



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Adaptive Sampling for High Throughput Data Using Similarity Measures

V. Bulaevskaya, A. P. Sales

May 8, 2015

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Adaptive Sampling for High Throughput Data Using Similarity Measures

Vera Bulaevskaya and Ana Paula Sales

September 30, 2013

Abstract

The need for adaptive sampling arises in the context of high throughput data because the rates of data arrival are many orders of magnitude larger than the rates at which they can be analyzed. A very fast decision must therefore be made regarding the value of each incoming observation and its inclusion in the analysis. In this report we discuss one approach to adaptive sampling, based on the new data point's similarity to the other data points being considered for inclusion. We present preliminary results for one real and one synthetic data set.

1 Introduction

In many modern-day problems, decisions are made based on enormous amounts of constantly streamed data. The rates at which these high throughput data arrive are so great that it is computationally impractical or even impossible to include every single incoming observation in the analysis. For example, in the context of cybersecurity and network monitoring, various network flow measurements are constantly updated and input into an algorithm that determines whether a threat or a problem is present in the system. Because of the overwhelming rates of data collection, only a fraction of the incoming observations can be processed by such an algorithm. Thus, a very fast decision must be made regarding inclusion or exclusion of each data point, creating a need for adaptive sampling. In this report, we discuss one possible approach to this, based on similarity measures, and present some preliminary results for one real and one synthetic data set.

The objective of adaptive sampling is to determine whether a given point should be discarded from training or updating the PF model in order to make the computational burden manageable. This decision clearly requires specifying a utility metric, or a measure of the value a given point will add to the quality of estimates of various quantities being modeled by the PF. All such utility metrics can be broadly divided into two groups. One group consists of metrics that are not specifically tied to the PF model in any way and can be used with any other modeling approach. We will refer to this type of metrics as “PF-agnostic”. This is in contrast to the utility metrics that are based on quantities within the PF model, or “PF-specific” metrics. In this report, we will focus on the PF-agnostic metrics, but will also briefly mention our current work on PF-specific metrics.

2 PF-Agnostic Metrics

For this type of metrics, we assume that we have a batch of observations from which we are to select a subset to use for training the PF. This affords us the possibility of judging any point’s utility relative to the other points in the batch. Note that this is in contrast to a fully streaming context in which a decision is made about a single observation on its arrival without reference to any other observations seen in the past or concurrently with the given observation.

All of the PF-agnostic metrics we consider are based on a given observation’s degree of similarity to the other observations in the batch. If an observation is very similar to many others in the batch, it may be redundant and may thus be a good candidate for being discarded. While clustering methods can be a highly accurate way to measure observations’ degree of similarity to one another, clustering algorithms are too computationally expensive to be used in settings where a decision about any given point must be made quickly, such as in our setting. We thus considered less ideal, but more tractable alternatives, the Euclidean distance and the cosine similarity measure.

2.1 Euclidean Distance

The Euclidean distance (ED) between two input vectors \mathbf{x}_1 and \mathbf{x}_2 is given by

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_i (x_{1i} - x_{2i})^2}. \quad (1)$$

The above definition is only valid for numerical inputs, so in the case of categorical inputs, these first have to be transformed to a numerical measurement. For example, in the case of email data described in Section 2.3.1, the inputs are text tokens encountered in an email. These can be transformed to frequencies of occurrences of each encountered token before computing a Euclidean distance between them.

Once EDs between a given point and all other points in the batch are computed, a summary statistic, e.g., the median or the minimum, of all such distances can be obtained. If one wants to remove a given number or fraction of points, one approach is to then simply discard the points with the lowest values of this statistic. In the presence of heavy clustering in the input space, however, this scheme can lead to a significant bias in the distribution of the output values in each cluster of the training set, i.e., the set of remaining points. This is because such a dropping scheme disproportionately favors points near the edges of each cluster. This bias in the training set will clearly result in biased predictions for the new data, making this approach problematic.

Therefore, instead of this deterministic approach, we considered a probabilistic modification of it. In particular, each point's distance summary statistic is first calculated and normalized to the range of all the points' values of this statistic in the dataset, resulting in a value between 0 and 1.

