STAT5003

Week 7 : Missing data and class imbalance

Dr. Justin Wishart

THE UNIVERSITY OF SYDNEY

functions covered

- functions
    - `mice::md.pattern`
    - `mice::marginplot`

Missing data

Mechanism for missing data

- **M**issing **C**ompletely **A**t **R**andom (**MCAR**)
  - E.g. Let's say we run a survey and some people don't want to give their age in the questionnaire, but this does not relate to any other variable (including their party preference)
- **M**issing **A**t **R**andom (**MAR**)
  - The missingness of the data in a variable is not related to the variable but related to some other variables.
  - E.g. In a survey, if people from a lower socioeconomic status may be less willing to provide salary information (but we know their SES status).
  - In other words. The salary value is missing not entirely due to the salary but conditional on the socioeconomic status.
- **M**issing **N**ot **A**t **R**andom (**MNAR**)
  - When data is not MCAR or MAR.
  - The missingness of the data is due to the value of the variable itself even after accounting for other variables
  - E.g. In a polling survey, if liberal voters are less likely to disclose how they intend to vote.

Identifying different types of missingness

Unfortunately, there is no statistical method to determine the mechanism of missingness

- You can guess the mechanism of missingness by knowing something about the data, and something about the data collection method
- To see if the data is MAR, can try to fit a classification model to predict missingness

Dealing with missing values

- For categorical data, "missing" can be a category.
  - For example, in a survey poll, if someone does not want to disclose who they want to vote for, can be in the category "undecided"
- Delete data with missing value. Two options.
  - Omit the variable with missing data.
  - Omit the observation with missing data.
  - Drawbacks are that you might be throwing away valuable information, or inadvertently introduce bias into the data
- Impute i.e. fill in the missingness.
  - Can replace missing values with the mean of the ones observed for that feature

