**WEEK 3**

1. One of the variable transformations as part of data processing is Normalization or Standardization, as described in our textbook. Using Weka with the Pima Indians onset of diabetes dataset (data/diabetes.arff), answer the following questions:
   * Describe when you would use each of Normalization and Standardization to your data
   * Applying the ‘Normalize’ filter using the default parameter values, how many distinct and unique values of the blood pressure ‘pres’ attribute are there?
   * Applying the ‘Standardize’ filter using the default parameter values, what are the minimum and maximum values of the blood pressure ‘pres’ attribute?

**2.** In real-world data, tuples (records, instances) with missing values for some attributes are a common occurrence. Describe at least three methods for handling this problem.

**3.** Question #3 of 2.6 Exercises in our textbook.  
  
The following two questions are using the **Weka**. You can find the datasets in the ***data*** folder where you have installed the Weka.

**4.**From the ***soybean*** dataset, answer the following questions:

1. How many instances are there?
2. How many attributes are there?
3. How many possible values are there for the *crop-hist* attribute?
4. Which of these attributes has *fungicide* as a possible value?
5. Examine the *soybean* ARFF file header and say when the dataset was first used?

**5.** Using the ***segment-challenge***dataset, find the number of records detecting 1) outliers & 2) extreme values respectively based on the Interquartile Range filtering algorithm in Weka. Use the following parameter values:

OutlierFactor = 3.8

ExtremeValuesFactor = 6.8

1.A.  
Normalization is the process of rescaling the dataset between the range of 0 and 1. Normalization is best used in cases where the distribution of the data is a non Gaussian (bell-curve) distribution. In general the proximity measures(similarity or dissimilarity between the object attributes) are transformed to range between 0 and 1 .  
The formula for normalization is:  
X'=x-xmin/xmax-xmin  
Where x= Original value of the feature  
x'= new value of the feature  
xmin=Minimum value of the feature  
xmax= maximum value of the feature  
Depending upon the range in which the values of the dataset lies which could be a large range say for example [0,infinity] this rescaling method would require a non linear transformation of the data and the object values will loose the original relationship with one another.  
In case of the Pima Indians onset of diabetes dataset Adaptive learning algorithm (neural network algorithm) compares the output values against the threshold of 0.488 and assigns them a value of 0 (negative) and 1(positive) which suggests this used normalization technique.

Standardization technique is used when the distribution of the dataset is a Gaussian distribution(bell curve), however not necessarily this is the case depending upon the algorithm we are using. This scaling method centers the values around the mean of the dataset with the value of Standard deviation being 1.  
The formula for Standardization is:  
X'= X-μ/σ

μ= Mean of the feature values  
σ= Standard deviation of the feature values  
Also the outliers in the dataset are unaffected by this scaling technique since it does not have a range between which the values should fall.

B.  
Normalize filter:  
Select Open URL>>Type the url address of the diabetes dataset  
From Filter >>Choose >> filters >> unsupervised >> attribute >> Normalize >> Apply  
Select "pres" attribute.  
Distinct= 47  
Unique= 8

C.  
Standardize filter:  
Select Open URL>>Type the url address of the diabetes dataset  
From Filter >>Choose >> filters >> unsupervised >> attribute >> Standardize >> Apply  
Select "pres" attribute.  
Minimum value= -3.57  
Maximum value= 2.733

2.  
Missing Values

A. Delete: The most basic strategy is to delete the row/column with null values. However it must be noted that this method is only useful when there is a large dataset and deletion would not cause the output be affected in one way or another.

B. Replace: This method replaces the missing values with the mean/median/mode of the dataset . For numeric attributes we replace the missing values with mean/median and for nominal attributes we can replace them with the mode of the dataset.

C.Predict:By considering the features present in the current dataset we can use prediction algorithms to predict the missing values. One such algorithm is Random-forest algorithm which finds the most applicable features among the set of the current features and decides if the feature would be contributing to the prediction process.

3.  
The boss is right.  
The marketing director is right in the sense of using ratio attribute to measure the customer satisfaction . However customer satisfaction cannot be measured based on only number of complaints associated with the product. This case is similar to how products are rated on Amazon, there are products that have the highest number of complaints/reviews but are sold the most and then there are products that have negligible to zero complaints but also arent that sellable.  
This issue can be resolved by taking into account the ratio of complaints to number of products sold.  
For example :  
Products X has 10 complaints and amount sold is 40.  
Products Y has 4 complaints and amount sold is 20.  
Products Z has 15 complaints and amount sold is 75.

Ratio :  
X= 10/40  
1/4  
Y= 4/20  
1/5  
Z= 15/85  
3/17

Z<Y<X`  
Thus even though the product Z has highest number of complaints as compared to X and Y, the product amount is sold more than either of them which means this product is being used by larger population which inturn means the product is good enough that more and more people are willing to try. And also if the product is able to impact a bigger set of people this adds a wide diversity of people using product.

b.  
The attribute type is Ratio Attribute. This is a quantitative type of attribute i.e it can be counted or calculated. It calculates the differences between the measurements but unlike Interval Attribute the attribute can contain null/zero value.

4.  
Soyabean Dataset  
a.Total Number of Instances: 683  
b.Number of Attributes: 36  
c.Possible Values for crop-hist attribute: 4  
=== Confusion Matrix ===  
a b c d <-- classified as  
28 9 16 12 | a = diff-lst-year  
11 70 37 47 | b = same-lst-yr  
20 32 104 63 | c = same-lst-two-yrs  
9 40 56 113 | d = same-lst-sev-yrs

d.Attribute: (Nom) seed-tmt  
=== Confusion Matrix ==  
a b c <-- classified as  
228 72 5 | a = none  
87 133 2 | b = fungicide  
21 11 3 | c = other  
e. Date: 11 July 1988

5.  
Number of records detecting outliers : 231  
Number of records detecting Extreme Values : 370