Let q denote a point's normalized distance summary statistic and f the desired fraction of points to be removed from the dataset. Assume that q is defined so that the probability of an observation being discarded is a decreasing function of q , such as in the case of the minimum distance (i.e., distance to a point's nearest neighbor). If the mean value of the statistic in the dataset $\bar{q} < f$, then the point is discarded with probability p given by

$$p = \begin{cases} 1 - cq & q \leq 1/c \\ 0 & q > 1/c \end{cases} \quad (2)$$

where c satisfies

$$(1 - cE(q \mid q \leq 1/c))P(q \leq 1/c) = f. \quad (3)$$

If, on the other hand, $\bar{q} \geq f$, each point is discarded with probability p given by

$$p = \begin{cases} c(1 - q) & q \geq 1 - 1/c \\ 1 & q < 1 - 1/c \end{cases} \quad (4)$$

where c satisfies

$$1 - (1 - c + cE(q \mid q \geq 1 - 1/c))P(q \geq 1 - 1/c) = f. \quad (5)$$

Note that the left-hand-sides of both (3) and (5) are just the expected values of p in (2) and (4), respectively. The parameter c in both cases is thus simply a tuning constant that ensures that the expected fraction of discarded points is equal to the target value f .

If the distance statistic q is chosen so that the probability of discarding an observation is an increasing function of the point's value of q , then the right-hand side of equations (2) and (4) is replaced by one minus itself.

2.2 Cosine Similarity

An alternative way to measure the degree of similarity between two vectors \mathbf{x}_1 and \mathbf{x}_2 is the cosine similarity (CS), given by

$$c(\mathbf{x}_1, \mathbf{x}_2) = \frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\|\mathbf{x}_1\| \times \|\mathbf{x}_2\|} = \frac{\sum_i x_{1i} \cdot x_{2i}}{\sqrt{\sum_i x_{1i}^2} \times \sqrt{\sum_i x_{2i}^2}}. \quad (6)$$

This measure is thus equal to the cosine of the angle between two inputs. As such, it is only applicable to non-scalar observations.

Unlike the Euclidean distance discussed above, this measure only reflects the relative orientation of the two vectors and does not take into account the magnitude of the individual vector components. Thus, two vectors that are multiples of each other will always have a cosine similarity equal to 1. In some applications, this feature may be a drawback.

On the other hand, unlike the Euclidean distance, cosine similarity is more efficient to compute for highly sparse inputs because only dimensions for which both vectors are non-zero are involved in the

calculation. This makes this measure attractive in applications, such as text mining and information retrieval [1].

As in the case of the Euclidean distance measure, we considered both a deterministic and a probabilistic approach to discarding points based on the cosine similarity measure.

2.3 Performance Study

To investigate the performance of the similarity measures discussed above as adaptive sampling approaches in the PF setting, we considered two examples of datasets with categorical outputs, which are described below.

2.3.1 Example 1: PU1 Corpus Dataset

The PU1 corpus dataset [2] consists of 1099 observations, where each observation corresponds to an email. The email classification (spam or legitimate, 481 and 618 observations, respectively) and the token words encountered in the email’s subject and body are recorded in each observation. The goal is to predict the email classification from the tokens in the subject and/or body of the email.

Since the similarity measures discussed above require numerical inputs, while the tokens in the subject and body of the email are categorical, these are first transformed to the frequency of each token’s occurrence. Note that the drawback of using this simple transformation is that the order of the tokens is not used in judging the degree of similarity. An alternative would be to use the frequency of n -grams, for example, in order to capture the token order, but this is much more burdensome computationally, so this option was not considered in this study.

