**Solutions to Chapter 1**

**AN INTRODUCTION TO DATA MINING AND PREDICTIVE ANALYTICS**

**Prepared by James Cunningham, Graduate Assistant**

1. **For each of the following, identify the relevant data mining task(s):**
   1. **The Boston Celtics would like to approximate how many points their next opponent will score against them.**

Estimation

* 1. **A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.**

Classification

Clustering

Description

* 1. **A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.**

Classification

* 1. **A political strategist is seeking the best groups to canvass for donations in a particular county.**

Clustering

Classification

Description

* 1. **A Homeland Security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.**

Prediction

* 1. **A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.**

Estimation

1. **For each of the following meetings, explain which phase in the CRISP-DM process is represented:**
   1. **Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.**

Model Evaluation Phase

* 1. **The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.**

Data Understanding Phase

* 1. **The data mining consultant meets with the Vice President for Marketing, who says that he would like to move forward with customer relationship management.**

Business Understanding Phase

* 1. **The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.**

Model Deployment Phase

* 1. **The analysts meet to discuss whether the neural network or decision tree models should be applied.**

Modeling Phase

1. **Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.**

Data mining requires human direction in order to be both effective and appropriate as problem-solving is a human process requiring human critical thinking every step of the way. As stated in the text, data mining without proper human direction is something that is ***very easy to do badly***. It is very easy to derive results that are damaging to business processes by (1) failing to understand the business problem at hand, (2) failing to understand the data sets at hand (and their interrelationships), (3) failing to select appropriate modeling techniques, and (4) failing to evaluate model results correctly.

One very popular fallacy is that data mining can be completely autonomous and thus requires little to no human direction. Applying data mining software features at random is bound to produce the wrong answer to the wrong question with the wrong data. In fact business decisions made based on inappropriate analyses are much more damaging and costly than those made based on no analysis at all. Also, once a model is deployed, it must be monitored for its efficacy and will most often need to be tuned over time.

1. **CRISP-DM is not the only standard process for data mining. Research an alternative methodology (Hint: SEMMA, from the SAS Institute). Discuss the similarities and differences with CRISP-DM.**

SEMMA is a process developed by the SAS Institute for conducting a data mining project. Each letter in the acronym SEMMA identifies a separate stage of the data mining process as follows:

**S**ample – The first stage in SEMMA entails extracting a representative sample of a much larger data set. Please note that this stage is optional and thus used at the discretion of the analyst.

**E**xplore – The second stage in SEMMA entails searching for unanticipated trends, patterns, and anomalies in order to gain an understanding of the data and develop ideas.

**M**odify – The third stage in SEMMA entails modifying the data set through a combination of selecting original variables and more importantly transforming variables and deriving new ones that would be most conducive to a data modeling exercise.

**M**odel – The fourth stage in SEMMA entails allowing the software to determine the best combination of variables that predict a desired outcome.

**A**ssess – The fifth and final stage in SEMMA entails evaluating model efficacy and estimating how well it will perform if deployed.

The CRISP-DM process was developed by a consortium pioneered by DaimlerChrysler, SPSS, and NCR and consists of six stages or *phases* as follows:

**Business Understanding** – The first phase entails gaining an understanding of the business problem at hand and translating this into a data mining problem to be solved and an initial solution approach. In direct contrast with SEMMA, we observe that CRISP-DM prescribes business-requirements development as an explicit activity and the specific data mining problem and solution approach as explicit deliverables whereas SEMMA does not. SEMMA prescribes delving right into the data set, which can lead to significant time wasted (that will most likely be proportional to the dimensionality of the data set being explored).

**Data Understanding** – The second phase entails determining how data will be collected and exploratory analysis. This phase is similar in nature to SEMMA’s Explore stage, but in contrast with SEMMA, the exploratory analysis activities of the CRISP-DM Data Understanding phase are conducted from the perspective of solving a particular data mining problem. Therefore, while exploration conducted in SEMMA’s Explore stage seems to be by pure brute-force, exploration conducted in CRISP-DM’s Data Understanding phase is done from the perspective of a specific data mining problem to be solved. In other words, the exploratory analysis in CRISP-DM’s Data Understanding phase is expected to be more effective and more efficient focusing on exploring correlations between predictors and interactions between predictors and a specific target variable.

