

WEEK 10 HOMEWORK

INSTRUCTIONS

- Every learner should submit his/her own homework solutions. However, you are allowed to discuss the homework with each other (in fact, I encourage you to form groups and/or use the forums) – but everyone must submit his/her own solution; you may not copy someone else's solution.
- The homework will be peer-graded.
- The homework grading scale reflects the fact that the primary purpose of homework is learning:

Rating	Meaning	Point value (out of 100)
4	All correct (perhaps except a few details) <u>with</u> a deeper solution than expected	100
3	Most or all correct	90
2	Not correct, but a reasonable attempt	75
1	Not correct, insufficient effort	50
0	Not submitted	0

Question 1

The breast cancer data set breast-cancer-wisconsin.data.txt from <http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/> (description at <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>) has missing values.

1. Use the mean/mode imputation method to impute values for the missing data.
2. Use regression to impute values for the missing data.
3. Use regression with perturbation to impute values for the missing data.
4. (Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using (1) the data sets from questions 1,2,3; (2) the data that remains after data points with missing values are removed; and (3) the data set when a binary variable is introduced to indicate missing values.