|  |  |
| --- | --- |
| Sr No | Question |
| 1 | **Variability Relationships**  The transform produced a rectangular distribution insofar as the actual values allowed for it. Each  variable was considered individually, and was redistributed individually without reference  to the distributions of other variables. In addition to this within-variable relationship, there  is also a critical between-variable relationship that exists for all of the variables. The  between-variable relationship expresses the way that one variable changes its value  when another variable changes in value. It is this multiple-way, between-variable  relationship that will be explored by any modeling tool. Since the chosen modeling tool is  going to explore these relationships between variables, it is critical to preserve them, so  far as possible, when replacing missing values.  Variability forms a key concept in deciding what values to use for the replacement.  Standard deviation is one measure of that variability. To  exemplify the underlying principles of preserving variability, consider a single variable  whose values are transformed into a rectangular distribution. Figure 8.1 shows the values  of such a variable. The 11 values of the original series are shown in the column headed  “Original sample.” Suppose that the value in series position 11 is missing and is to be  replaced. Since in this example the actual series value is present, it is easy to see how  well any chosen estimator preserves the relationships.    **Figure:** Estimating the value of position 11, given only the values in positions  1 through 10.  Position 11 has an actual value of 0.6939. The mean for the original 11-member series is  0.4023. If the value for position 11 is to be estimated, only positions 1 through 10 are used  since the actual value of position 11 is assumed unknown. The third column (headed  “Position 11 missing”) shows the 10 values used to make the estimate, with the mean and  standard deviation for these first 10 positions shown beneath. The mean for these first 10  positions is 0.3731. Using this as the estimator and plugging it into position as an estimate  of the missing value (position 11) changes the standard deviation from about 0.2753 to  0.2612. The column mean is unchanged (0.3731). Using the mean of instances 1 through  10 has least disturbed the *mean* of the series. The actual value of the “missing” value in  position 11 is 0.6939, and the mean has estimated it as 0.3731—a discrepancy of 0.3208.  Suppose, however, that instead of using the mean value, a value is found that least  disturbs the standard deviation. Is this a more accurate estimate, or less accurate?  Position 11 in column five uses an estimate that least disturbs the standard deviation.  Comparing the values in position 11, column four (value not disturbing mean as  replacement) and column five (value not disturbing standard deviation as replacement),  which works best at estimating the original value in column two?  The column four estimator (preserving mean) misses the mark by  0.6939 – 0.3731 = 0.3208.  The column five estimator (preserving standard deviation) only misses the mark by  0.6622 – 0.6939 = 0.0317.  Also, preserving the standard deviation (column five) moved the new mean closer to the  original mean value for all 11 original values in column one. This can be seen by comparing the mean and standard deviation for each of the columns. The conclusion is  that preserving standard deviation does a much better job of estimating the “true” mean  and provides a less biased estimate of the “missing” value.  Was this a convenient and coincidental fluke? Figure 8.2 shows the situation if position 1  is assumed empty. The previous example is duplicated with values for preserving mean  and standard deviation shown in separate columns. As before, generating the  replacement value by preserving standard deviation produces a less biased estimate.    Why is it that preserving variability as measured by standard deviation produces a better  estimate, not only of the missing value, but of the original sample mean too? The reason  is that the standard deviation reflects far more information about a variable than the mean  alone does. The mean is simply a measure of central tendency.  The standard deviation reflects not just central tendency, but also information about the  variability within the variable’s distribution. If the distribution is known that knowledge  contributes to determining a suitable replacement value. It is this use of additional  information that produces the better estimate.  Preserving variability works well for single missing values in a variable. If multiple values  are missing, however, the estimator still produces a single estimate for all missing values,  just as using the mean does. If both positions 1 and 11 were missing in the above  example, any estimator preserving standard deviation would produce a single estimate to  replace both missing values. So, there is only one single replacement value to plug into all  the missing values. Does this cause any problem? |
|  |  |