The dataset was randomly partitioned into a training and validation set, each consisting of 990 and 109 observations, respectively. We considered the following adaptive schemes for discarding the observations:

1. ED used in a deterministic fashion (ED-D)
2. ED used in a probabilistic fashion (ED-P)
3. CS used in a deterministic fashion (CS-D)
4. CS used in a probabilistic fashion (CS-P)
5. K-means clustering of points (KM)

6. Spherical k-means clustering of points (SKM)

For the ED-P and the CS-P schemes, we used the minimum distance, or the point’s Euclidean distance to its nearest neighbor (NN), and the maximum similarity measure, respectively, as the summary statistic. In the KM scheme, the points are clustered using the k-means algorithm [3] and the points furthest from the cluster centroid are removed, with the number removed from each cluster equal to the target dropping fraction f of the given cluster’s size. The SKM scheme is exactly the same, but the clustering is performed using the spherical k-means algorithm [4]. The number of clusters for both clustering algorithms was chosen to maximize the average of the clusters’ silhouette metrics, which is a measure of how tightly grouped the data are within each cluster [5].

As alluded to earlier, the rationale for considering the KM and the SKM schemes is that we expect these two methods to provide an upper bound on the performance of similarity measures in terms of the resulting predictive accuracy. These can be ideal methods for judging points’ similarity to one another, but they are computationally too expensive to be used in contexts where a decision about any given point must be made quickly, as is the case in the high-throughput data setting. Thus, including them in this study allows us to benchmark the similarity methods against some of the best, but too costly, alternatives.

The training set was reduced by removing 25%, 50% or 75% of all the points, either randomly or adaptively using schemes 1–6 above. Using both the subject and body token frequencies to assess an email’s degree of similarity is more complicated than using just the subject or the body token frequencies alone, so we decided to use just one of the two as a starting point. Since a typical email has considerably more body than subject tokens, we chose using the body token frequencies to compute the similarity measures. However, the training of the PF on the full and reduced datasets was done using both subject and body tokens. The tokens and the email classification were all modeled using multinomial distribution in the PF implementation.

Thus, for each of the 30 partitions, the PF was trained on the full training set and the reduced training sets, obtained by dropping observations randomly, as well as according to each of the 6 adaptive schemes listed above, for each of the dropping fractions we considered. The trained PF model was then used to make predictions of the probability of an email being legitimate for each email in the valida-

tion set based on its subject and body tokens. A Receiver Operating Characteristic (ROC) curve can then be constructed to describe the probability of labeling a legitimate email as legitimate (true positive) as a function of declaring a spam email as legitimate (false positive). The area under this curve, known as the area under the curve (AUC), is one commonly used summary statistic of the ROC curve. Higher values of AUC indicate better average accuracy of predictions.

Figure 1 shows the values of the AUC for the 30 experiments described above. The first observation that can be made about the results shown in the figure is that in terms of the AUC values, random dropping of points (shown in red) performs as well or better than using the full training set (shown in grey). This indicates that this particular training set is not well suited for testing adaptive sampling since an ideal test case is one in which dropping the points at random would cause the prediction quality to deteriorate relative to using the full training set.

Nevertheless, a few features of these results are still worth pointing out. For example, while the ED-D (shown in green) is one of the best performers when the dropping fraction is 25% and 50%, its performance is very poor when the dropping fraction is at its high level of 75%, for the reasons described in Section 2.1. Indeed, the probabilistic counterpart of ED-D, ED-P (shown in blue), performs much better for 50% and 75% dropping fractions (and is the best performer for 75% dropping fraction, outperforming even the KM dropping scheme, shown in yellow), confirming the expectation that motivated the use of the ED-P metric. Thus, even though random sampling performed as well as using a full training set, the ED-P metric was an even better option in this case.

Also, the cosine similarity measures, CS-D (shown in cyan) and CS-P (shown in magenta), are both quite poor performers in this dataset although, unsurprisingly, when the dropping fraction is high, the probabilistic version, CS-P, performs better than its deterministic counterpart, CS-D, for the 75% dropping fraction. Finally, as expected, the KM scheme (shown in yellow) is one of the best performers although, as mentioned previously, it is outperformed by the ED-P scheme when the dropping fraction is at its highest level.

