Alemi F, et al., Fortune J Health Sci 2024
DOI:10.26502/fjhs.165

has more of both X_1 and X_2. Partial sets are all the Means that have less values of either X_1 or X_2. In Figure 1 these are B, C, and D. For example, B has less of X_1 than case P and same value on X_2 as case P.

A more complex situation arises when the Mean has some features less and other features more than the case P. The Mean A is a mixed Mean, it has level a_2 on X_2, which exceeds p_2 for the case P. It has level a_1 on X_1 which is less than p_1 for case P. Similarly, Mean E is a mixed Mean, compared to P, it exceeds on X_1 but is less on X_2. These mixed Means can belong to either partial or excessive category, depending on the net impact of Means where just X_1 and X_2 are present. If $Y(X_1 = 0, X_2 = a_2) > Y(X_1 = a_1, X_2 = 0)$, then Mean A is classified as excessive because the impact of excess on X_2 is higher than the impact of loss on feature X_1. Similarly, Mean E is assigned to excessive set if the net effect of having more of X_1 and less of X_2 is positive, i.e., $Y(X_1 = 0, X_2 = e_2) > Y(X_1 = e_1, X_2 = 0)$. Once all Means are classified into partial and excessive sets, then the outcome for patient P is bounded by the maximum outcome in the partial set and the minimum outcome in the excessive set. The middle of the upper and lower bound is an estimate of the outcome for case P.

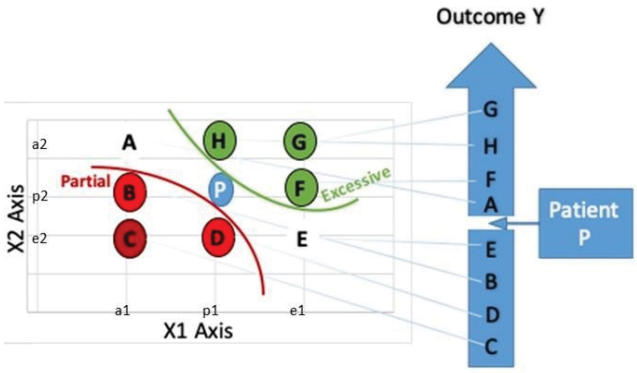

Figure 1: Partial and Excessive Sets in Matches to Patient P in Two Dimensions

For a more clinical example, consider predicting mortality from two factors: severity of cardiac illnesses and severity of infectious diseases in the patient's medical history. We agree that more severe illnesses lead to more mortality, thus the relationship is monotone. Suppose in a training set we have 8 types of patients, as in Figure 1, each with differing levels of cardiac and infectious illnesses. Suppose that a patient shows moderate levels of severity for both cardiac and infectious diseases. Then, C, B, and D are partial types because they have at least one feature less severe than patient P and no feature worse than patient P. Furthermore, F, G, and H are excessive Means or types of patients as they have at least one feature that exceeds in severity than patient P and no feature that is less severe. The patient types A and E are mixed types, as one feature indicates more severe and another less severe illness. These two mixed types are reclassified into excessive or partial types based on the relative severity of types with just cardiac and infectious diseases. Suppose, A is classified as excessive and E as partial. The algorithm calculates the probability of mortality for patient P from the maximum of partial patient types, in this case, E and the minimum of excessive patient types, in this case A.

Note that in these comparisons, we never used feature-based reasoning. We did not create a model, e.g., a regression equation, about how much features X_1 or X_2 contribute to the outcome Y. We also did not use Euclidian distance (squared root of sum of squared differences on each feature) between cases to judge the similarity of the cases. The entire inference was based on outcomes in various cases; we examined Means where X_1 or X_2 had different levels. The outcome for case P was inferred from the outcome for Means in the training set of data, without calculating the contribution of each feature to the outcome.

