Estimating the Mean Number of K-Means Clusters to Form

Robert A. Murphy, Ph.D.

e-mail: robert.a.murphy@wustl.edu

Abstract: Utilizing the sample size of a dataset, the random cluster model is employed in order to derive an estimate of the mean number of K-Means clusters to form during classification of a dataset.

Keywords and phrases: probability, random cluster, K-means, machine learning, classification.

1. Introduction and Related Work

In traditional \(K\)-Means classification, a set of data is assumed to have formed correlated clusters of data points, if \(K\) points are randomly chosen beforehand to represent classes and each of the remaining data points are assigned to the class containing the closest class representative. A survey \cite{29} of different approaches for specifying an appropriate choice of \(K\) typically results in one of eight methods being used. Values of \(K\) are chosen from a set or range \([1, 3, 4, 5]\) and \([8, 10, 15, 16, 26, 28]\), specified by a user \cite{21}, determined in a later processing step via a closed form equation or approximation technique \cite{16}, made equivalent to the number of generators of a synthetic set of test data \cite{5}, determined via statistical measures \cite{5, 14, 17, 19, 27, 30}, made equivalent to the number of classes which form after performing another clustering preprocessing step \cite{6, 22}, approximated through visual inspection of plotted data points \cite{3, 6}, applying a neighborhood measure \cite{22}.

Each of the previously-mentioned techniques has its advantages and drawbacks. As such, in this hybrid, \(K\)-means data classifier, several of the aforementioned methods are employed to find that a set of data will form clusters, if the measured distances between data points (or some common point in each class) are below a certain threshold. With the assumption that the data points are randomly generated throughout some bounded region according to a certain probability distribution, we identify an upper bound on a chosen distance measure of regularity such that clusters of data points form when their measured distances from a common weighted-center is less than the upper bound.

With the choice of a distance measure of regularity, the data points are assumed to be generated into a bounded subset \(B' \subset \mathbb{R}^L\). Let \(B \subset \mathbb{R}^2\) be a bounded region of unit area. Using graph-based methods, we can define a bijective projection mapping of the data points from \(B'\) into \(B\) by partitioning \(B\) into a fixed number of structures of the same size and shape such that disjoint class regularities are maintained, and each data point in \(B'\) maps to exactly one structure in the partition of \(B\). From this partition, we can calculate the maximum size of each structure such that disjoint class regularities are maintained after projection.

A finite, independent sampling from the normal distribution is assumed, which is then ordered by applying some suitable order statistic, producing a dependent, beta-distributed sampling \cite{23}.

Therefore, given the fixed size of the region \(B\), the ordered sample is projected into \(B\), while...
maintaining disjoint class regularities, if the amount of separation between classes is predetermined and the number of classes is known.

2. Setup

2.1. Embedding and Distance Threshold

In [25], it is proven in corollary (7.4.39) that an ordered set of data, which is assumed to be spatially uniformly distributed, will form clusters, i.e. classes, if the measured distances between data points (or some common data point in each class) are below a certain threshold, which is computed as a function of the number of data points sampled from the total population of data. Now, most datasets would be too computationally expensive to order. So, we can get around this limitation by making certain assumptions.

We can make the assumption that the total population of data is normally distributed by making use of the Central Limit Theorem [18]. In essence, this theorem states that any set of randomly distributed data with finite mean and variance will tend to be normally distributed as the sample size grows large. We assume a statistically representative sample of the total population of data according to the normal distribution, as represented by a sequence of independent, identically distributed random variables, \( X_1, X_2, \ldots, X_J \). An order statistic is applied to the these random variables so that \( X_{k_1} < X_{k_2} < \ldots < X_{k_J} \), where \( \sigma(i) = k_i \) for \( 1 \leq i \leq J \) is a permutation resulting in the ordering of \( \{X_i\}_{i=1}^J \). Note that, \( \{X_{k_i}\}_{i=1}^J \) is a sequence of dependent random variables, since for each \( i \in [2, J] \), the random variable \( X_{k_i} \) depends upon \( X_{k_j} \) for all \( j < i \). Now, \( \{X_{k_i}\}_{i=1}^J \) is a (less computationally expensive) ordering of a sample which was drawn from a statistically representative sample of the entire population.

