SOMS4101: Data Analysis and Visualisation

The why and how of machine learning in medical sciences

Adam Dunn

THE UNIVERSITY OF SYDNEY

Learning objectives

At the completion of this module, you will be able to:

1. **Select appropriate methods to answer questions about complex datasets**
2. **Understand how machine learning is used in health and medical research**
3. Apply scripts in R to analyse complex datasets
4. **Effectively report on analyses using data visualisation techniques**

Outline

Machine learning

- Supervised and unsupervised machine learning
- Classifiers and methods

Feature representations

- Transforming data
- Dimensionality reduction approaches

Training and evaluation

- Cross-validation, over-fitting

Reporting

- Reporting of accuracy measures

Why should I learn about machine learning?

- Nearly ***every field*** has had ML introduced, from history/literature through to medicine/public health
- Researchers now have ***access*** to much larger volumes of data; changes the feasibility of manual vs automatic (consider one biomarker vs finding a gene signature)
- Traditional high-impact ***medical and science journals*** are increasingly willing to review and accept ML methods
- Don't need to know the algorithms, only really need to ***know when and how*** to apply ML methods to applications

Outline

Machine learning

- Supervised and unsupervised machine learning
- Classifiers and methods

Feature representations

- Transforming data
- Dimensionality reduction approaches

Training and evaluation

- Cross-validation, over-fitting

Reporting

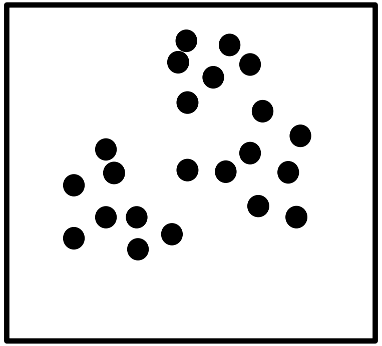
- Reporting of accuracy measures

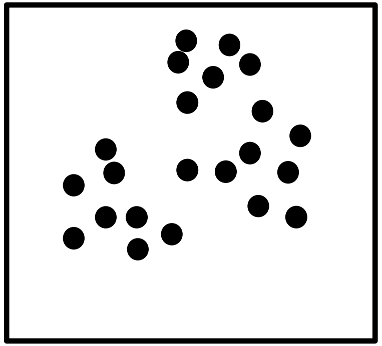
Preliminaries

- ***Example:*** A single case from a dataset, such as a patient, an image, or a document
- ***Label:*** The ground truth or predicted value assigned to an outcome of interest, such as a diagnosis (in statistical analysis, these are the values of the response/outcome variables)
- ***Features:*** The set of characteristics that are used to train and predict an outcome (in statistical analysis, these are factors or explanatory variables); an example has lots of features
- ***Classification task:*** The overall problem structure related to predicting an outcome from a set of features
- ***Training data:*** A subset of examples from the dataset, which is used to train a model (more on validation data later)

Supervised and unsupervised machine learning

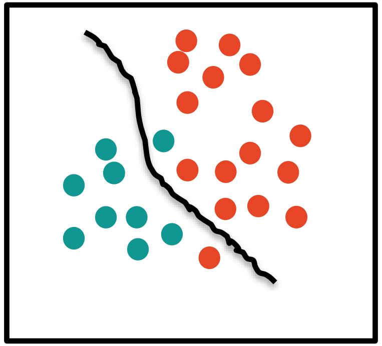
- **Supervised machine learning:** where we have a set of ground truth *labels* for each *example* in our dataset.
  - *Classification:* predicting classes/categories including binary; methods include logistic regression, naïve Bayes, support vector machines, decision trees, random forest, neural networks, …
  - *Regression:* predicting continuous values; methods include linear regression, elastic net regression, LASSO, support vector regressors

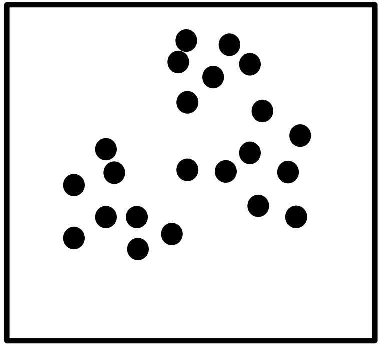




Supervised and unsupervised machine learning

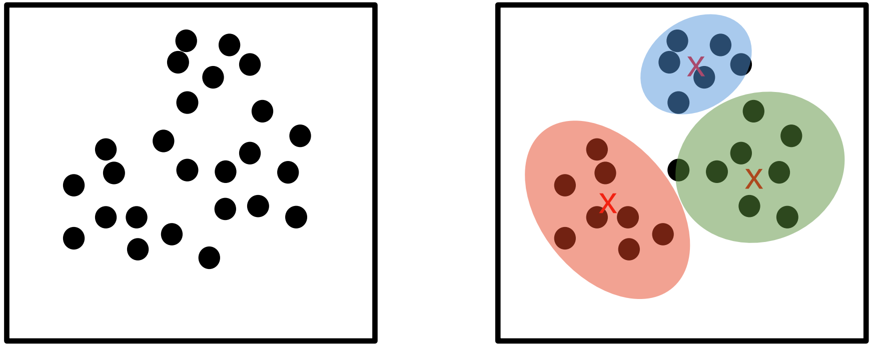
- **Supervised machine learning:** where we have a set of ground truth *labels* for each *example* in our dataset.
  - *Classification:* predicting classes/categories including binary; methods include logistic regression, naïve Bayes, support vector machines, decision trees, random forest, neural networks, …
  - *Regression:* predicting continuous values; methods include linear regression, elastic net regression, LASSO, support vector regressors





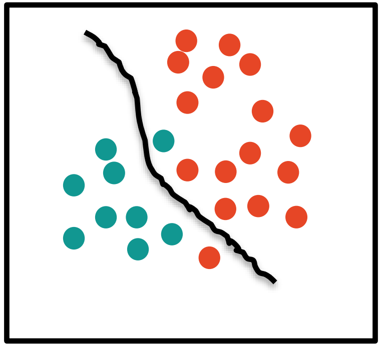
Supervised and unsupervised machine learning

- **Unsupervised machine learning:** where we do not *necessarily* have an outcome of interest and want to characterise our data.
  - *Clustering:* finding natural groups; k-means, agglomerative clustering, affinity propagation, …
  - *Dimensionality reduction:* where the number of dimensions of the feature space is reduced by removing uninformative features or mapping features into fewer dimensions



The classification task

- ***Classification task:* Given a set of examples, use the features of those examples to predict a set of labels.**
  - Are there models that would help us create a clinical decision making tool? (random forest; ***interpretability***)
  - Which features are most informative? (ablation testing)
  - Which feature representation and algorithm/method works best?



Representing features – transforming data

Binary data: the presence or absence of a characteristic, such as the presence or absence of a gene in a patient
Categorical data (nominal): features that can take on three or more values where the order is not important
Categorical data (ordinal): features that can take on three or more values where the order is important
Numerical data: a feature that is defined by a continuous value

- Assign values 0, 1, 2, … to categorical data within one feature?
- One-hot encoding, where each value becomes a new binary feature?
- R has encoders built in, and if you are interested you can ask
  (in python these are in sklearn)

Representing features – dimensionality reduction

When we have datasets that involve thousands to millions of features, it is hard to build robust models, especially when the data are very sparse.

– ***Feature selection:*** methods for selecting informative features; includes measuring information gain or entropy across features without knowing the label, or iteratively test for accuracy with/without each feature

– ***Mapping/projection:*** methods for mapping all features into fewer dimensions, including PCA, LDA, autoencoders (& visualisation methods like t-SNE and UMAP also do this)

Your turn (5 minutes)

Find any published example where dimensionality reduction is used to support visualisation of data, ***post the image into the discussion forum*** as a reply with a URL, name the method used to map it into two dimensions, and (optionally) comment on what the figure shows and how it contributes to answering the main aims of the research in the article.

Outline

Machine learning

- Supervised and unsupervised machine learning
- Classifiers and methods

Feature representations

- Transforming data
- Dimensionality reduction approaches

Training and evaluation

- Cross-validation, over-fitting

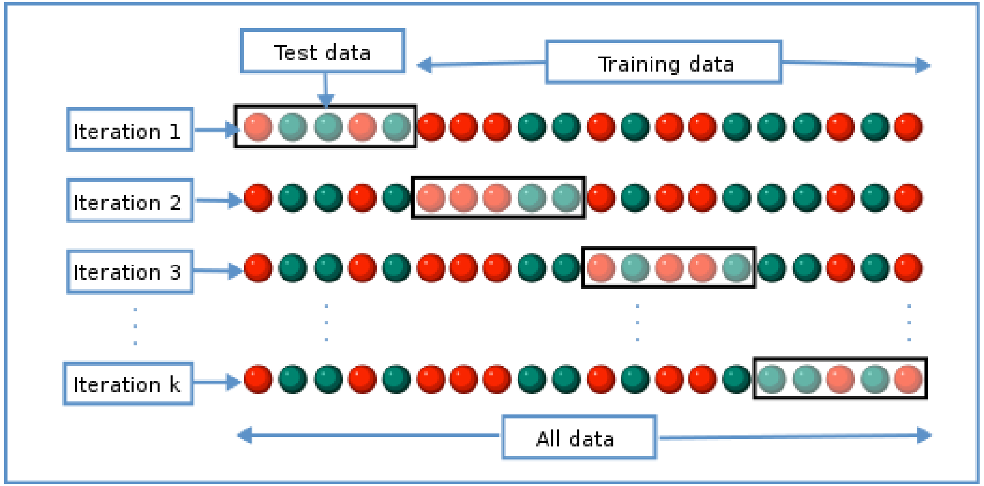
Reporting

- Reporting of accuracy measures

Training and testing – cross-validation

Cross validation: a kind of resampling process used to avoid over-fitting; and a way to evaluate how well a model should perform in unseen data.

***K-fold cross validation*:** the most common form used in practical machine learning; performance often reported by averaging across all folds.



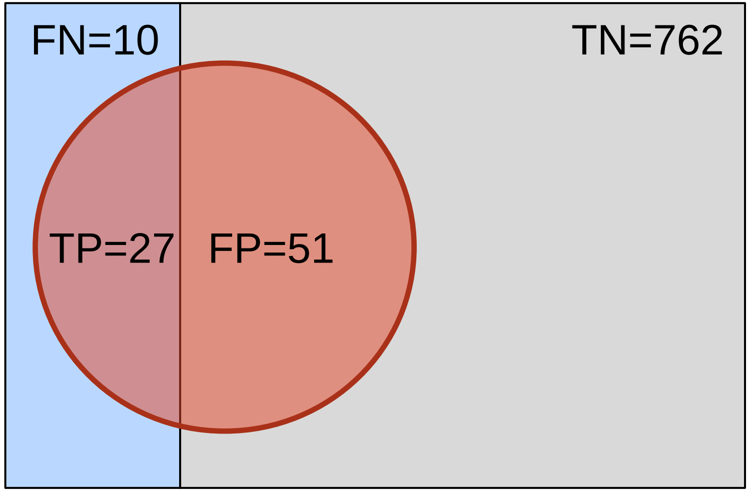
https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Reporting – performance measures

Accuracy: the number of correct predictions divided by the total number of predictions = **94%,** but this is an ***unbalanced dataset***

From 840 participants in the study cohort, 37 had a diagnosis of pancreatitis in their EHR. We want to use previously prescribed medications to predict pancreatitis from combinations of prescribed medications.

Model predicts pancreatitis in 78 participants (27 TP, 51 FP, 762 TN, 10 FN) = 93.9% accurate but 51 taken off medications they might need?