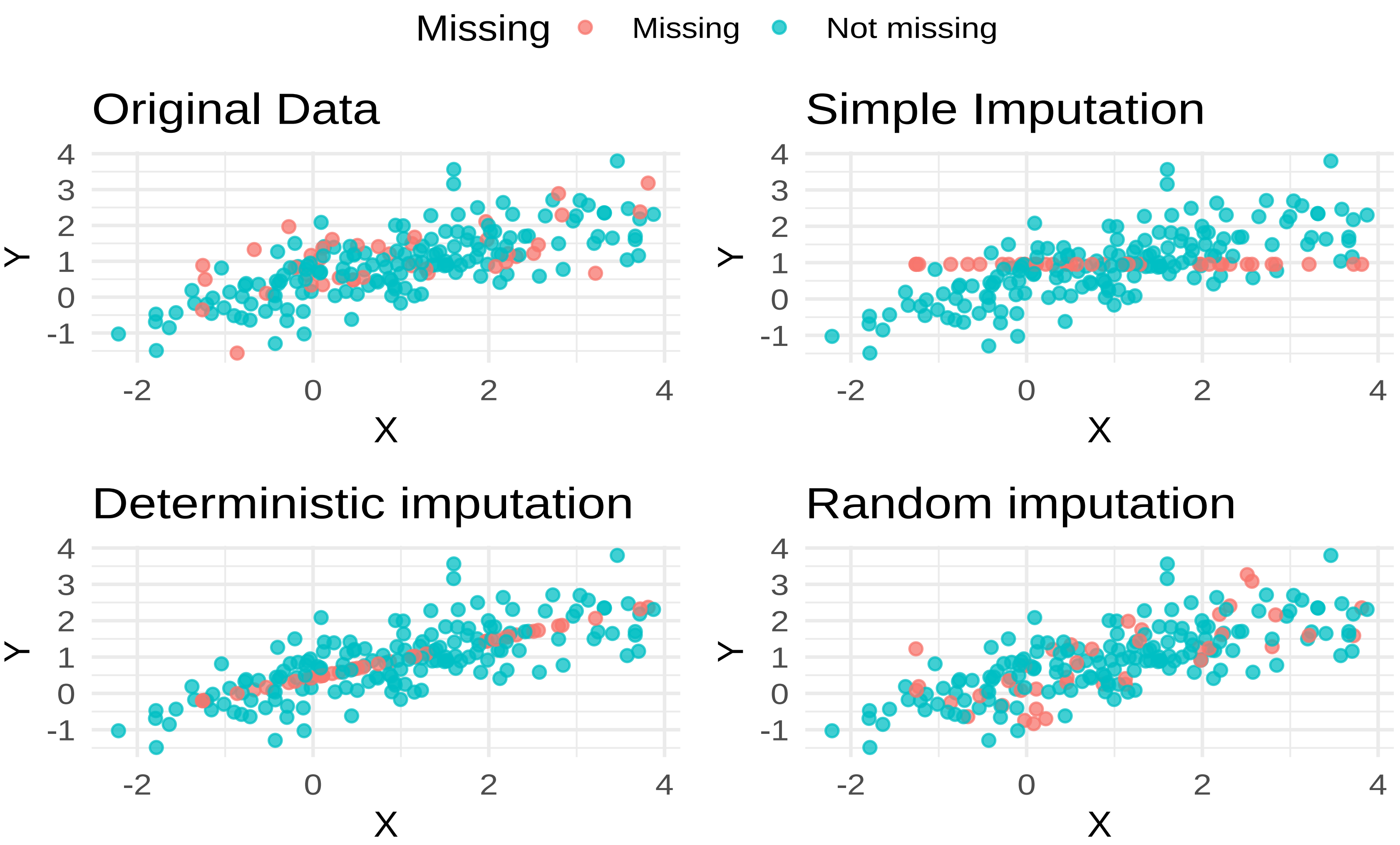
Single imputation

- Single imputation replace the missing value with a single value.
- Examples:
  - Replace the missing values of a feature with the mean/median value of that feature
  - Use a predictive method for filling in the missing values e.g. regression trees, kNN
  - Replace the missing value with the last observed value for that feature
  - With single imputation, once the missing data is added back, it is treated as valid observed data, hence the uncertainty in the missing value data is lost.

Multiple imputation

- Multiple imputation accounts for uncertainty in the imputation process.
- Generally follows three steps:
  - Impute the data k times (this can be done using a single imputation method)
  - Perform analysis (e.g. regression) on each of the k imputed data sets
  - Pool the k results together
- Multiple Imputation by Chained Equation (MICE) is a popular method
  - See van Buuren and Groothuis-Oudshoorn (2011)

Basic impute, deterministic imputation, random imputatation



Other practical suggestions

- It is highly recommend that you visualize your data to look for patterns of missingness
- Be wary of variables with high proportion of missing data. However, this might not be a problem if imputation is applicable and performs well.
- Some algorithms can cope with missingness (e.g. decision trees) and so you may not need to do imputation
- If you believe the pattern of missingness is informative, you can include it as a dummy variable

R packages for dealing with missing values

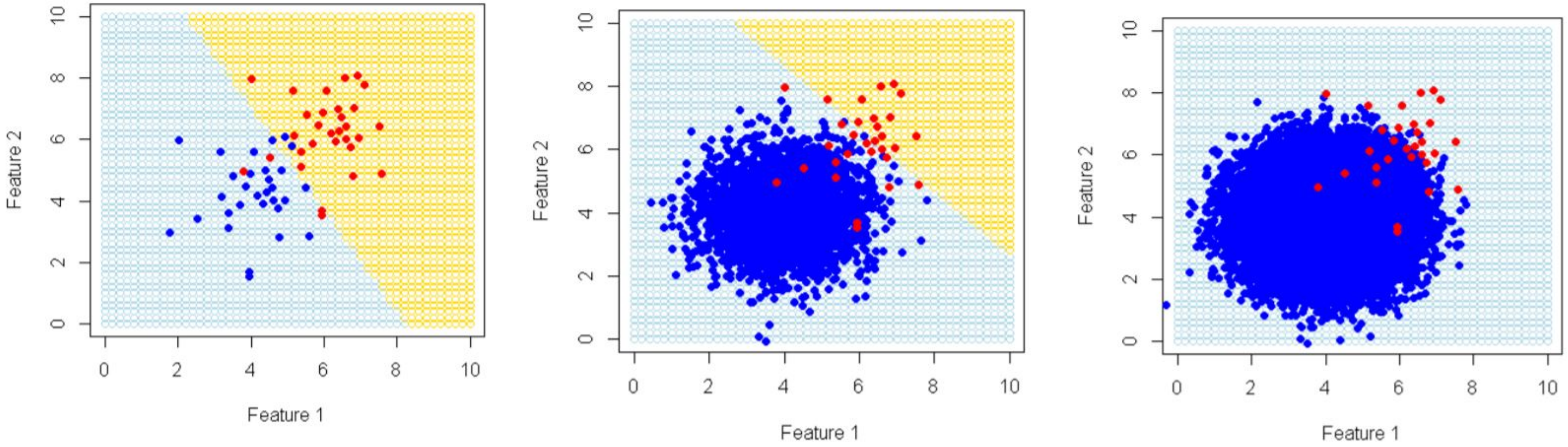
- Impute (Bioconductor package)
  - KNN imputation, written for microarray data
- MICE
  - Multiple imputation
- missForest
  - Uses Random Forest (in upcoming tree based module) to predict the missing values
  - Can be used for continuous and categorical data
- Amelia II
  - Multiple imputation

Class imbalance

Class imbalance

- Let's say we have a classification problem to detect credit card fraud, but only 1% of transactions are fraud.
- If you use accuracy as the metric to optimize, then just by classifying every transaction as not-fraud will get you to 99% accuracy!

