explained Variance. The "elbow" is indicated by the red circle. The number of clusters chosen should therefore be 4.

The method can be traced to speculation by Robert L. Thorndike in 1953.[3]

X-means clustering

In statistics and data mining, **X-means clustering** is a variation of k-means clustering that refines cluster assignments by repeatedly attempting subdivision, and keeping the best resulting splits, until some criterion is reached.[4] The Bayesian information criterion is used to make the splitting decision.[5]

Information criterion approach

Another set of methods for determining the number of clusters are information criteria, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), or the Deviance information criterion (DIC) — if it is possible to make a likelihood function for the clustering model. For example: The *k*-means model is "almost" a Gaussian mixture model and one can construct a likelihood for the Gaussian mixture model and thus also determine information criterion values.[6]

An information–theoretic approach

Rate distortion theory has been applied to choosing *k* called the "jump" method, which determines the number of clusters that maximizes efficiency while minimizing error by information-theoretic standards.[7] The strategy of the algorithm is to generate a distortion curve for the input data by running a standard clustering algorithm such as k-means for all values of *k* between 1 and *n*, and computing the distortion (described below) of the resulting clustering. The distortion curve is then transformed by a negative power chosen based on the dimensionality of the data. Jumps in the resulting values then signify reasonable choices for *k*, with the largest jump representing the best choice.

The distortion of a clustering of some input data is formally defined as follows: Let the data set be modeled as a *p*-dimensional random variable, *X*, consisting of a mixture distribution of *G* components with common covariance, Γ. If we let $c_1 \dots c_K$ be a set of *K* cluster centers, with c_X the closest center to a given sample of *X*, then the minimum average distortion per dimension when fitting the *K* centers to the data is:

$$d_K = \frac{1}{p} \min_{c_1 \dots c_K} E[(X - c_X)^T \Gamma^{-1} (X - c_X)]$$

This is also the average Mahalanobis distance per dimension between *X* and the set of cluster centers *C*. Because the minimization over all possible sets of cluster centers is prohibitively complex, the distortion is computed in practice by generating a set of cluster centers using a standard clustering algorithm and computing the distortion using the result. The pseudo-code for the jump method with an input set of *p*-dimensional data points *X* is:

```
JumpMethod(X):
    Let Y = (p/2)
    Init a list D, of size n+1
    Let D[0] = 0
    For k = 1 ... n:
        Cluster X with k clusters (e.g., with k-means)
        Let d = Distortion of the resulting clustering
        D[k] = d^(-Y)
    Define J(i) = D[i] - D[i-1]
    Return the k between 1 and n that maximizes J(k)
```

The choice of the transform power $Y = (p/2)$ is motivated by asymptotic reasoning using results from rate distortion theory. Let the data *X* have a single, arbitrarily *p*-dimensional Gaussian distribution, and let fixed $K = \lfloor \alpha^p \rfloor$, for some *α* greater than zero. Then the distortion of a clustering of *K* clusters in the limit as *p* goes to infinity is α^{-2}. It can be seen that asymptotically, the distortion of a clustering to the power $(-p/2)$ is proportional to α^p, which by definition is approximately the number of clusters *K*. In other words, for a single Gaussian distribution, increasing *K* beyond the true number of clusters, which should be

one, causes a linear growth in distortion. This behavior is important in the general case of a mixture of multiple distribution components.

Let X be a mixture of G p-dimensional Gaussian distributions with common covariance. Then for any fixed K less than G, the distortion of a clustering as p goes to infinity is infinite. Intuitively, this means that a clustering of less than the correct number of clusters is unable to describe asymptotically high-dimensional data, causing the distortion to increase without limit. If, as described above, K is made an increasing function of p, namely, $K = \lfloor \alpha^p \rfloor$, the same result as above is achieved, with the value of the distortion in the limit as p goes to infinity being equal to α^{-2}. Correspondingly, there is the same proportional relationship between the transformed distortion and the number of clusters, K.

Putting the results above together, it can be seen that for sufficiently high values of p, the transformed distortion $d_K^{-p/2}$ is approximately zero for $K < G$, then jumps suddenly and begins increasing linearly for $K \geq G$. The jump algorithm for choosing K makes use of these behaviors to identify the most likely value for the true number of clusters.

Although the mathematical support for the method is given in terms of asymptotic results, the algorithm has been empirically verified to work well in a variety of data sets with reasonable dimensionality. In addition to the localized jump method described above, there exists a second algorithm for choosing K using the same transformed distortion values known as the broken line method. The broken line method identifies the jump point in the graph of the transformed distortion by doing a simple least squares error line fit of two line segments, which in theory will fall along the x-axis for $K < G$, and along the linearly increasing phase of the transformed distortion plot for $K \geq G$. The broken line method is more robust than the jump method in that its decision is global rather than local, but it also relies on the assumption of Gaussian mixture components, whereas the jump method is fully non-parametric and has been shown to be viable for general mixture distributions.