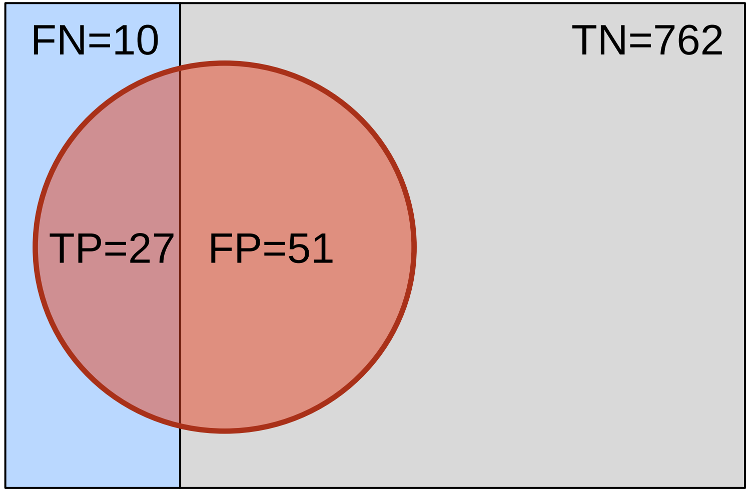


Reporting – performance measures

F1-score: the harmonic mean of precision and recall, and equivalent to 2*(precision*recall)/(precision+recall) = **46%**; where precision = TP/(TP+FP)= 34.6%; and recall = TP/(TP+FN) = 73.0%

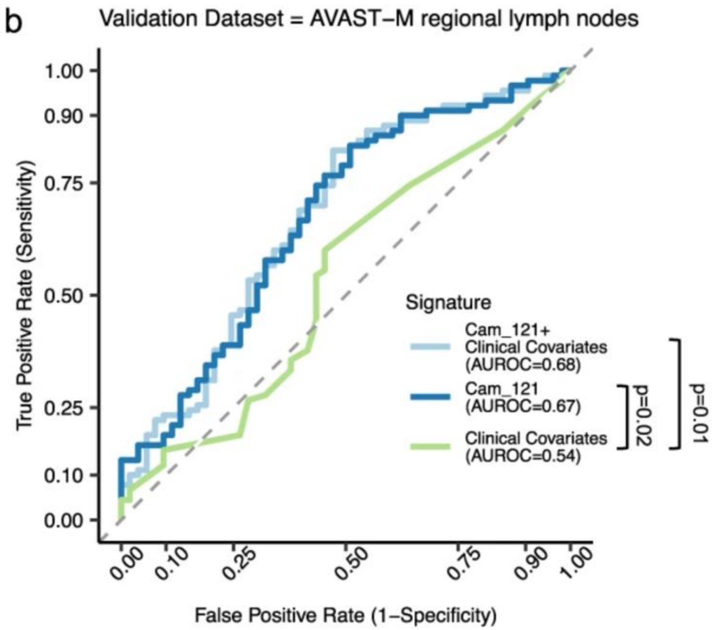
From 840 participants in the study cohort, 37 had a diagnosis of pancreatitis in their EHR. We want to use previously prescribed medications to predict pancreatitis from combinations of prescribed medications.

Model predicts pancreatitis in 78 participants (27 TP, 51 FP, 762 TN, 10 FN) = 93.9% accurate but 51 taken off medications they might need?



Reporting – performance measures

AUC: Area under the receiver operating characteristic (ROC) curve, which is also called the c-statistic, and not to be mistaken with area under the precision-recall curve, which is different.



We can use the different points on the curve to examine the trade-off between recall (sensitivity) and specificity (*not* precision)

Outline

Machine learning

- Supervised and unsupervised machine learning
- Classifiers and methods

Feature representations

- Transforming data
- Dimensionality reduction approaches

Training and evaluation

- Cross-validation, over-fitting

Reporting

- Reporting of accuracy measures