Description of the Algorithm

This algorithm repeatedly uses a concept known in the literature as Preferential Independence to organize the data [5]. In preferential independence, shared features of a Mean do not affect the order of outcomes of the Means. If S is a set of shared features and N is a set of features not shared, then $Y(S = s, N = n)$ is the outcome for the Mean with shared features "s" and unshared features "n". Preferential independence assumes that the order of outcomes for any two Means stays the same, independent of the shared features. Changing shared features "s" to "t" does not change the order of outcomes of the two Means:

$$Y(S = s, N = n) > Y(S = s, N = m) \Leftrightarrow$$

$$Y(S = t, N = n) > Y(S = t, N = m)$$

The monotone organization of features, discussed earlier, requires that a Mean with an added feature should not have a lower outcome. If n indicates the feature set of the starting Mean, then monotone relationship requires:

$$n + 1 > n \Leftrightarrow Y(S = n + 1, N = 0) \geq Y(S = n, N = 0)$$

Preferential independence allows us to generalize the monotone relationship to a variety of pairwise comparisons of Means. Our experience to date has shown that in almost all databases, the monotone relationship does not hold for all pairwise comparisons.

Step 1: Remove Order Violations from the Training Set. The first step in the algorithm is to remove Means in which adding a new feature violates the assumed monotone relationship. Violation of the required order identifies Means that are exceptions to the general model developed here to organize the data. These Means are not following the general pattern in the data and should be treated separately. If a new case matches the exceptions, then the model is ignored, and the outcome of the exceptional Mean is used to predict the

Alemi F, et al., Fortune J Health Sci 2024
DOI:10.26502/fjhs.165

outcome for the new case. To identify the exceptions, these steps can be followed:

1. Using the training set of data, start with the Mean where no feature is present. Call this the reference Mean
2. Add a feature to the reference Mean, creating a new Mean with an added feature than the starting Mean. We refer to these types of Means as "Excessive Mean." Look up the outcome for the constructed Excessive Mean in the data.
3. Test if the pair of Means have an order violation, where the Mean with added feature has a lower outcome than the Mean without it. Small variations in data can be ignored as all data have random variations that violate order requirements. The purpose of searching for exceptions is to identify large and consistent order violations.
4. Repeat steps 2 and 3 until all pairs of Means have been examined.

The Means with the highest number of order violations are assigned to the Exception set. When an order violation is observed, it is not clear if the violation is due to the reference Mean having an unusually high outcome or the Excessive Mean having an unusually low outcome. We propose to focus on the Mean with the largest number of order violations.

Step 2: Assign Means in Training Set to Excessive, Partial, and Mixed Sets. To predict the outcome for a new case, the Means in the training data (excluding exceptions) are classified into three types: partial, excessive, or mixed. If there is a Mean that exactly matches every feature of the new case, then there is no need for the algorithm and the prediction can be made from the outcome of this exact match. Similarly, if the new case has an exact match to Means in the exception set, then the outcome for the exception is used as the predicted value for the new case. A partial set, shown as Y^- is a set of Means where some of the features present in the patient are not matched in the Mean. An excessive set, shown as Y^+ is a set of Means, where all features present in the new case are matched and one or more additional features are present in the Mean but not in the new case. A Mean is mixed if some features in the new case are absent in the Mean; as well as some features in the Mean are absent in the new case. This set of Means is shown as $Y^{\pm}$. A Mean in this set is shown as having three sets of features:

$$Y(X_s = s, X_p = p, X_e = e).$$

It has features that are shared with the patients shown as having level "s", a set of features that are in the new case but absent in the Mean, shown as "p", and a set of features that are in the Mean but not in the new case, shown as "m."

Step 3: Re-classify Mixed Means. One classifies Means in $Y^{\pm}$ to either Y^- or Y^+ based on two methods. If the classification of the Mean to Excessive set creates an order violation, then the Mean is classified as Partial. This occurs if the re-classified Excessive has an outcome that is lower than the highest outcome for Means in the Partial set. Also, the reverse holds: if the classification of the Mean to partial creates a new order violation, then it is classified as Excessive. This occurs if the reclassified Partial has an outcome that is higher than the lowest outcome of the Means in the Excessive set.