To make an assumption of uniformity of the ordered, dependent sample, we first note that the Beta distribution is the probability distribution of an order statistic of normally distributed random variables [23]. We can take the Beta distribution shape parameters to be \( \alpha = 1 = \beta \), since we injectively map each data point to exactly 1 partition of a uniformly partitioned, 2-dimensional, bounded region, while maintaining class relationships between data points.

Let \( T^2 \) be the total number of sampled data points, with each data point having \( L \) attributes. The attributes of each of the \( T^2 \) are normalized to have values in the interval [0, 1]. What we now have are \( T^2 \) data points, each with \( L \) attributes, which are represented by \( T^2 \) points in an \( L \)-dimensional hypercube of unit area. Let \( J = M^2 \) and assume that we have a smaller dataset of size \( M^2 \), which is further sampled from the \( T^2 \) data set. The value \( K = N^2 \) will be estimated later, as a function of \( M \).

2.2. Partition of the Bounded Region

Let \( B \) be the unit square in the 2-dimensional plane. Theorem (7.3.38) from [25] gives the minimum number of hexagons required to partition \( B \) into hexagons of equal size such that there are \( K = N^2 \) disjoint partitions totaling \( J = M^2 \) hexagons.

**Theorem 1** Assume that there are \( J = M^2 \) samples and \( K = N^2 \) classifications for the samples. The minimum number of hexagons required to partition the unit square into \( K = N^2 \) disjoint regions such that \( J = M^2 \) is the sum total of all hexagons in the disjoint regions is given by

\[
S(M, N) = M^2 + (N - 1)^2 + 2MN.
\]
In the proof of theorem 1, it is shown that the amount of separation between classes is defined by the row and column of unoccupied hexagons of length $M/N$ which surrounds each of the disjoint classes, with data points in each class being correlated to points in hexagons sharing a common side and to those points at the end of a path of correlated points. The idea is to use the result of the theorem to calculate, as a function of $M$ and $N = N(M)$, an estimate of the size of a prototypical hexagon which will be used to partition $B$. The size of the prototypical hexagon is directly determined by the size of its smallest circumscribing circle.

### 2.3. Upper Bound on the Distance Measure

Assume that the representation of each data point in $B$ is a node in a graphical network. In [12], it is stated and proven that there is a critical probability of hexagons containing a node such that it is no longer possible to have disjoint clusters of nodes when this critical probability is exceeded. Since the size of $B$ is fixed, then to decrease the probability while maintaining $K = N^2$ disjoint contiguous clusters of nodes, the size of each hexagon must decrease while increasing the number of hexagons in the boundaries of the disjoint regions. In this way, the ratio of the total number of occupied hexagons to the total number of hexagons will be less than this critical probability. Note that we used uniformity of the nodes throughout $B$ so that the approximate number of nodes in a cluster of hexagons is proportional to the ratio of the number of hexagons in the cluster divided by the number of hexagons in the entire region, $B$. Also, note that the minimum number of hexagons required for separation is given by theorem (1), so that the common radius of the circle that can inscribe any one of these hexagons is of size

$$R(M, N) = \frac{1}{2\sqrt{S(M, N)}},$$

thereby necessarily indicating that

$$B(M, N) = 2 \ast R(M, N)$$

is the diameter of the circle.

**Lemma 2** $R(M, N)$ is decreasing for increasing $M$ and $N$.

**Proof** By eq. (1), $S(M, N)$ is increasing for increasing $M$ and $N$. Consequently, by eq. (2), $R(M, N)$ is decreasing for increasing $M$ and $N$.

**Theorem 3** Denote the critical probability as $p_c$. With probability 1, every node in $B$ is contained within one contiguous cluster if and only if the probability of any one hexagon containing a node, exceeds $p_c$. Otherwise, all clusters are disjoint with probability 1.