The AUC is just one statistic reflecting the quality of a PF-trained model. To make a more complete assessment of the benefit of adaptive sampling, one should compare the estimates of the joint and marginal token distributions that result from the full dataset and each of the re-

duced datasets (indeed, the AUC value is based on such an estimate). This comparison is likely to reveal more differences, particularly if the distribution of the full training set contains some subtle features that a reduced training dataset would be unlikely to capture unless it is carefully chosen through adaptive sampling. The capability to obtain detailed density estimates is currently being implemented, and this comparison will be completed in the near future.

2.3.2 Example 2: Gaussian Mixture Test Data

One situation in which random sampling can perform very poorly is when the data are heavily clustered, with some of the clusters containing relatively small fractions of the data. In this case, if the dropping fraction is large relative to the fraction of the data in the small cluster, dropping observations randomly can result in a reduced training set containing very few or even no points from the small cluster, thus heavily biasing the model. Adaptive sampling, on the other hand, is much more likely to capture the points in the small cluster.

To compare the performance of the adaptive sampling measure to that of random sampling, we designed a training dataset containing 800 observations. Each observation consists of two features. One is a mixture of two Gaussians, given by $0.97N(0.1, 1) + 0.03N(8, 1)$. The other feature is a categorical label, 1 or 2, and is completely determined by the cluster membership of the Gaussian feature. This dataset is shown in Figure 2. As can be seen from the plot, there are two very well separated clusters, with one significantly smaller (20 observations, or 2.5% of the data) in size than the other (780 observations, or 97.5% of the data), and the cluster membership is equal to the categorical label (one shown in red and the other in green).

The dataset was randomly divided into a training set, containing 600 observations, and a validation set, containing 200 observations. Then as in the case of the email data, the dataset was reduced using random and adaptive sampling schemes. Since there is only one continuous feature, we cannot use the cosine similarity measure or spherical k-means in this dataset, so we only considered ED-D, ED-P and KM methods. The dropping fractions we considered were 25%, 50%, 75%, 90% and 95%. The PF model was then trained on the full and each of the reduced datasets and the labels in the validation set were predicted using this model. This experiment was repeated 200

times.

Because of a very simple and clustered structure of this dataset, there are only 2 possible AUC values, 1 and 0.5. If there is at least one observation from each cluster in the training set, the prediction will be perfect, resulting in AUC value of 1. If, on the other hand, one cluster is completely left out of the training set, prediction for the observations in that cluster in the validation set will always be wrong by definition. This will result in the AUC value of 0.5, corresponding to a random guess classifier. Thus, random sampling is particularly risky in this setting when the dropping fraction is high relative to the percentage of points in the smaller cluster.

In fact, one can analytically compute the probability of not capturing any observations from the smaller cluster in a dataset reduced via random sampling, as the number of points from the smaller cluster in such a setting follows the Hypergeometric distribution. Specifically, for a given dropping fraction f , the probability of a random sample of size $600(1 - f)$ containing no observations from the smaller cluster is given by

$$\frac{\binom{585}{600(1-f)}}{\binom{600}{600(1-f)}} \quad (7)$$

since the training set contains $0.025 \times 600 = 15$ points from the smaller cluster and, accordingly, 585 points from the larger cluster (the training sets were drawn to preserve this distribution in each experiment). This probability is clearly an increasing function of f and for our chosen dropping fractions $f = 0.25, 0.5, 0.75, 0.9$ and 0.95 , this probability is equal to 0, 0, (both to 5 decimals) 0.01, 0.2 and 0.46, respectively.

These theoretical values are very close to what we observed in our simulations: using random sampling to reduce the dataset resulted in the AUC values of 0.5 in 0%, 0%, 0.5%, 16% and 43% of the 200 experiments for the 5 dropping fractions, respectively (AUC was 1 in the remainder of the experiments for each of the 5 fractions). All the other dropping schemes (and, of course, the full training set) resulted in the AUC value of 1 100% of the time because, as explained before, they all captured the data from the smaller cluster no matter how small the training set was. Table 1 summarizes these results in terms of the percentage of experiments that resulted in the AUC value of 1 for each of the schemes and dropping fractions.

