BMEG3105: Data analytics for personalized genomics and precision medicine — Fall 2023 -Lecture 15

Lecturer: Yu LI (李煜) from CSE Liyu95.com, liyu@cse.cuhk.edu.hk
Date: 25 October 2023
Topic: multi-Omics Overview

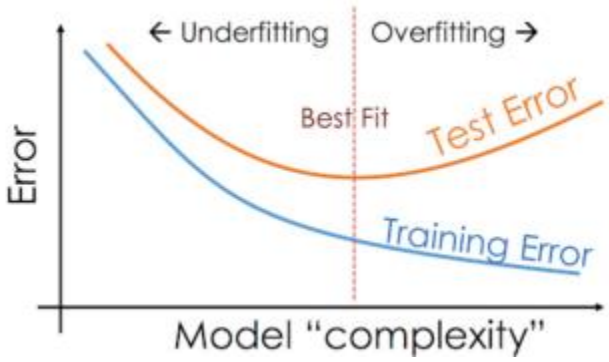
Lecture Outline

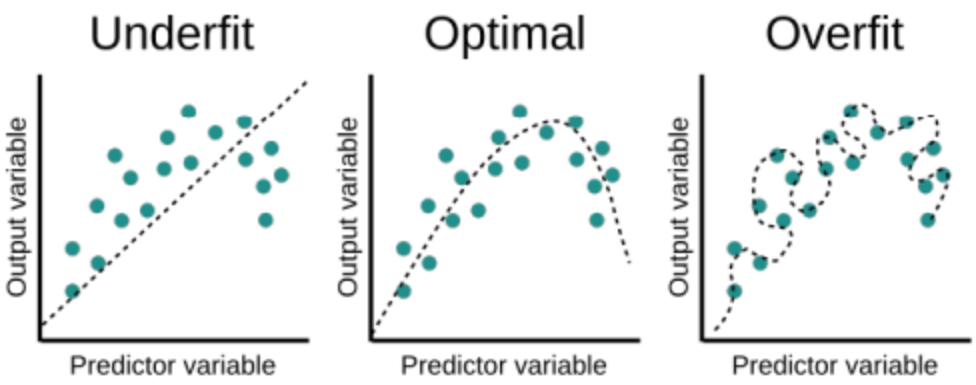
1. Model overfitting
2. Multi-omics overview
3. Statistical testing

Model Overfitting

Definition

- Statistically: the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably
- Machine learning: the method is more complex than the problem and too complicated that it may fit the noise in the data, such that it can perform well on the training dataset but does not perform well on the testing dataset,