In order to not exceed the critical probability, while maintaining the $K = N^2$ classes of $J = M^2$ data points, the common radial size $r_0$ of each circle must be less than $R(M, N)$. By thm. (3), the clusters will be disjoint with probability 1. The following corollary to thm. (1) follows from these statements and lemma (5).

**Corollary 4** Let $H$ be a hexagon of size such that a circle of radius $r_0 = r_0(M, N)$ can be circumscribed, where

$$r_0 \leq R(M, N).$$
If $B$ is partitioned into copies of $H$, then with probability 1, it follows that $K = N^2$ is the mean number of disjoint clusters of contiguous hexagons in the region $B$ that are occupied by the $J = M^2$ data points.

With $r_0$ given by corollary (4), the size of the prototypical hexagon can be calculated for re-partitioning $B$ through each classification of the $T^2$ data set. Furthermore, corollary (4) guarantees that the classes will remain distinct, with probability 1, through each new classification.

**Lemma 5** For a fixed number $(J = M^2)$ of uniformly distributed data points in $B$ and for any $\rho \in (0, p_c]$, with $p_c = 1 - 2 \sin(\pi/18)$,

$$\frac{M^2}{S(M,N)} = \frac{M^2}{M^2 + (N-1)^2 + 2MN} = \rho$$

(3)

determines the expected number $K = N^2$ of disjoint classes to form such that $J = M^2$ is the total of all occupied hexagons across all classes.

**Proof** At the risk of ambiguity, let $N^2$ denote both the random variable and the expectation of the random variable which takes the number of formed classes as its value. Because $B$ is partitioned by hexagons, it is shown in [11, Chapter 3] that $p_c = 1 - 2 \sin(\pi/18)$. By uniformity, the mean number of data points in each class is $M^2/N^2$. By theorem (3), each class will be disjoint and each hexagon in $B$ will be as large as possible if $B$ is partitioned into $S(M,N)$ hexagons of equal size. Also, by theorem (3), the probability of any of the $M^2$ hexagons being populated with a data point has to be less than or equal to $p_c$ in order that the expected classes form with probability 1, resulting in eq. (3). For any $\rho \in (0, p_c]$, $K = N^2$ is found by solving eq. (3) to obtain $K = N^2$ as the least integer which is not less than the integer part of a non-negative solution to eq. (3), for fixed, positive $M^2$.

3. Conclusions

By deriving an upper bound on a distance measure between points in a bounded region or some common point such that when the measured distances fall below the derived threshold, it is seen that segregated classes of correlated points form with high probability. By estimating an integer solution to an algebraic equation (3), the mean number $K = N^2$ of classes is obtained, which can be used in the equation of the upper bound (2) on the distance measure during $K$-means classification of the $J = M^2$ data points into approximately $K$ classes.
References