This example is clearly not realistic in a sense that due to very heavy clustering, there is no loss in predictive ability when using a

Sampling scheme	Dropping fraction				
	25%	50%	75%	90%	95%
Random sampling	100	100	99.5	84	57
Full and AS schemes	100	100	100	100	100

Table 1: Percentage of 200 experiments that resulted in the AUC value of 1 for each of the sampling schemes and dropping fractions. The remaining experiments resulted in the AUC value of 0.5 for the reasons explained in the text.

reduced training set obtained with adaptive sampling schemes compared to the full training set. In real datasets, this will almost never be the case. However, the example does illustrate the benefit of adaptive sampling, using similarity measures in particular, over random sampling. High dropping fractions that lead to a significant deterioration in random sampling’s predictive ability may indeed correspond a realistic scenario in which the data arrive at an extremely high rate relative to the analysis rates one can reasonably attain.

2.4 Summary and Future Work

The two examples discussed above suggest that using certain similarity measures for adaptive sampling can be beneficial relative to random sampling and in some cases even relative to using the full dataset, as in the case of the email data. In particular, the probabilistic version of the Euclidean distance metric, ED-P, emerged as the best performer for higher dropping fractions in that example. The Gaussian test set example illustrated the risk associated with random sampling and high dropping fractions in cases of clustered data with very imbalanced cluster counts.

More datasets need to be considered to validate these findings and in particular, datasets in which random sampling leads to a deterioration in the model’s quality, as this is the motivation for carrying out adaptive sampling. Moreover, as mentioned above, the comparison should be made in terms of the density estimates produced by the different sampling methods, rather than AUC values alone, as such a comparison is more refined and is more likely to expose any differences in the quality of the resulting models. Both of these tasks are the subject of current work.

The similarity measures and, more generally, any PF-agnostic methods have two limitations. As discussed previously, they can only be used in a batch mode: any one point's utility can only be assessed relative to a set of points in a batch rather than by itself in isolation as it arrives. However, the ability to make a decision about a point in a streaming fashion is clearly a very desirable feature. In addition, the agnostic aspect of the PF-agnostic methods means that they do not directly leverage the information that is of particular interest in the PF context.

For this reason, we are currently working on implementing a PF-specific metric. Since the overarching goal of the PF is density estimation, one way to assess a given point's value is to predict its likelihood or, equivalently, its log-likelihood without using the PF (since the point of using adaptive sampling is precisely to avoid using the PF on every observation). This approach will be covered in detail in a separate report.

Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was funded by the LLNL Laboratory Directed Research and Development (LDRD) Grant No. 11-ERD-035.

References

- [1] A. Singhal, “Modern information retrieval: A brief overview,” *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp. 35–43, 2001.
- [2] I. Androutsopoulos, J. Koutsias, K. Chandrinos, and D. Spyropoulos, “An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal email messages,” in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Athens, Greece), 2000.
- [3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [4] K. Hornik, I. Feinerer, M. Kober, and C. Buchta, “Spherical k-means clustering,” *Journal of Statistical Software*, pp. 1–22, 2012.
- [5] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *Computational and Applied Mathematics*, pp. 53–65, 1987.

AUC for Email Data with Subj AS and Subj-Body Training (30 Exp)