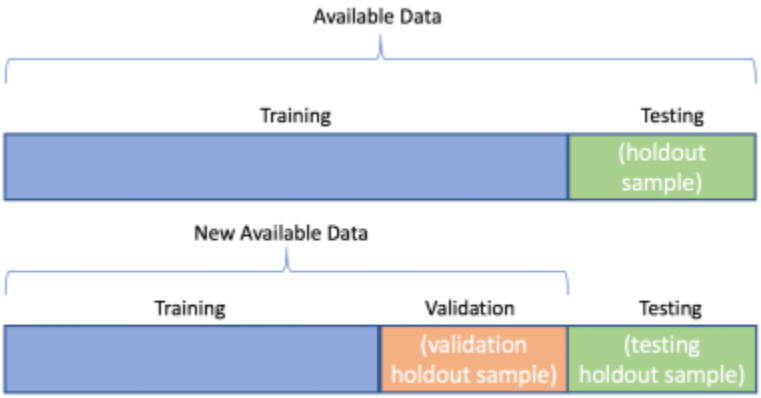




Evaluation of model and detecting overfitting

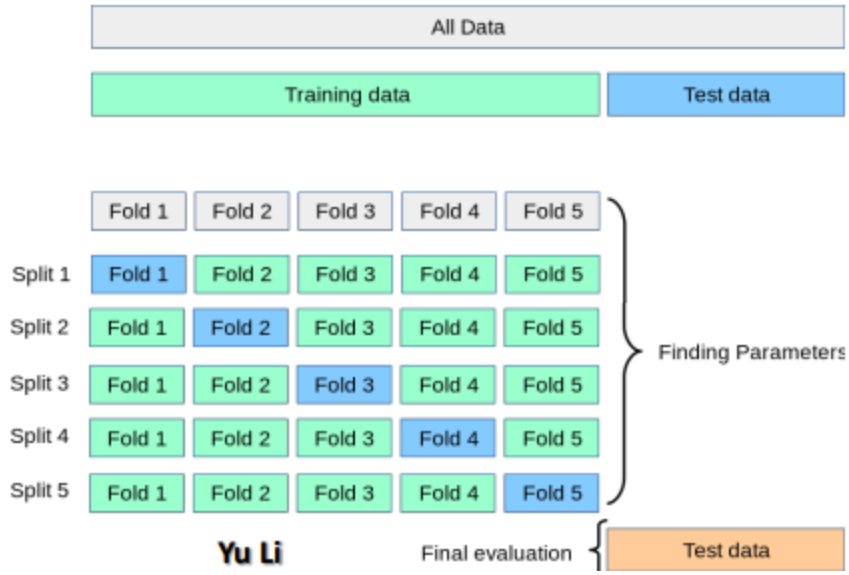
1. Train-validation-test split (more preferred in big data)

- Train: 70%
- Validation: 15%
- Test: 15% (the model should not be trained on the testing data)



2. **Cross-validation (fine if model is light)**

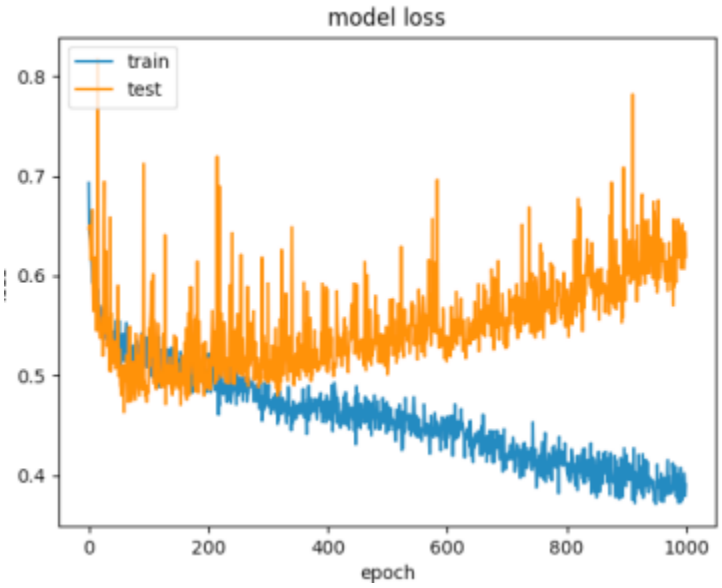
- 5-fold validation
- Leave-one-out
- Reliable evaluation
- Expensive



Evaluation criteria

1. **Loss function**

- The difference between training loss and validation loss
- The performance may be OK even if overfitting



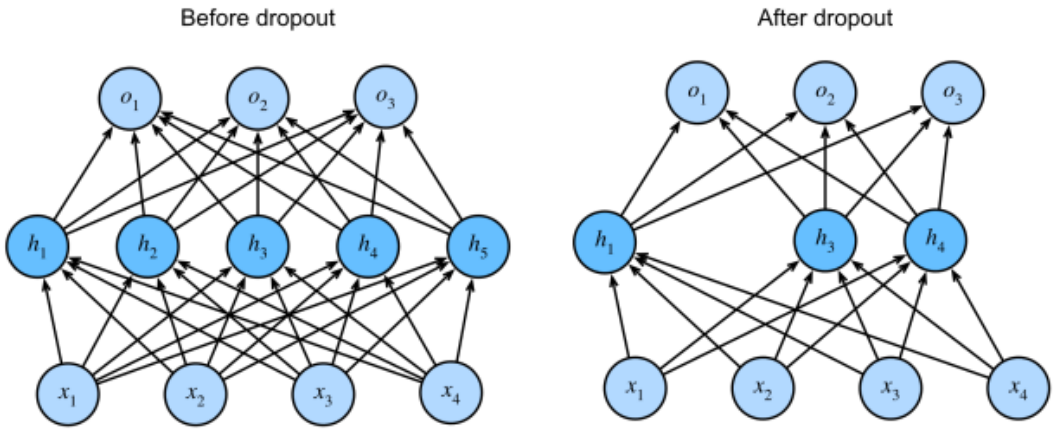
2.Performance

- Precision, recall, F1 score
- Make sure all of them have reasonable values (no bug), handle all the bugs first
- Performance on training dataset has the increasing trend
- The difference between training dataset and validation dataset may increase over the time

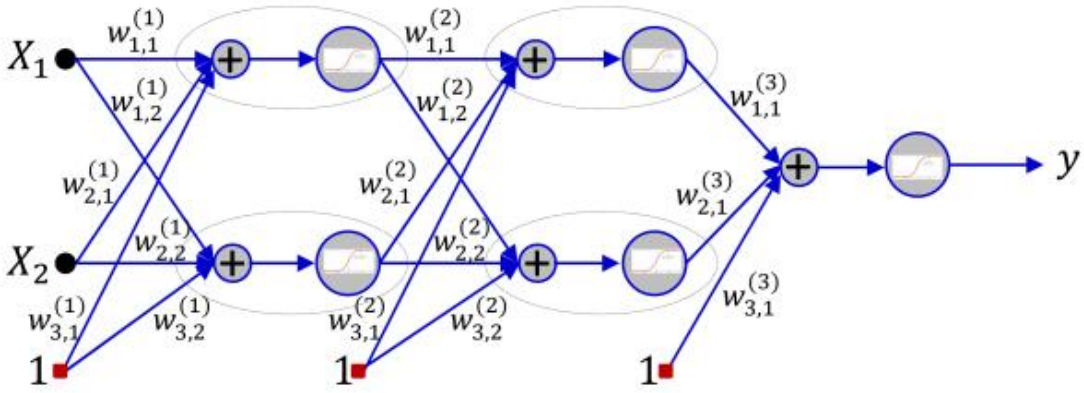
Source of overcomplexity and solutions

- Data Too little, not reflect the true distribution (Noise which the things that mislead the model to learn a wrong distribution are not necessarily bad) ➔ add some noise, data augmentation, transfer learning (add some noise to the label, if the noise is little and uniformly distributed)
- Model too large, too many useless parameters ➔ Transfer Learning
    - Transfer learning :
    - 1. The data size that the model has seen is enlarged
    - 2. The model has already learned useful features
    - 3. We may only fine-tune a few final layers
- Connectivity too strong, co-adaptation ➔ don't rely on one node or feature too much or discard node and edges stochastically during training /Add noise to the model to increase smoothness
  (rate usually 0.5-0.8)

Dropout



- Parameter value range Too large, model too flexible ➔ Panelize over the large weight values (weight decay regularizer)



Weight decay Regularizer

$$L' = L + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

During GD update:

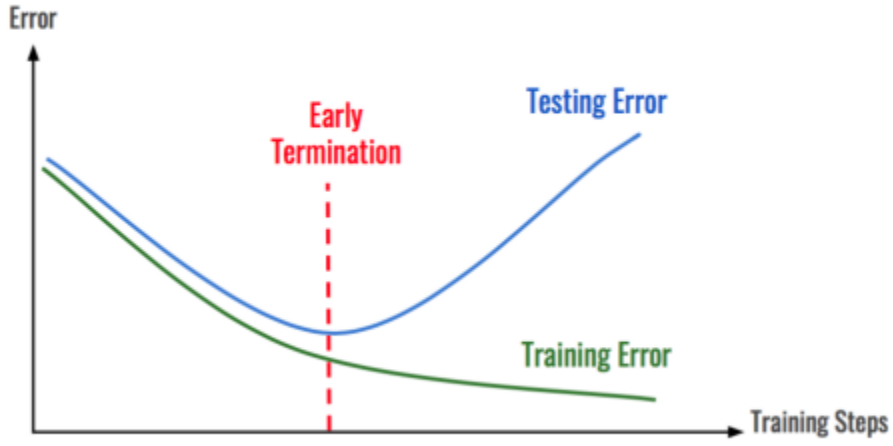
$$w_{ij}^k = w_{ij}^k - \Delta w_{ij}^k - \alpha\lambda w_{ij}^k$$

$$\Delta w_{ij}^k = \alpha \frac{\partial Div(Y,d)}{\partial w_{ij}^{(k)}}$$

The algorithm dec
time training the n

mics **Yu Li**

- Training time Too long, tend to overfitting The error may not reflect the accuracy

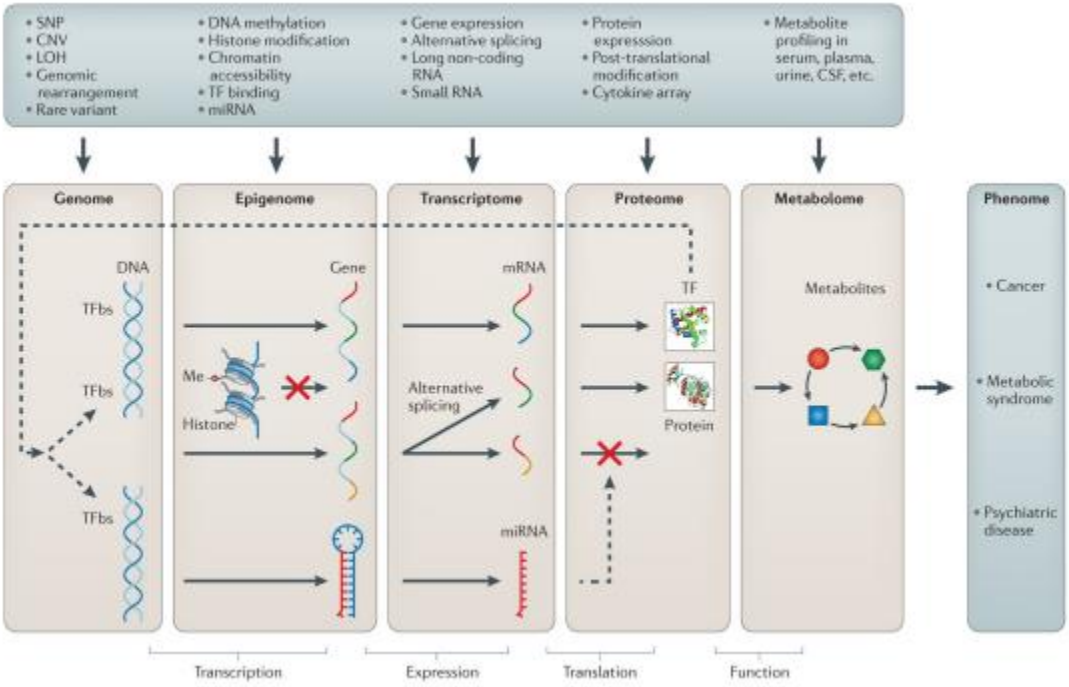


Multi-omics

- A longitudinal big data approach for precision health

- Omics: Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms and to study biological entities in large scale

Types



- Genome pipeline
- Epigenome pipeline
- Transcriptome pipeline

Key-message

- Multi-omics data analysis can be very tedious nowadays but the core techniques are the same
- Sequence alignment and comparison
- Dimension reduction and visualization
- Clustering and classification

Statistical Testing

- Differential gene expression analysis: Statistical analysis to discover quantitative changes in expression levels between experimental groups. For a given gene, whether the gene expression difference is significant, other than due to natural random variation

1.T-test

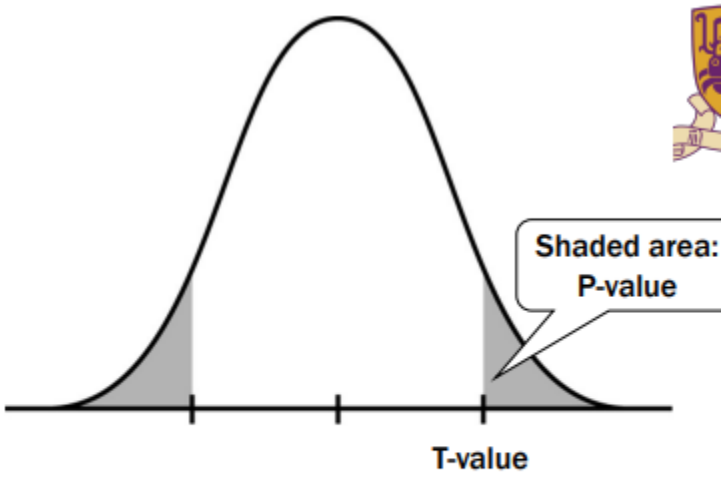
- A kind of standard statistical test procedure

- The purpose of t-test: Is there a significant difference between two sets of data?
- Calculate a test statistic based on the mean and variance of the data
- Test statistic follows a Student's t-distribution

P-value

- the probability that the result from the data occurred by chance
- Along with test statistic, t-value
- The smaller p-value is, the more confident we are
- Example:

  Unpaired two tailed t test: p-value smaller than 0.5 ➔ the two gene expression are not significantly different from each other

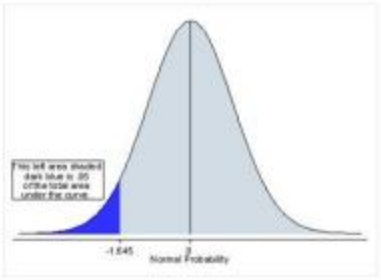


$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{15.09 - 13.00}{\sqrt{\frac{146.69}{11} + \frac{18.22}{10}}} = \frac{2.09}{3.894} = 0.54$$

Types of t-test

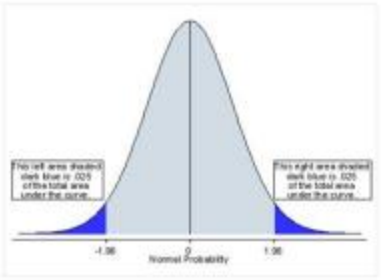
- One-tailed test VS two-tailed test
- Two-tailed test: different or the same
- One-tailed test: greater, larger, smaller, at least

One-tailed t-test



A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

Two-tailed t-test



A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

The formula to calculate t-value ,the formula to translate t-value to p-value can be different

But the t-test procedure is the same

If the p-value is smaller than 0.5, the genes' gene expression are significantly different.

Gene enrichment analysis

Question: how to identify pathways related with type-II diabetes?

2.Testing association

- For a given pathway, we have genes related to it or not
- For the type-II diabetes, we have genes related to it or not
- If the pathway is related to type-II diabetes
- The number of genes (not) related to both should be high (a,d)
- The number of genes related to just one should be low (b, c)

	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213

How large they should be to say they are related confidently?-

Fisher's exact test

- Fisher's exact test is a statistical significance test used in the analysis of contingency tables
- P-value can be calculated exactly from the table
- Recall t-test. We calculate a t-value

- Based on a distribution, we get the p-value
- Suppose pathway and type-II diabetes are independent

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

p = 0.5802 > 0.05 ➔ This pathway is not related to type-II diabetes