Lec10- Feature Selection & Dimension Reduction

Recap

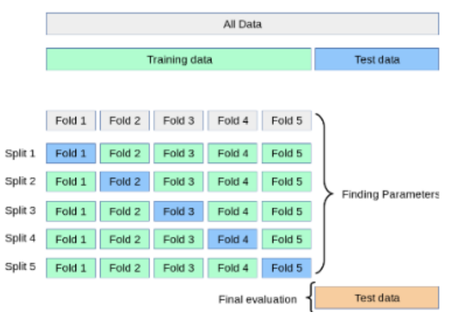
Binary classification evaluation

Actual clusters	Predicted clusters: The same	Predicted clusters: Not the same
The same	a(TP)	b(FN)
Not the same	c(FP)	d(TN)

- Accuracy = (a + d) / (a + b + c + d)
- Precision = a / (a + c) — Among the predicted positive samples, how many of them are correct
- Recall = a / (a + b) — How many actual positive samples are predicted to be positive
- F1-score = 2 * precision * recall / (precision + recall) — The weighted average of precision and recall
- Balanced accuracy = 0.5 * (a / (a + b) + d / (c + d)
- Value is not absolute. Context matters.

Cross-validation

a procedure to measure the performance of models



e.g. 5-fold cross-validation

- 10 data points: P1 - P10
- 5-fold: P1-2, P3-4, P5-6, P7-8, P9-10
  The grouping can be random
- Procedure
  1. P1-2's results based on the model from P3-10
  2. .....
  3. P9-10's results based on the model from P1-8
  4. Averaging

Multi-class evaluation

- Still using accuracy, precision, recall, F1 score and so on — considering each class as a binary classification problem
- aggregate multiple values into one value
  - Macro-average — The low-performance of small classes will show up in Macro-average
  - Micro-average

Evaluate clustering

- In clustering, correct as long as two similar cells are in the same cluster
- evaluate a pair of cells
- confusion matrix

Actual clusters	Predicted clusters: The same	Predicted clusters: Not the same
The same	a(TP)	b(FN)
Not the same	c(FP)	d(TN)

Why feature selection & dimension reduction

Bio-data can be huge, noise, unrelated, and duplicated

- **huge**
  - gene expression profile: 25,000 genes
  - single-cell RNA-seq with 13.7 million cells in human body
  - Excel storage: 13700000 * 25000 = 1198750MB =1.2TB
- **duplicated**
  - Irrelevant genes: we do not have to include them in our analysis
  - Highly correlated genes: we do not have to include all of them
  - Some genes are complementary: form pathways. Combine them into one value may be more useful

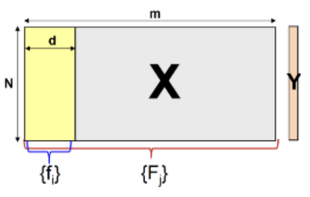
Benefit

- **Data compression** — Efficient storage and retrieval
- **Improve prediction performance** — Remove unrelated inputs
- **Understand the prediction results** — What genes are related to the cancer prediction
- **Facilitate data visualization (Lec 17)** — 25000D to 2D, understand the distance between cells visually

Feature Selection/ Extraction

Thousands to millions of low-level features: select/extract the most relevant one to build better, faster, and easier to understand learning machines.

- **Feature Selection**
  - Choose the best subset genes from all the genes
  - **Feature ranking**
  - **Feature subset selection: Filter and Wrapper**
- **Feature Extraction**
  - Extract new features by linear or non-linear combination of the original features — New feature = Gene 1 + Gene 2
  - New features may not have physical interpretation / meaning (usually for non-linear)
  - e.g. PCA, SVD, Isomap, LLE, CCA



d<<m
using label Y -> supervised
without label Y -> unsupervised

Feature Ranking

Build better, faster, and easier to understand learning machines

- **Eliminate useless features (distracters).**
- **Rank useful features.**
- **Eliminate redundant features.**

Discover the most relevant features w.r.t. target label

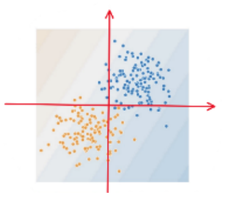
Find genes that discriminate between healthy and disease patients

How to measure which ones are useful?

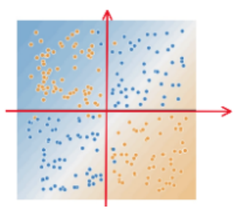
- **Correlation between feature and class** — Weight VS Gender = 0.714
- **Mutual information I(i)** — The higher I(i), the attribute is more related to the class
- **Fisher score F** — The higher F the attribute is more related to the class

Issues of individual features ranking

- **Relevance VS Usefulness**
  - Relevance does not imply usefulness
  - Usefulness does not imply relevance
- **Leads to the selection of a redundant subset** — k best features != best k features; Height & Weight < Height & Major
- **A variable that is useless by itself can be useful with others** — Salary: occupation +age
- **example (2 features x & y)**



even only use 1 feature, classification is fine



only consider 1 feature (x or y).
all the data are mixed, we cannot separate with 1 feature

Propose a classifier: x*y

- x*y > 0 blue
- x*y < 0 yellow

Subset Feature Selection

Filter

- **Classification performance is not involved in the selection loop**
- **Variance thresholds:** — Features with a higher variance contain more useful information
  - age
  - height
- **Information gain** — Features should be different

	G1	G2	G3	G4	Cancer
S1	10	2	6	8	Yes
S2	10	3	7	8	Yes
S3	10	4	8	6	No
S4	10	5	9	5	No

example — only present G3 & G4

- G1: do not have variance, not useful => deleted
- G2 & G3: too similar => delete one
- choose G4 by variance thresholds

Wrapper

- **Using the classification performance to guide selection**
- **Computational expensive**
- **Recursive feature elimination**
- **Sequential feature selection**

	G1	G2	G3	G4	Cancer
S1	10	2	6	8	Yes
S2	10	3	7	8	Yes
S3	10	4	8	6	No
S4	10	5	9	5	No

example

1. No feature
2. Find the first best feature using cross-fold validation
3. Add the second feature using cross-fold validation
4. .....
5. Until the new feature does not improve the performance