Inspect a SVM



- Notice the negative (blue) class swamps the classifier
- Boundary moves and disappears favouring only the negative class.

Use different metric

- F_1 score

$$F_1 = \frac{2TP}{2TP + FP + FN} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Area Under Curve (AUC) for the ROC
- Cohen's Kappa
    - $\kappa = \frac{p_0 - p_e}{1 - p_e}$
    - Compares expected to observed accuracy

What is Cohen's Kappa metric?

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

- $p_o =$ observed agreement of classes = Overall Accuracy
- $p_e =$ hypothetical expected agreement

Consider the simple case of binary classification (two classes) with n cases. Call the two judges, judges A and B.

- Judge A could denote the observed class assignment labels.
- Judge B could denote the predicted class assignment labels

$$p_o = \text{Proportion of Agreement between judge } A \text{ and } B = \text{This is the accuracy.}$$

Cohen's Kappa: Empirical chance measure

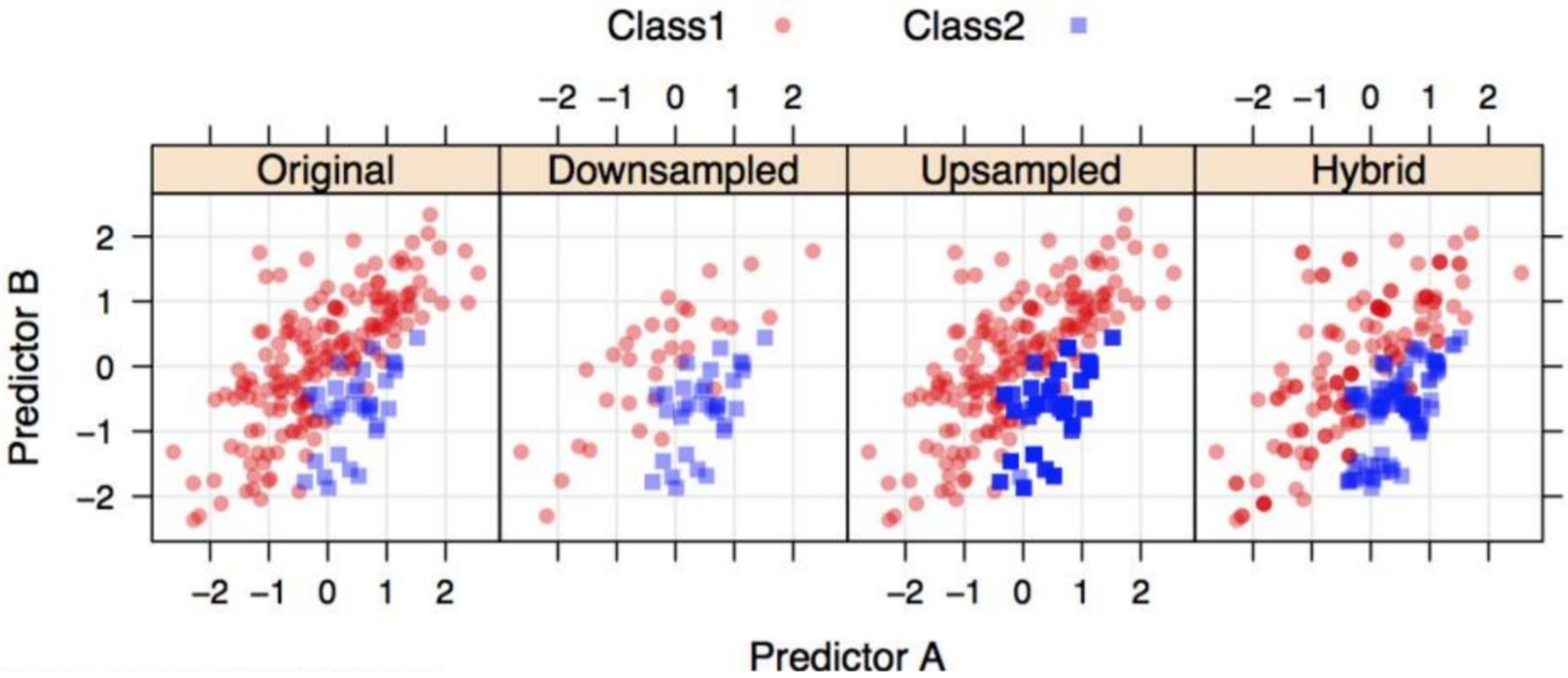
The other metric p_e is the empirical chance these judges give the classification by chance.