Not all re-classifications create order violations. If it is possible to reclassify the Mean to either Partial or Excessive without order violations, then one keeps the order of cases constructed from the mismatched features.

$$Y(X_s = s, X_p = p, X_e = e) \in Y^+ \Leftrightarrow Y(X_s = 0, X_e = e) > Y(X_s = 0, X_p = p)$$

In these situations, the order of the mixed case is decided based on what is known as "Corner Means." An Excessive Corner Mean is composed of just the additional features in the mixed Mean and no other feature being present. A Partial Corner Mean is composed of only missing features in the mixed Mean and no other features. By preferential independence, the comparison of Excessive and Partial Corner Means establishes the assignment of the mixed Mean. Corner Means occur often and the order among them can be observed in the data. If this is not the case, these corner Means must be estimated through data balancing. An SQL code for estimate the net impact of Corner Means is available through the first author [6].

Step 4: Estimate the Outcome for New Case. The outcome for the new case is predicted as the average of the partial Means with the highest outcome and Excessive mean with the lowest outcome:

$$Y = \begin{cases} Maximum(Y^-)/2\,, if Y^+ is null \\ (1 + Minimum(Y^+))/2\,, if Y^- is null \\ \dfrac{(Maximum(Y^-) + Minimum(Y^+))}{2}, Otherwise \end{cases}$$

There are situations in which no partial or no excessive matches exist. These situations occur at border points, where there are no lower or upper bounds. If the partial set Y^- is null, then 0 is assumed. If excessive set Y^+ is null, then 1 is assumed.

Optimality of the Algorithm: A mathematical proof of optimality can be found in Keeney and Raiffa [7], where they use mutual preferential independence to create a mathematical model that preserves order of outcomes. In such a model, the maximum of partial and the minimum of excessive Means are the closest Means to the new case. The assumption of preferential independence and monotone positive relationship between features and the outcome, implies that Means in the excessive set have higher outcome than the new case. If the Mean in the partial set has an outcome that exceeds the minimum of the excessive set, then preferential independence is violated. Therefore, the lowest outcome among the Means in the excessive set is the upper bound for the new case.

Alemi F, et al., Fortune J Health Sci 2024
DOI:10.26502/fjhs.165

Similarly, the Mean with the maximum outcome in the partial set is closest to the outcome of the new case. Since these two Means have the outcomes closest to the new case, therefore the estimated average of maximum and minimum outcome for the new case is optimal.

Identifying Interaction Terms

Some investigators assign meaning to feature weights in regression equations. These investigators may want to examine feature weights, even if 2NM algorithm is available. These investigators can use the procedures of the 2NM algorithm to identify which interaction terms might affect the outcome variable, then derive the regression model with the interaction term. To find out what interactions may exist in the data, the data is transformed as proposed for application of 2NM algorithm, then preferential independence plots are created. These plots are organized by examining. The effect of a variable (present or absent) on the outcome, as a function of different shared features. One line is plotted for cases with the shared features. Another line is plotted for cases without the features. The X-axis is the shared features, plotted in order of increasing outcome. Two lines are plotted. The first line is the outcome for the case with just shared features and the variable being examined. The second line is the outcome for the case with shared features and the variable being examined. If there is no interaction term between the variable and the shared features, then the two lines must be parallel or nearly parallel throughout the range. Changing the shared features should not change the impact of the variable. If there is an interaction term in the data involving the variable, then the two lines could be diverging or converging. If the two lines cross each other then it indicates a violation of preferential independence. Preferential independence plots are reported in previous papers [8]. Preferential Independence plots show at what shared features there is a change in impact of a variable. Typically, a large number of interaction terms are identified and, for ease of use, the common denominator among the interaction terms is used. Once possible interaction terms have been identified, then a regression with the combination of main effects and interaction effects can capture the non-linearity in the data.