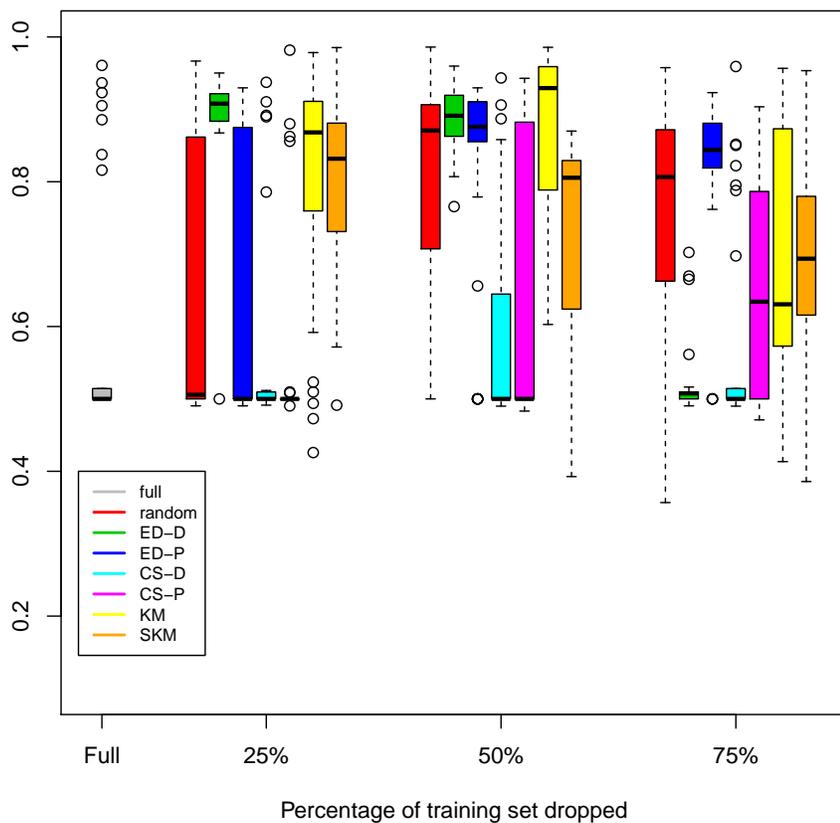


Figure 1: Boxplots of the 30 AUC values for the full training sets and reduced training sets using each of the dropping schemes and each of the dropping fractions.

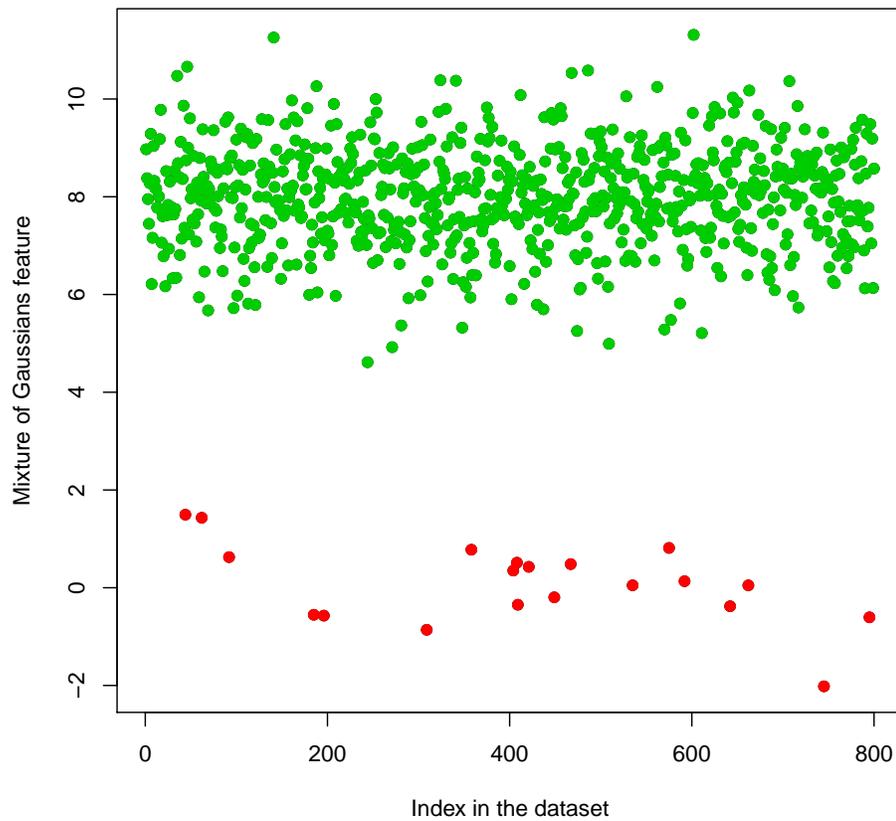


Figure 2: The Gaussian mixture dataset, consisting of 20 observations in one cluster and 780 observations in the other. The y-axis show the value of the continuous Gaussian mixture feature. The two colors indicate the categorical feature label, which is perfectly correlated with the cluster membership.