**Data Preparation** - The third phase entails all of the actions (e.g. selections, transformations, derivations, etc.) needed to develop a data set that is most conducive to a data modeling exercise. This phase is similar to SEMMA’s Modify stage, but contrast with SEMMA, the preparation activities conducted in the CRISP-DM Data Preparation phase are done so with a specific data mining problem and target modeling approach in mind. This is a critical distinction between the two processes. As an example, if we have data that is highly inter-correlated or *multicoliear*, we can leverage a dimensional transformation such one produced via Principal Components Analysis (PCA) to eliminate the multicolinearity, but only for certain types of modeling approaches. Therefore, since the CRISP-DM Data Preparation phase has a target modeling approach in mind when preparing data, it can leverage advanced transformational techniques like PCA appropriately and is thus superior to the SEMMA Modify stage.

**Modeling** - The fourth phase entails the human-directed application of multiple modeling techniques in order to (1) optimize the balance between model bias and model variance and (2) maximize the ability for these models to operate effectively on new observations. While this is similar to SEMMA’s Model stage, the CRISP-DM Modeling phase is human-directed whereas SEMMA’s Model stage appears to be autonomous with little or no human direction. As stated in the text, autonomous data mining is a dangerous practice.

**Evaluation** – The fifth phase entails thorough evaluation of both the (1) constructed models for their efficacy and performance as well as the (2) approach used to construct the models to ensure that the constructed models actually solve the business problem at hand. While this phase is similar to SEMMA’s Assess stage, the CRISP-DM Evaluation phase verifies that the models constructed actually solve the business problem at hand. Since SEMMA does not prescribe formal definition of the business problem to be solved, the SEMMA Assess stage may actually result in a model that performs well but operates on the wrong target variable and corresponding predictors and thus has little or no business value.

**Deployment** – The sixth and final phase entails preparing the model results so that it can be leveraged by the business sponsor. For simpler data mining projects, this may entail generating a report that the sponsor may use to base business decisions off of. For more complex projects, this may entail implementation of the final model in a commercial rules-engine software package. In direct contrast with SEMMA, there is no corresponding stage in the SEMMA process prescribing model deployment.

**Solutions to Chapter 2**

**DATA PREPROCESSING**

**Prepared by James Cunningham, Graduate Assistant**

1. **Describe the possible negative effects of proceeding directly to mine data that has not been preprocessed.**

Neglecting to preprocess the data adequately before data modeling begins will likely produce data models that are unreliable and whose results should be considered dubious as best. Performing data cleaning and data transformation during the data preparation phase is absolutely necessary for successful data mining efforts.

For example, suppose you are analyzing a data set that includes a person’s Age and Date\_of\_Birth attributes, and you want to calculate the average Age. Now, if 5% of the records contain a value of 0 for Age, the mean value would be very misleading and inaccurate. One solution to this problem would be to derive Age for the zero-based records based on information contained in the Date\_of\_Birth variable. Now, the mean value for Age is more representative of those persons in the data set.

1. **Refer to the income attribute of the five customers in Table 2.1, before preprocessing.**
2. **Find the mean income before preprocessing.**

The mean value for Income before preprocessing is 38,999.80 and is derived by the possible inclusion of Income values -40,000 (erroneous) and 100,000 (possible outlier).

1. **What does this number actually mean?**

In this case the mean value has little meaning because we are combining real data values with erroneous values.

1. **Now, calculate the mean income for the three values left after preprocessing. Does this value have a meaning?**

However, the mean value for Income produced by values 75,000, 50,000, and 10,000 (9,999 rounded to nearest 5,000) is 45,000. The latter value is certainly more representative of the true mean for Income, now that the records containing questionable values have been excluded.

1. **Explain why zip codes should be considered text variables rather than numeric.**

Zip codes should be considered text variables because they cannot be quantified on any numeric scale. Even their order has no numerical significance.

1. **What is an outlier? Why do we need to treat outliers carefully?**

Consider a set of numerical observations and the center of this observation set. An outlier is an observation that lies much farther away from the center than the majority of the other observations in the set.

We must treat outliers carefully because they can cause us to misrepresent the true center of an observation set incorrectly if they lie significantly farther away from the other observations in the set.

1. **Explain why a birthdate variable would be preferred to an age variable in a database.**

A birthdate variable is preferable to an age variable in a database because (1) one can always derive age from birthdate by taking the difference from the current date, and (2) age is relative to the current date only and would need to be updated continuously over time in order to remain accurate.

1. **True or false: All things being equal, more information is almost always better.**

The answer is true. In general, more information is almost always better. The more information we have to work with, the more insight into the underlying relationships of a particular domain of discourse we can glean from it.