The silhouette method

The average silhouette of the data is another useful criterion for assessing the natural number of clusters. The silhouette of a data instance is a measure of how closely it is matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster, i.e. the cluster whose average distance from the datum is lowest.[8] A silhouette close to 1 implies the datum is in an appropriate cluster, while a silhouette close to −1 implies the datum is in the wrong cluster. Optimization techniques such as genetic algorithms are useful in determining the number of clusters that gives rise to the largest silhouette.[9] It is also possible to re-scale the data in such a way that the silhouette is more likely to be maximised

at the correct number of clusters.[10]

Cross-validation

One can also use the process of cross-validation to analyze the number of clusters. In this process, the data is partitioned into v parts. Each of the parts is then set aside at turn as a test set, a clustering model computed on the other $v - 1$ training sets, and the value of the objective function (for example, the sum of the squared distances to the centroids for k-means) calculated for the test set. These v values are calculated and averaged for each alternative number of clusters, and the cluster number selected such that further increase in number of clusters leads to only a small reduction in the objective function. [11]

Finding number of clusters in text databases

In text databases, a document collection defined by a document by term D matrix (of size m by n, m: number of documents, n: number of terms) number of clusters can roughly be estimated by the following formula $\frac{mn}{t}$ where t is the number of non-zero entries in D. Note that in D each row and each column must contain at least one non-zero element. [12]

Analyzing the kernel matrix

Kernel matrix defines the proximity of the input information. For example, in Gaussian Radial basis function, determines the dot product of the inputs in a higher-dimensional space, called feature space. It is believed that the data become more linearly separable in the feature space, and hence, linear algorithms can be applied on the data with a higher success.

The kernel matrix can thus be analyzed in order to find the optimal number of clusters .[13] The method proceeds by the eigenvalue decomposition of the kernel matrix. It will then analyze the eigenvalues and eigenvectors to obtain a measure of the compactness of the input distribution. Finally, a plot will be drawn, where the elbow of that plot indicates the optimal number of clusters in the data set. Unlike previous methods, this technique does not need to perform any clustering a-priori. It directly finds the number of clusters from the data.

External links

- Clustergram – cluster diagnostic plot (http://www.r-statistics.com/2010/06/clustergram-visualization-and-diagnostics-for-cluster-analysis-r-code/) – for visual diagnostics of choosing the number of (k) clusters (R code)

- Eight methods for determining an optimal k value for *k*-means analysis (https://stackoverflow.com/a/15376462/1036500) – Answer on stackoverflow containing R code for several methods of computing an optimal value of k for *k*-means cluster analysis

Bibliography

1. See, e.g., David J. Ketchen Jr; Christopher L. Shook (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique". *Strategic Management Journal*. **17** (6): 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G.
2. See, e.g., Figure 6 in
   - Cyril Goutte, Peter Toft, Egill Rostrup, Finn Årup Nielsen, Lars Kai Hansen (March 1999). "On Clustering fMRI Time Series". *NeuroImage*. **9** (3): 298–310. doi:10.1006/nimg.1998.0391. PMID 10075900.
3. Robert L. Thorndike (December 1953). "Who Belongs in the Family?". *Psychometrika*. **18** (4): 267–276. doi:10.1007/BF02289263.
4. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" (PDF). Retrieved 2016-08-16.
5. "The X-Alter algorithm : a parameter-free method to perform unsupervised clustering" (PDF). Retrieved 2016-08-16.
6. Cyril Goutte, Lars Kai Hansen, Matthew G. Liptrot & Egill Rostrup (2001). "Feature-Space Clustering for fMRI Meta-Analysis". *Human Brain Mapping*. **13** (3): 165–183. doi:10.1002/hbm.1031. PMID 11376501. see especially Figure 14 and appendix.
7. Catherine A. Sugar; Gareth M. James (2003). "Finding the number of clusters in a data set: An information-theoretic approach". *Journal of the American Statistical Association*. **98** (January): 750–763. doi:10.1198/016214503000000666.
8. Peter J. Rousseuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*. **20**: 53–65. doi:10.1016/0377-0427(87)90125-7.
9. R. Lleti; M.C. Ortiz; L.A. Sarabia; M.S. Sánchez (2004). "Selecting Variables for *k*-Means Cluster Analysis by Using a Genetic Algorithm that Optimises the Silhouettes". *Analytica Chimica Acta*. **515**: 87–100. doi:10.1016/j.aca.2003.12.020.
10. R.C. de Amorim & C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". *Information Sciences*. **324**: 126–145. doi:10.1016/j.ins.2015.06.039.
11. See e.g. "Finding the Right Number of Clusters in k-Means and EM Clustering: v-Fold Cross-Validation". *Electronic Statistics Textbook*. StatSoft. 2010. Retrieved 2010-05-03.
12. Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". *ACM Transactions on Database Systems*. **15** (4): 483. doi:10.1145/99935.99938. especially see Section 2.7.
13. Honarkhah, M; Caers, J (2010). "Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling". *Mathematical Geosciences*. **42** (5): 487–517. doi:10.1007/s11004-010-9276-7.

- Ralf Wagner, Sören W. Scholz, Reinhold Decker (2005): The Number of

Clusters in Market Segmentation, in: Daniel Baier, Reinhold Decker; Lars Schmidt-Thieme (Eds.): Data Analysis and Decision Support, Berlin, Springer, 157–176.

Categories: Cluster analysis | Clustering criteria