- i.e. Computes the change that Judge A and Judge B agree on a randomly picked element.
  - Chance that either both judges classify positive class or both judges classify negative class.
  - Assuming they pick at random.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 980000  20000
##          1      0      0
##
##                Accuracy : 0.98
##                  95% CI : (0.9797, 0.9803)
##     No Information Rate : 0.98
##     P-Value [Acc > NIR] : 0.5019
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.00
##             Specificity : 0.00
##          Pos Pred Value : 0.98
##          Neg Pred Value :  NaN
##              Prevalence : 0.98
##          Detection Rate : 0.98
##    Detection Prevalence : 1.00
##       Balanced Accuracy : 0.50
##
##        'Positive' Class : 0
##
```

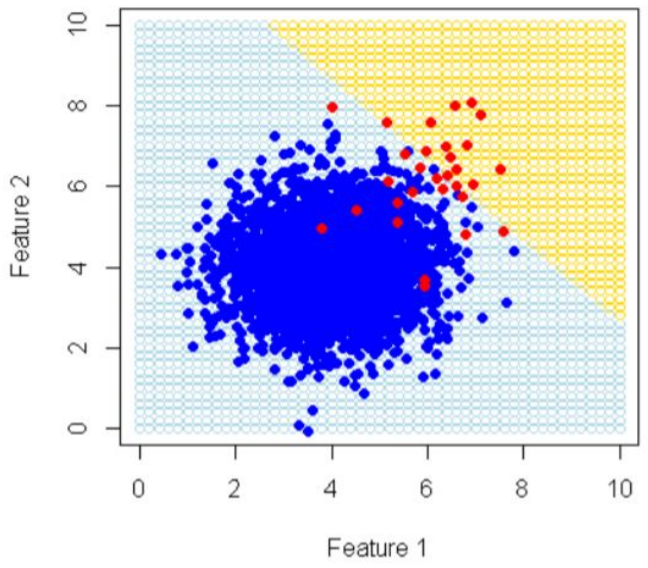
```
##    pred
## obs      0      1
##   0 980000      0
```

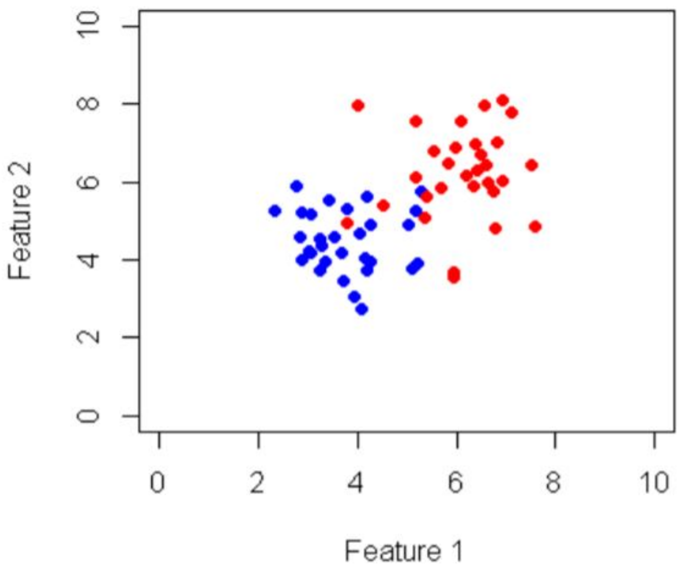
Alternatively, modify the input data



- Random up-sampling
- Disadvantages:
  - creates duplicated and/or artificial instances
  - can introduce bias and/or noise to the original data

Down-sampling





- Advantage is it does not introduce duplicates and/or artificial instances
- Disadvantages:
  - Not all data points are used.
  - Potentially removing useful information.
- Better choice for data with very high class imbalance.

Create synthetic samples of the minority class

- Synthetic Minority Over-sampling Technique (SMOTE) is a popular algorithm
- It creates synthetic samples from the minority class by:
    - Finding the k-nearest-neighbours for minority class observations
    - Randomly choosing one of the k-nearest-neighbours, then using it to create a similar but random new observation
- Be careful you split your data into training/validation before doing any oversample/SMOTE. Otherwise, you will leak information from training to validation data set.
- The `R` package `DMwR` implements SMOTE
    - See Torgo (2010)

References

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC. URL: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). "mice: Multivariate Imputation by Chained Equations in R". In: *Journal of Statistical Software* 45.3, pp. 1-67. URL: https://www.jstatsoft.org/v45/i03/.