1. **Explain why it is not recommended, as a strategy for dealing with missing data, to simply omit the records or fields with missing values from the analysis.**

It is not recommended to omit records or fields from an analysis simply because they have missing values. The rationale for this recommendation is that omitting these fields and records may cause us to lose valuable insight into the underlying relationships that we may have gleaned from the partial information that we do have.

1. **Which of the four methods for handling missing data would tend to lead to an underestimate of the spread (e.g., standard deviation) of the variable? What are some benefits to this method?**

Replacing a missing value by the attribute value’s mean artificially reduces the measure of spread for that particular attribute. Although the mean value is not necessarily a typical value, for some data sets this form of substitution may work well. Specifically, the effectiveness of this technique depends on the size of the variation of the underlying population. In other words, the technique works well for populations having small variations, and works less effectively for populations having larger variations.

Several benefits to leveraging this method include (1) ease of implementation (i.e. only one value to impute), (2) preservation of the standard error (i.e. no additional residual error is introduced).

1. **What are some of the benefits and drawbacks for the method for handling missing data that chooses values at random from the variable distribution?**

By using the data values randomly generated from the variable distribution, the measures of center and spread are most likely to remain similar to the original; however, there is a chance that the resulting records may not make intuitive sense.

1. **Of the four methods for handling missing data, which method is preferred?**

Having the analyst choose a constant to replace missing values based on specific domain knowledge is overall, probably the most conservative choice. If missing values are replaced with a flag such as “missing” or “unknown”, in many situations those records would ultimately be excluded from the modeling process; that is, all remaining valid, potentially important, values contained in those records would not be included in the data model.

1. **Make up a classification scheme which is inherently flawed, and would lead to misclassification, as we find in Table 2.2. For example, classes of items bought in a grocery store.**

|  |  |
| --- | --- |
| **Breakfast** | **Count** |
| Cold Cereals | 72 |
| Sugar Smacks | 1 |
| Cheerios | 2 |
| Hot Cereals | 28 |
| Cream of Wheat | 3 |

Using the table above, the “Breakfast” categorical attribute contains 5 apparent classes. However, upon further inspection the classes are discovered to be inconsistent. For example, both “Sugar Smacks” and “Cheerios” are cold cereals, and “Cream of Wheat” is a hot cereal. Below, the cereals are now classified according to one of two classes, “Cold Cereals” or “Hot Cereals.”

|  |  |
| --- | --- |
| **Breakfast** | **Count** |
| Cold Cereals | 75 |
| Hot Cereals | 31 |

1. **Make up a data set, consisting of the heights and weights of six children, in which one of the children is an outlier with respect to one of the variables, but not the other. Then alter this data set so that the child is an outlier with respect to both variables.**

In the table below, Child #1 is an outlier with respect to Weight only. All children in the table are close in Height differing at most by 9 inches. However, all children except for Child # 1 are close in Weight differing at most by 7 pounds. Child #1 is an outlier as the Weight differs by 18 pounds from the second-heaviest child (Child #6), making this right-tailed difference in Weight greater than the entire Weight range for the other five children.

|  |  |  |
| --- | --- | --- |
| Child | Height (in) | Weight (lbs) |
| 1 | 49 | 100 |
| 2 | 50 | 75 |
| 3 | 52 | 77 |
| 4 | 55 | 79 |
| 5 | 57 | 80 |
| 6 | 58 | 82 |

In the table below, Child #1 is an outlier with respect to both Height and Weight. All children except for Child #1 in the table are close in Height differing at most by 8 inches and are close in Weight differing at most by 7 pounds. Child #1 is an outlier for both Height and Weight as the Height differs by 14 inches from the second-shortest child (Child#2) (which is greater than the entire Height range of the other five children), and the Weight differs by 18 pounds from the second-heaviest child (Child #6) (which is greater than the entire Weight range of the other five children).

|  |  |  |
| --- | --- | --- |
| Child | Height (in) | Weight (lbs) |
| 1 | 36 | 100 |
| 2 | 50 | 75 |
| 3 | 52 | 77 |
| 4 | 55 | 79 |
| 5 | 57 | 80 |
| 6 | 58 | 82 |

**Use the following stock price data (in dollars) for Exercises 13–18**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | **7** | **20** | **12** | **75** | **15** | **9** | **18** | **4** | **12** | **8** | **14** |

1. **Calculate the mean, median, and mode stock price.**

The ***mean*** is calculated as the sum of the data points divided by the number of points as follows:

Mean Stock Price = (10+7+20+12+75+15+9+18+4+12+8+14) / 12 = 204 / 12 = $**17**.

