Next Up Previous Contents Index

Next: Model-based clustering **Up:** K-means **Previous:** K-means **Contents** **Index**

Cluster cardinality in K-means

We stated in Section 16.2 that the number of clusters K is an input to most flat clustering algorithms. What do we do if we cannot come up with a plausible guess for K?

A naive approach would be to select the optimal value of K according to the objective function, namely the value of K that minimizes RSS. Defining $\mathrm{RSS}_{min}(K)$ as the minimal RSS of all clusterings with K clusters, we observe that $\mathrm{RSS}_{min}(K)$ is a monotonically decreasing function in K (Exercise 16.7), which reaches its minimum 0 for $K=N$ where N is the number of documents. We would end up with each document being in its own cluster. Clearly, this is not an optimal clustering.

A heuristic method that gets around this problem is to estimate $\mathrm{RSS}_{min}(K)$ as follows. We first perform i (e.g., $i=10$) clusterings with K clusters (each with a different initialization) and compute the RSS of each. Then we take the minimum of the i RSS values. We denote this minimum by $\widehat{\mathrm{RSS}}_{min}(K)$. Now we can inspect the values $\widehat{\mathrm{RSS}}_{min}(K)$ as K increases and find the ``knee'' in the curve - the point where successive decreases in $\widehat{\mathrm{RSS}}_{min}$ become noticeably smaller. There are two such points in Figure 16.8 , one at $K=4$, where the gradient flattens slightly, and a clearer flattening at $K=9$. This is typical: there is seldom a single best number of clusters. We still need to employ an external constraint to choose from a number of possible values of K (4 and 9 in this case).

A second type of criterion for cluster cardinality imposes a penalty for each new cluster - where conceptually we start with a single cluster containing all documents and then search for the optimal number of clusters K by successively incrementing K by one. To determine the cluster cardinality in this way, we create a generalized objective function that combines two elements: *distortion* , a measure of how much documents deviate from the prototype of their clusters (e.g., RSS for K-means); and a measure of *model complexity* . We interpret a clustering here as a model of the data. Model complexity in clustering is usually the number of clusters or a function thereof. For K-means, we then get this selection criterion for K:

$$K = \arg\min_K [\mathrm{RSS}_{min}(K) + \lambda K] \tag{195}$$

where λ is a weighting factor. A large value of λ favors solutions with few clusters. For $\lambda = 0$, there is no penalty for more clusters and $K = N$ is the best solution.

The obvious difficulty with Equation 195 is that we need to determine λ. Unless this is easier than determining K directly, then we are back to square one. In some cases, we can choose values of λ that have worked well for similar data sets in the past. For example, if we periodically cluster news stories from a newswire, there is likely to be a fixed value of λ that gives us the right K in each successive clustering. In this application, we would not be able to determine K based on past experience since K changes.

A theoretical justification for Equation 195 is the *Akaike Information Criterion* or AIC, an information-theoretic measure that trades off distortion against model complexity. The general form of AIC is:

$$\textbf{AIC:} \qquad K = \arg\min_K [-2L(K) + 2q(K)] \tag{196}$$

where $-L(K)$, the negative maximum log-likelihood of the data for K clusters, is a measure of distortion and $q(K)$, the number of parameters of a model with K clusters, is a measure of model complexity. We will not attempt to derive the AIC here, but it is easy to understand intuitively. The first property of a good model of the data is that each data point is modeled well by the model. This is the goal of low distortion. But models should also be small (i.e., have low model complexity) since a model that merely describes the data (and therefore has zero distortion) is worthless. AIC provides a theoretical justification for one particular way of weighting these two factors, distortion and model complexity, when selecting a model.

For K-means, the AIC can be stated as follows:

$$\textbf{AIC:} \qquad K = \arg\min_K [\mathrm{RSS}_{min}(K) + 2MK] \tag{197}$$

Equation 197 is a special case of Equation 195 for $\lambda = 2M$.

To derive Equation 197 from Equation 196 observe that $q(K) = KM$ in K-means since each element of the K centroids is a parameter that can be varied independently; and that $L(K) = -(1/2)\mathrm{RSS}_{min}(K)$

(modulo a constant) if we view the model underlying K-means as a Gaussian mixture with hard assignment, uniform cluster priors and identical spherical covariance matrices (see Exercise 16.7).

The derivation of AIC is based on a number of assumptions, e.g., that the data are . These assumptions are only approximately true for data sets in information retrieval. As a consequence, the AIC can rarely be applied without modification in text clustering. In Figure 16.8 , the dimensionality of the vector space is $M \approx$ 50,000. Thus, $2MK > 50{,}000$ dominates the smaller RSS-based term ($\widehat{\mathrm{RSS}}_{min}(1) < 5000$, not shown in the figure) and the minimum of the expression is reached for $K=1$. But as we know, $K=4$ (corresponding to the four classes China, Germany, Russia and Sports) is a better choice than $K=1$. In practice, Equation 195 is often more useful than Equation 197 - with the caveat that we need to come up with an estimate for λ.

Exercises.

- Why are documents that do not use the same term for the concept car likely to end up in the same cluster in K-means clustering?
- Two of the possible termination conditions for K-means were (1) assignment does not change, (2) centroids do not change (page 16.4). Do these two conditions imply each other?

Next Up Previous Contents Index

Next: Model-based clustering **Up:** K-means **Previous:** K-means **Contents** **Index**

This is an automatically generated page. In case of formatting errors you may want to look at the PDF edition of the book.
2009-04-07