Results

Simple Simulated Data

We show the accuracy of predictions using a simulated data set with a single standardized variable X. We focus on a single-variable simulation because it is easier to visualize the performance of 2NM in a single variable than in a multiple-variable simulation. Let us assume that increases in X never decrease values of Y and X is positively and monotonely related to Y. Figure 2 shows such a simulated relationship. The observed data, shown in blue dots, shows that the values of Y either stay the same or increase as X increases. The

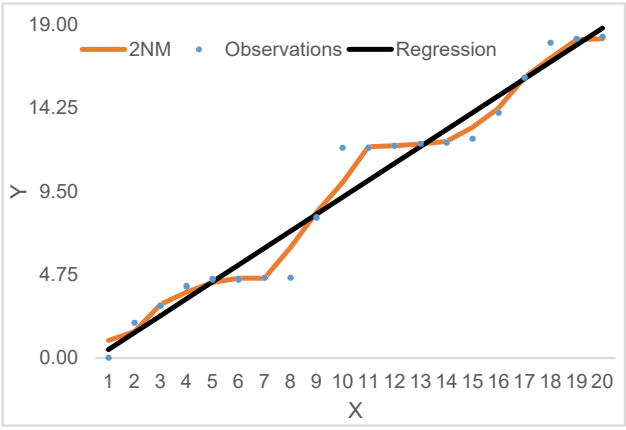

Figure 2: Regression and 2NM Fit to Simple Simulated Data with 1 Variable

black line shows the fit of a regression equation to the data. This fit is an optimal fit to the single feature in the simulation. The 2NM relies on case comparisons and in this simulation, the cases have only one variable. Thus, for any point in the data, the 2NM is calculated as the average of the point before and after it (the orange line). The R^2 associated with the fit of the regression line to the data was 0.85; in contrast the R^2 associated with the fit of 2NM was 0.88. The 2NM had a better fit, as it moved closer to the data points, when the relationship was not linear.

Application to Real Data

We tested the accuracy of predictions on a large database organized to guide what is likely to occur in a nursing home residents' life. The outcome of interest was probability of death in the next 6 months. The sample included 296,051 residents in Veterans Affairs nursing homes called Community Living Centers (CLCs). The period of study included 1/1/2000 through 9/10/2012. These data included a comprehensive assessment of residents' disabilities. Independent variables were gender (M), and whether the resident had feeding (E), bathing (B), grooming (G), dressing (D), bowel continence (L), bladder continence (U), toilet use (T), transfers (S), and walking (W) disabilities. To simplify, we show a combination of features present and assume that all features not mentioned were absent. Thus, MUE shows a male patient with urine continence and feeding disability. WB shows a female resident with walking and bathing disabilities. By policy, assessments were done within 14 days of admission, at least quarterly, or sooner when there has been an event such as hospitalization, or when the nursing home staff identify change in the resident's status. In our data, there were two peaks in the distribution of assessments; one for residents assessed every month (75,994 residents) and the other for residents assessed every 3 months (42,904 residents). The average time between assessments was 115 days (standard deviation of 235 days).

Alemi F, et al., Fortune J Health Sci 2024
DOI:10.26502/fjhs.165

The assessments were grouped into 1,346 unique combinations of functions and gender. The rare combination with less than 10 cases were ignored, resulting in 758 unique combinations of gender and disabilities, what in 2NM terminology is referred to as reliable Means. Ten percent of the data (67 cases) were set aside to test the accuracy of predictions. The remaining 90% were used for training of 2NM predictions and creating the reliable Means. The first step in the analysis is to check that the assumptions of 2NM are met, in particular we checked that for all Means in the training set adding another condition to the Mean does not reduce the probability of mortality. Table 1 lists situations where preferential independence was violated. Because of the large sample of data, we used both effect size (greater than 0.10 drop in the mortality rate) and statistical significance (Z>1.98). For example, in the first row in the table we see that adding "Grooming (G)" disability to "Older, Male, OM" patients with no disability will lead to a -0.66 drop in the mortality rate. Obviously, adding a disability should not reduce mortality rates and, in fact, out of 202 times where grooming disability was added to another condition only on 6 occasions the rate of mortality dropped. It almost always made the mortality rate worse. A quick analysis of Table 1 shows that violations of preferential independence occur mostly in patients who have no disabilities (shown as "males older than 74 years OM", "females older than 74 years OF", "males less than 65 years YM", "females less than 65 years YF", "males 65-74 years M", "females 65-74 years F"). In addition, violations of preferential independence also occurred in older male patients with all disabilities present but not urine incontinence or toileting disability (shown as OMERGBWDL). These situations cannot be modeled with 2NM. We set aside these situations as exceptions and focused on portions of the data where there was no violation of preferential independence. This reduced the Means available in the training set from 688 to 681 Means. In the Means remaining in the training set, there were no large and statistically significant violations of preferential independence.