The ***median*** is calculated by placing the prices in order and (a) selecting the middle value if the number of points is odd, or (b) taking the average of the two middle values if the number of points is even. Since we have twelve points, median is calculated as follows:

Median Stock Price = mean of center values {4,7,8,9,10,**12,12**,14,15,18,20,75} = 24/2 = $**12**.

The ***mode*** is calculated as the value that occurs the most often in the set and is calculated as follows:

Mode Stock Price = highest frequency of {4,7,8,9,10,**12,12**,14,15,18,20,75} = $**12**.

1. **Compute the standard deviation of the stock price. Interpret what this number means.**

The ***standard deviation*** represents the expected distance of a point chosen at random from a data set to the center of that set and is calculated by taking the square root of the ***variance***. The variance is the average of the sum of squared distances of each point from the data-set mean. Given that the mean is $17 (see Exercise #13) for this set, the variance for the set of stock prices is calculated as follows:

Stock Price Variance (Var) =

(4-17)2+(7-17)2+(8-17)2+(9-17)2+(10-17)2+(12-17)2+(12-17)2+(14-17)2+(15-17)2+(18-17)2+(20-17)2+(75-17)2 =

(-13)2 + (-10)2 + (-9)2 + (-8)2 + (-7)2 + (-5)2 + (-5)2 + (-3)2 + (-2)2 + (1)2 + (3)2 + (58)2 =

169 + 100 + 81 + 64 + 49 + 25 + 25 + 9 + 4 + 1 + 9 + 3364 = 3900 / 12 = **325 $2**.

Taking the square root of the Variance, the Standard Deviation (SD) is calculated as follows:

Stock Price Standard Deviation (SD) of Stock Price = √(325) = **±$18.03**.

Since the mean is $17 and the standard deviation is plus/minus $18.03, the expected price of a stock drawn at random from the set of twelve stocks is expected to lie mathematically between ($17–$18.03) = **-$1.03** (i.e. $0.01 since we assume that a stock price can never be less than one penny USD) and ($17+$18.03) = $**35.03**.

As we can see, each stock with the exception of the one priced at $75 is priced within this range.

1. **Find the min-max normalized stock price for the stock worth $20.**

Min-Max normalization scales an observation relative to the data-set’s range resulting in a value between 0 and 1 (this value has no units) and is formulated as follows:

**MinMaxXi = [Xi – Min(X)] / [Max(X) – Min(X)]**

Therefore, the min-max normalized stock price of $20 is calculated as follows:

MinMax($20) = ($20 - $4) / ($75 - $4) = ($16) / ($71) = **0.2254**.

1. **Calculate the midrange stock price.**

The midrange stock price is the central price for the entire price range and is formulated as follows:

**MidRangeX = [Max(X) + Min(X)] / 2**

For the problem at hand we have as follows:

MidRangeX = ($75 + $4) / 2 = ($79) / 2 = **$39.5**

1. **Compute the Z-score standardized stock price for the stock worth $20.**

Z-Score standardization scales an observation where the mean value is zero, the SD is 1 and most values lie between -4 and 4 (this value has no units) and is formulated as follows:

**Z-Score(X) = [Xi – Mean(X)] / |SD(X)|**

Given the mean of $17 (see Exercise #13) and |SD| of 18.03 (see Exercise #14), The Z-Score for the stock price of $20 is calculated as follows:

Z-Score($20) = ($20 - $17) / $18.03 = ($3) / $18.03 = **0.1664**.

Please note that this value makes sense as it is slightly greater than zero just as $20 is slightly greater than $18.03.

1. **Find the decimal scaling stock price for the stock worth $20.**

Decimal standardization scales an observation to a value between -1 and 1 (this value has no units) and is formulated as follows:

**Decimal(Xi) = Xi / 10d**

where d is the number of digits in the observation in the data set having the largest absolute value. Since the largest stock price is $75, d = 2 as there are two digits in this price. The decimal standardization is then calculated as follows:

Decimal($75) = $75 / $102 = $75 / $100 = **0.75**

1. **Calculate the skewness of the stock price data.**

Skewness is the lack of normalization of a Z-Score-standardized distribution and is measured using the following formula: