Data Science Project: Epigenetic Classifier for Twin Zygosity

Zexi Liu, Haoze Xia, Atefeh Zamani, Ni Zhang, Tianyu Zhou

Supervised by:

Dr. Jiadong Mao and Prof. Michael Kirley

Overview

What is zygosity of Twins and why are they important?

- Monozygotic twin (MZ) or identical twin
- Dizygotic twin (DZ) or fraternal twin

Importance:

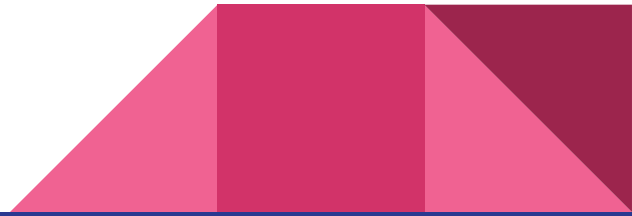
- Cheaper classification method compared to traditional questionnaires from hospitals
- It helps better understand disorders that’s correlated with twins, e.g. spine disorders.

Overview

Aim 1: Test the provided classifier by van Dongen et al. [2021] on a new data set to check its performance.

Aim 2: Train a new classifier based on machine learning methods to obtain a classifier with better performance compared to classifier provided by van Dongen et al. [2021].

We **finished Aim 1** within the time frame of the 1st semester and will present results in the following section



Group Members

Zexi Liu - Review Review Literature

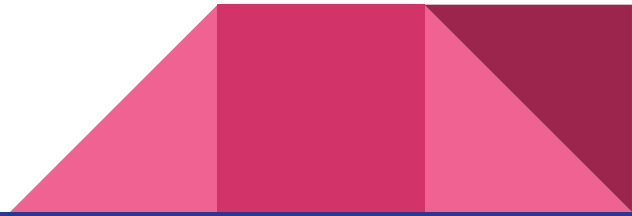
Atefeh Zamani - Review Literature

Haoze Xia - Coding

Ni Zhang - Coding

Tianyu Zhou - Coding

Agile development cycle



Host Organisation

Twins Research Centre of Epidemiology and Biostatistics UoM

- Dr. Shuai Li and Prof. John Hopper

What is the Data Science Problem?

- Data linkage of medical data gives more opportunity for research in recent years

- The organisation wants to explore machine learning techniques to this sudden available data, and this project is to explore options.

DataSets

- In van Dongen et al. [2021],
  - One dataset to train (1957 Samples), Five datasets for testing (4485 Samples).

	N Total	N MZ Twins	N DZ Twins	N Family Members
NTR	1957	924	1033	237
E-Risk	1164	470	694	-
FTC	1708	559	1149	-
TwinksUK	492	395	97	-
BSGS	**356**	**134**	**222**	**257**
NTR	765	564	201	

Table 1: Datasets analysed in van Dongen et al. (2021)

Difficulties and Classifiers

Difficulties:

- Dimension of the datasets:
  - Big datasets (450K columns)
  - Reducing the dimension of the data
    - Meta-Analysis
      - Monozygotic twinning is associated with a persistent DNA methylation profile in adult somatic tissues.
      - This MZ-signature comprises 834 CpG sites.

Classifiers:

- Classifier 13: Classify based on their zygosity (MZ or DZ)
- Classifier 14: Classify MZ twins from the rest (DZ twins and family members).

Datasets

- The datasets we have access to
  - E-Risk (1658 Samples), 4 other datasets (1921 Samples)

	N Total	N MZ Twins	N DZ Twins	N XZ Twins	N Family Members
E-Risk (NCBI-GSE105018)	1658	860	608	190	-
AMDTSS (NCBI-GSE100227)	479	132	132	-	215
BSGS (NCBI-GSE56105)	**614**	**135**	**223**	**-**	**256**
Denmark (NCBI-GSE73115)	180	94	86	-	-
E-MTAB (E-MTAB-1866648)	648	240	408	-	-

Table 2: The datasets that will be used in this research

Findings - Summarized Results

- Classifier 13 & 14 have produced similar results apart from the proportion of non-twins correctly predicted.
- Since classifier 14 was trained from the dataset containing family members while classifier 13 wasn't, classifier 14 would therefore have a better performance when predicting non-twins.
- The average AUC is around 0.65 and the proportion of DZ twins correctly predicted is low.

Average values of the five testing datasets							
	AUC	proportion MZ twins correctly predicted	proportion DZ twins correctly predicted	proportion non-twins correctly predicted	proportion MZ twins incorrectly predicted	proportion DZ twins incorrectly predicted	proportion non-twins incorrectly predicted
Classifier 13	0.645	0.762	0.481	0.440	0.238	0.519	0.560
Classifier 14	0.647	0.665	0.571	0.724	0.335	0.429	0.276

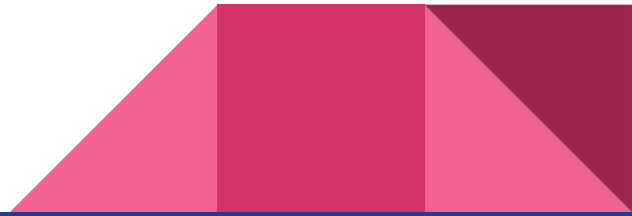
Findings - Detailed Results

Classifier 13	N total	N MZ twins	N DZ twins	N non-twin	AUC	proportion MZ twins correctly predicted	proportion DZ twins correctly predicted	proportion non-twins correctly predicted	proportion MZ twins incorrectly predicted	proportion DZ twins incorrectly predicted	proportion non-twins incorrectly predicted
NTR - testing	1100	522	339	239	0.751	0.843	0.469	0.444	0.157	0.531	0.556
BSGS - nature	606	131	220	255	0.796	0.916	0.450	0.302	0.084	0.550	0.698
E-Risk	1658	852	612	0	0.728	0.684	0.683	NaN	0.316	0.317	NaN
BSGS	614	135	223	256	0.784	0.904	0.404	0.313	0.096	0.596	0.688
Denmark	180	94	86	0	0.544	0.596	0.570	NaN	0.404	0.430	NaN
AMDTSS	479	132	132	215	0.648	0.742	0.545	0.567	0.258	0.455	0.433
E-MTAB	648	240	408	0	0.522	0.882	0.201	NaN	0.118	0.799	NaN

Classifier 14	N total	N MZ twins	N DZ twins	N non-twin	AUC	proportion MZ twins correctly predicted	proportion DZ twins correctly predicted	proportion non-twins correctly predicted	proportion MZ twins incorrectly predicted	proportion DZ twins incorrectly predicted	proportion non-twins incorrectly predicted
NTR - testing	934	520	334	80	0.766	0.808	0.572	0.625	0.192	0.428	0.375
BSGS - nature	606	131	220	255	0.799	0.901	0.468	0.451	0.099	0.532	0.549
E-Risk	1658	852	612	0	0.739	0.621	0.740	NaN	0.379	0.260	NaN
BSGS	614	135	223	256	0.774	0.637	0.798	0.848	0.363	0.202	0.152
Denmark	180	94	86	0	0.563	0.755	0.395	NaN	0.245	0.605	NaN
AMDTSS	479	132	132	215	0.667	0.750	0.530	0.600	0.250	0.470	0.400
E-MTAB	648	240	408	0	0.492	0.564	0.390	NaN	0.436	0.610	NaN

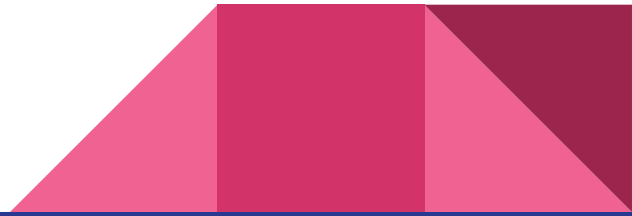
Data Science Pipelines

- Data engineering
  - Dimension reduction - from 450k variables
- Machine learning
  - SVM, Adaboost, Random forests, etc.
- Information output
  - Result tables & Figures



Plan for Coming Semester

- Dimension reduction methods
  - PCA, Auto encoder, etc.
- Training classifiers using different machine learning methods
  - Support vector machine (SVM)
  - Logistic regression
  - Bayesian algorithm
  - Random forests
  - Adaptive boosting (Adaboost)
  - Stochastic gradient boosting
  - Deep learning
  - Stacking and Embedding Models



Key Challenges

1. The size of the dataset
    - Find a new method to load the dataset in local machine
    - Cloud services
        - Spartan
2. High dimensional dataset
    - Do some research and testing different machine learning algorithms
3. Unbalanced training data
    - Re-structure the training data
        - Over-sampling for minority class
    - Random Forest

Key tasks, milestones and roles

Thanks for watching!