| 2 | As mentioned before, what are the basic issues that must be resolved in data  reduction? Again, we provide a series of questions associated with the correct answer  related to each type of task that belongs to the data reduction techniques:  • How do I reduce the dimensionality of data?—Feature Selection (FS).  • How do I remove redundant and/or conflictive examples?—Instance Selection  (IS).  • How do I simplify the domain of an attribute?—Discretization.  • How do I fill in gaps in data?—Feature Extraction and/or Instance Generation. |
|  |  |
| 3 | The alternative to Gaussian model for a long tail data distribution is===== Log Normal Distribution. |
|  |  |
| 4 | The alternative to Gaussian model for a long tail data distribution is===== power law distribution. |
|  |  |
|  |  |
|  |  |
| 5 | In case of linear transformation where y= (x-Xmin)/ (xmax-xmin) the requirement is to transform the quantity to fall whenever the original quantity goes up can be achieved by the equation: 1- y |
|  |  |
| 6 | The variance of random variable X is defined by  Var(X) = E[(X-m)2], where, m=E(X)  Then the Var(X) is given by  Var(X)= E(X)2 – (E(X))2  Variance will never be Zero if there is Uncertainty.  Properties of Var(X)= Ea2)-(E(a))2 = a2-a2=0  Var(aX+b)=a2.Var(X) = a2 . Var(X) |
|  |  |
|  | Standardization becomes Normalization for large data set. |
|  |  |
|  | Var= 100 and Var= 1000 which dataset is good?  X1, ….X100 Mu  X1……X1000 Mu  Var(X1+……+X1000)/1000  = (1/1000000)(1000 Lambda)  = Lambda/1000  Var(X1+……+X100)/100  = (1/10000)(100 Lambda)  = Lambda/100  Therefore we go for more samples. |
|  |  |
|  | Var(X+Y)= var(X) + Var(Y) + 2 Cov(X,Y)====under all circumstances  Var(X+Y)= var(X) + Var(Y) === iff X & Y are independent |
|  |  |
|  | To establish relationship between X & Y, assume that a large set of data is collected in such a case it can be assumed that X & Y jointly follow bivariate normal distribution. |
|  |  |
|  | Statistically we believe that for large data set two variable X & Y have linear relationship. |
|  |  |
|  | Corr(U,V) = cov(U,V)/ sqrt(var(U).var(Y)) where, U=X-Y and V=Y+Z  = cov(X.Y).cov(X.Z).cov(Y.Z) – Var(Y) |
|  |  |
|  | We have two estimatations  h1= (X1,X2, …. Xn) and h2=  9X1,X2,….Xn) are two estimatators of Theta. Which one is more efficient?  h1 is said to be more efficient then h2, if Var(h1) < Var(h2) |
|  |  |
|  | What should be the minimum value of variance?  The MVUE: Minimum Variance Unbiased Estimator is sufficient. However, sufficient may not be efficien. (Rao’s Theory) |
|  |  |
|  | We have two data sets one consisting of large samples and other small samples  First data set is better than second if ===  Second is better than first if ===== |
|  |  |
|  |  |
|  | The total variance is thus the difference between the average of the squared magnitude  of the data points and the squared magnitude of the mean (average of the  points). (pg 21) |
|  |  |
|  | Mean of cantered data matrix == Zero |
|  |  |
|  | Sample Mean is Unbiased An estimator ˆθ is called an unbiased estimator for  parameter θ if E[ˆθ] = θ for every possible value of θ. (pg 48) |
|  |  |
|  | It is worth noting that variance is in fact the second moment about the mean,  corresponding to r = 2, which is a special case of the r-th moment about the mean  for a random variable X, defined as E [(x − μ)r]. (pg. 53) |
|  |  |
|  | In other words, the sample mean ˆμ varies or deviates from the mean μ in proportion  to the population variance σ2. However, the deviation can be made smaller by  considering larger sample size n. (pg 55) |
|  |  |
|  | Sample Variance is Biased, but is Asymptotically Unbiased The sample  variance in (2.10) is a biased estimator for the true population variance, σ2, i.e.,  E[ˆσ2] 6= σ2. (pg. 55) |
|  |  |
|  | Covariance The covariance between two attributes X1 and X2 provides a measure  of the association or linear dependence between them, and is defined as  σ12 = E[(X1 − μ1)(X2 − μ2)]  = = E[X1X2] − E[X1]E[X2]  If X1 and X2 are independent random variables, then we conclude that their  covariance is zero. This is because if X1 and X2 are independent, then we have  E[X1X2] = E[X1] · E[X2]  which in turn implies that  σ12 = 0  However, the converse is not true. That is, if σ12 = 0, one cannot claim that X1 and  X2 are independent. All we can say is that there is no linear dependence between  them, but we cannot rule out that there might be a higher order relationship or dependence between the two attributes. (pg. 58) |
|  |  |
|  | Correlation The correlation between variables X1 and X2 is the standardized covariance,  obtained by normalizing the covariance with the standard deviation of each  variable. (pg 59) |
|  |  |
|  | Entropy Entropy, in general, measures the amount of disorder or uncertainty in a  system. In the classification setting, a partition has lower entropy (or low disorder)  if it is relatively pure, i.e., if most of the points have the same label. On the other  hand, a partition has higher entropy (or more disorder) if the class labels are mixed,  and there is no majority class as such. (pg. 543) |
|  |  |
|  | If the partition is pure, then the probability of the majority class is 1 and the  probability of all other classes is 0, and thus, the Gini-index is 0. On the other  hand, when each class is equally represented, with probability P(ci|D) = 1  k , then  the Gini-index has value k−1  k . Thus, higher values of the Gini-index indicate more  disorder, and lower values indicate more order in terms of the class labels. (pg. 544) |
|  |  |
|  |  |
|  |  |
|  |  |