Examples shown in Table 2 demonstrate the way the 2NM prediction works. In a new case MGWD (a patient who is under 65 years, male, with grooming, walking, and dressing disabilities), we searched the database and found the closest partial and excessive Means. The 5 Means provided differ from MGWD by one feature. MGW, MWD, and MGD are partial matches. MGBWD and MGTWD are excessive matches. MGBWD is not possible as its mortality rate is less than MGD, even though it has more disabilities. Thus, it violates the assumption of preferential independence. We ignore this Mean. Now we have a set of partial matches and another set of excessive matches. The maximum mortality rate for partial matches is MGD. The minimum mortality rate for excessive matches is MGTWD. Therefore, we predict that patient MGWD has a mortality rate of 0.025. In reality, it had a mortality rate of 0.015.

For patient MSGWDL (a patient who is under 65, male, with transfer, grooming, walking, dressing disabilities, and bowel incontinence), we follow the same procedure. The six Means identified differ from the patient by one feature. First, we remove Means that violate preferential independence: MRGTBWDL and MGBWDL. These two Means have more disabilities than MRGBWL but have lower mortality rate, a violation of preferential independence. The two nearest Means are MSGBWL, the maximum of partial set, and MRGBWDLU, the minimum of excessive set. For this patient, we predict a mortality rate of 0.018. The observed rate was 0.030. For our final example in Table 2, we have a patient OMWL (between 65 and 74 years, male, with walking disability and bowel incontinence). Again, the Mean MRGTBWDL is not possible as it has more features but lower mortality rate than OML. After eliminating this Mean, the maximum mortality rate for the partial set is OML and the minimum mortality rate for the excessive set belongs to OMWLU. Therefore, we predict this patient will have a mortality rate of 0.016 and in fact it had a mortality rate of 0.020.

In 130,428 set-aside validation cases, the 2NM had a McFadden Pseudo R-square of 0.51. A linear logistic regression was trained on the 1,174,218-training sample used by 2NM, using the linear combination of the same set of variables (age, gender, and disabilities). On the validation cases, the McFadden Pseudo R-squared for the linear logistic regression was 0.09.

Discussion

This paper shows that the optimal number of Means to use in k-Means methods is 2, if the data are transformed to fit a series of assumptions. These assumptions create a monotone positive relationship between the features and the outcome of interest. These assumptions also create a data set where the order of preferences among combination of features is the same in any subset of data. Data can be organized to meet these assumptions. Once the data are transformed to meet the assumptions of 2NM, then we divide Means into excessive and partial sets. Mixed Means (both partial and excessive) can be re-assigned based on the net impact of their partial and excessive feature sets. The maximum outcome of the partial and minimum outcome of the excessive Mean estimates the outcome for a new case. The procedure we used was accurate in a simple 1 variable simulation, showing how 2NM captures non-linear aspects of the data. We also compared 2NM and linear Logistic Regression in a large database of nursing home resident's disabilities. In predicting 6-month mortality in 10% set aside test cases, 2NM was more accurate than linear Logistic regression (Pseudo R-square of 0.51 versus

0.09). The 2NM procedure captured non-linearity among the variables in the model. The linear regression model did not.

Traditionally, Nearest Neighbor and k-Means have relied on a distance metric to predict the outcome for a new case. In this approach, one typically defines closeness based on the Euclidean distance calculated over all the features. Euclidian distance is arbitrary; one could have specified non-Euclidean measures of distance. Furthermore, the distance measure is often used with no regard for interaction among the features, a procedure known to reduce accuracy [9, 10]. Some investigators have tried to consider the non-linearity of the features by creating separate models for sub-classes in the data or through Kernel methods [11-13]. These efforts continue to weigh various features, albeit in new and novel ways but produce limited improvements in accuracy [14]. In contrast, in 2NM, closeness is measured on a single dimension: the outcome. The outcome is the only dimension where closeness matters, closeness in all other multidimensional spaces is irrelevant. By using feature-based measures of distance, Nearest neighbor and K-means undermine a key advantage in case-based reasoning. The approach presented here requires no feature weighting. It predicts entirely through aggregate, case-based comparisons. The assignment of cases into exceptions, excessive, and partial is done entirely through contrasting outcomes in pairs of cases. Since cases reflect the interaction among the features, then 2NM considers the interaction among the features. It does not disaggregate the case into features and then uses a formula (typically only main effects) to re-combine the features into a distance measure. The 2NM procedure may be most relevant in high dimensional data, where it is often difficult to specify all interactions among the features. The proposed algorithm may improve accuracy because it relies on case comparisons without feature-by-feature comparisons.

Limitations of the Algorithm

For the algorithm to be accurate, the Means in the training set must have values in the entire range of possible outcomes. Sufficient data should exist to have a variety of Means, ranging in probability of outcome from 0 to 1. For example, if we are examining probability of mortality, then we should have Means, or group of cases, where no one dies; where everyone dies, and different mixtures in between these two extremes. In predicting the outcome for a new case, a dense, uniform, distribution of Means and their probability of the outcome guarantees that the two Nearest Means for the new case are close to each other and thus a more precise prediction can be made.

Funding: No Funding.

Ethics approval: Earlier versions of this paper have benefitted from exchanges with Janusz Wojtusiak, Ph.D., Allison Williams, Ph.D., Cari Levy, M.D

What Is Known on This Topic?

- Case-based reasoning, such as k-means or Nearest Neighbor, take into account interactions among the features of the case, except when measuring distance among cases.
- Accuracy of case-based reasoning and feature-based predictions are similar

What This study Adds?

- This study reports accuracy of a new case-based reasoning that does not rely on feature-based distance calculations.

References

1. Pelleg D, Moore AW. X-Means: extending K-Means with efficient estimation of the number of clusters. In: ICML (2000): 727–734.
2. Bischof H, Leonardis A, Selb A. MDL principle for robust vector quantisation. Pattern Analysis & Applications 2 (1999): 59–72.
3. Hamerly G, Elkan C. Learning the K in K-Means. In: Advances in Neural Information Processing Systems (2003).
4. Raykov YP, Boukouvalas A, Baig F, Little MA. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS One 11 (2016): e0162259.
5. Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge University Press, Cambridge UK (1993).
6. Alemi, F. et al. Big Data in Healthcare: Statistical Analysis of Electronic Health Records, Chicago, Il, Health Administration Press (2020).
7. Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge University Press, Cambridge UK (1993).
8. Alemi F, Levy C, Citron BA, Williams AR, Pracht E, Williams A. Improving Prognostic Web Calculators: Violation of Preferential Risk Independence. J Palliat Med 12 (2016): 1325-1330.
9. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus 5 (2016): 1304.
10. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus 5 (2016): 1304.
11. Liu Y, Ge SS, Li C, You Z. k-NS: a classifier by the distance to the nearest subspace. IEEE Trans Neural Netw 22 (2011): 1256-68.

Alemi F, et al., Fortune J Health Sci 2024
DOI:10.26502/fjhs.165

12. Cevikalp H, Neamtu M, Barkana A. The kernel common vector method: a novel nonlinear subspace classifier for pattern recognition. IEEE Trans Syst Man Cybern B Cybern 37 (2007): 937-51.

13. Zhang P, Peng J, Domeniconi C. Kernel pooled local subspaces for classification. IEEE Trans Syst Man Cybern B Cybern 35 (2005): 489-502.

14. Veenman CJ, Reinders MJ. The nearest subclass classifier: a compromise between the nearest Mean and nearest neighbor classifier. IEEE Trans Pattern Anal Mach Intell 27 (2005): 1417-29.