[1] Al-Daoud, M.B.; Venkateswarlu, N.B.; Roberts, S.A. (1996), *New Methods for the Initialisa- tion of Clusters*, Pattern Recognition Letters, 17, pp. 451-455, 1996
[2] Alpaydin, E. (2010), *Introduction to Machine Learning*, Second Edition, The MIT Press.
[3] Bilmes, J.; Vahdat, A.; Hsu, W.; Im, E.J. (1996), *Empirical Observations of Probabilistic Heuristics for the Clustering Problem*, Technical Report, TR-97-018, International Computer Science Institute, Berkeley, California, 1997
[4] Battou, L.; Bengio, Y. (1995), *Convergence Properties of the K-Means Algorithm*, Adv. Neural Inf. Processing Systems, 7, pp. 585-592, 1995
[5] Bradley, S.; Fayyad, U.M. (1998), *Refining Initial Points for K-Means Clustering*, Proceedings of the Fifteenth International Conference on Machine Learning (ICML ’98) (Ed. J. Shavlik), Madison, Wisconsin, pp. 91-99, 1998
[6] Cai, Z. (2001), *Technical Aspects of Data Mining*, PhD Thesis, Cardiff, University, 2001
[7] Carroll, D.E.; Goel, A. (2004), *Lower Bounds for Embedding into Distributions over Excluded Minor Graph Families*, Lecture Notes in Computer Science, Volume 3221, pp. 146-156, 2004
[8] Castro, V.E.; Yang, J. (2000), *A Fast and Robust General Purpose Clustering Algorithm*, Proceedings of the Fourth European Workshop on Principles of Knowledge Discovery in Databases and Data Mining (PKDD 00), Lyon, France, pp. 208-218, 2000
[9] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977), *Maximum Likelihood for Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, Series B (Methodological), Volume 39, No. 1., pp. 1 - 38, 1977
[10] Fritzke, B. (1997), *The LBG-U Method for Vector Quantization - An Improvement Over LBG Inspired by Neural Networks*, Neural Processing Lett., 5 (1), pp. 35-45, 1997
[11] Grimmett, Geoffrey (1999), *Percolation*, Springer-Verlag.
[12] Grimmett, Geoffrey (2006), *The Random-Cluster Model*, Springer-Verlag.
[13] Guyon, Xavier. (1995), *Random Fields on a Network: Modeling, Statistics and Applications*, Springer.
[14] Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. (2002), *Cluster Validity Methods. Part I. SIGMOD Record*, http://www.acm.org/sigmod/record/, Volume 31 (2), 2002
[15] Han, J.; Kamber, M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, California.
[16] Hansen, L.K.; Larsen, J. (1996), *Unsupervised Learning and Generalisation*, Proceedings of IEEE Conference on Neural Networks, Washington, DC, June 1996, pp. 25-30, 1996
[17] Hardy, A. (1996), *On the Number of Clusters*, Comput. Statist. Data Analysis, Volume 20, pp. 83-96, 1996
[18] Hogg, R.V.; McKean, J.W.; Craig, A.T. (2005), *Introduction to Mathematical Statistics*, Pearson Prentice Hall.
[19] Ishioka, T. (2000), *Extended K-Means with an Efficient Estimation of the Number of Clusters*, Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2000), Hong Kong, PR China, December 2000, pp. 17-22
[20] Kar, A. (2003), *Weyl’s Equidistribution Theorem*, Resonance, pp. 30 - 37, 2003
[21] Kerr, A.; Hall, H.K; Kolub, S. (2002), *Doing Statistics with SPSS*, Sage, London.
[22] Kothari, R.; Pitts, D. (1999), *On Finding the Number of Clusters*, Pattern Recognition Lett., Volume 20, pp. 405-416, 1999
[23] Mameli, V.; Musio, M. (2013), *A Generalization of the Skew-Normal Distribution: The Beta
Skew-Normal, Communications in Statistics - Theory and Methods, Volume 42, pp. 2229-2244, 2013

[24] Meester, Ronald; Roy, Rahul (1996), Continuum Percolation, Cambridge University Press.

[25] Murphy, Robert (2011), Partial Connectivity in Wireless Sensor Networks with Applications, UMI Proquest.

[26] Pelleg, D.; Moore, A. (1999), Accelerating Exact K-Means Algorithms with Geometric Reasoning, Proceedings of the Conference on Knowledge Discovery in Databases (KDD 99), San Diego, California, pp. 277-281, 1999

[27] Pelleg, D.; Moore, A. (1999), X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters, Proceedings of the 17th International Conference on Machine Learning (ICML 2000), Stanford, California, pp. 727-734, 2000

[28] Pena, J.; Lazano, J.A.; Larranaga, P. (1999), An Empirical Comparison of Four Initialisation Methods for the K-Means Algorithm, Pattern Recognition Lett., Volume 20, pp. 1027-1040, 1999

[29] Pham, D.T.; Dimov, S.S.; Nguyen, C.D. (2005), Selection of K in K-Means Clustering, Proc. IMechE: J. Mechanical Engineering Science, Volume 219, Part C, pp. 103-119, 2005

[30] Theoridis, S.; Koutroubas, K. (1998), Pattern Recognition, Academic Press, London.