Εξόρυξη Δεδομένων, Φροντιστήριο 21/02/2013, Αντικείμενο: Προεπεξεργασία

1. Data cleaning

1.1. Missing values

Για τη συμπλήρωση των χαμένων τιμών που υπάρχουν στα δεδομένα έχουν προταθεί διάφορες τεχνικές. Συμπληρώστε τις 5 ελλείπεις πλειάδες, με τις εξής μεθόδους:

i. με χρήση καθολικής μεταβλητής
ii. με τη μέση τιμή του γνωρίσματος
iii. με τη μέση τιμή του γνωρίσματος για τις πλειάδες της ίδιας κλάσης.

	Class	Attribute_1	Attribute_2	Attribute_3
Item_1	0	119		34
Item_2	0	143	5	
Item_3	0	128	7	42
Item_4	1	153		35
Item_5	1	121	3	

1.2. Smoothing by binning

Έστω τα αριθμητικά δεδομένα: 8, 10, 10, 12, 15, 18, 19, 20, 21, 23, 24, 28.
Αφού τα διαχωρίσετε σε bins ίσου βάθους=4, να ομαλοποιήσετε τα δεδομένα χρησιμοποιώντας:

(α) το μέσο όρο των bins
(b) τα όρια τιμών των bins.

---

2. Data Integration

2.1 Correlation Analysis on nominal data

Suppose that the ratio of male to female students in the Science Faculty is exactly 1:1, but in the Pharmacology Honours class over the past ten years there have been 80 females and 40 males. Is this a significant departure from expectation?

It is given that concerning the χ^2 table, the "critical value" for {$p = 0.05$ and *1 degree of freedom*} is 3.84.

	Female	Male	Total
Observed numbers (O)			

2.2 Correlation Analysis on numerical data

Suppose two stocks A and B have the following values in one week: