BMEG3105 Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI (李煜) from CSE

Wednesday, 12 October 2022

Lecture 11: Dimension reduction

Review of last lecture

- Dimension reduction can
  - Compress data size for efficient storage and retrieval
  - Improve prediction performance by removing unrelated input
  - Simplify model for easier interpretability and understanding
  - Facilitate data visualisation
- Feature selection (choose the best subsets features), methods include
  - Feature ranking
  - Filter
    - Do not involve classification performance
    - Apply filter criteria (e.g. only features with variance higher than threshold stay)
    - Allow different features remain, i.e. information gain
  - Wrapper
    - Use classification performance to guide selection
    - Computational expensive
    - Recursive feature elimination
    - Sequential feature selection
      - Greedy algorithm
      - Do NOT guarantee global best combination of features selected
- Feature extraction

New course arrangement

- Add a few more topics about data (e.g. neural networks, overfitting)
- Do NOT compress the biological application part (module 2)

Today's content

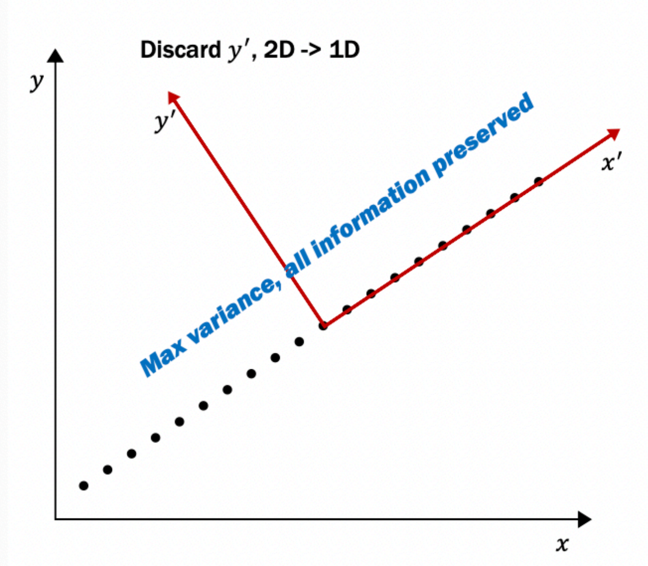
1. Dimension reduction: PCA
2. Neural networks

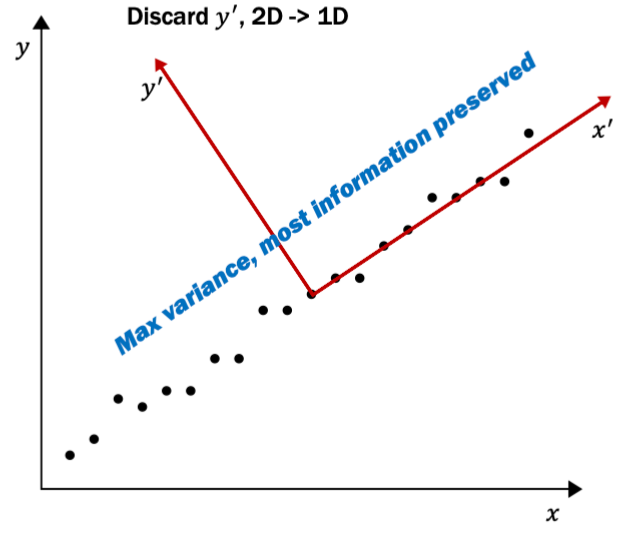
1.1 Dimension reduction

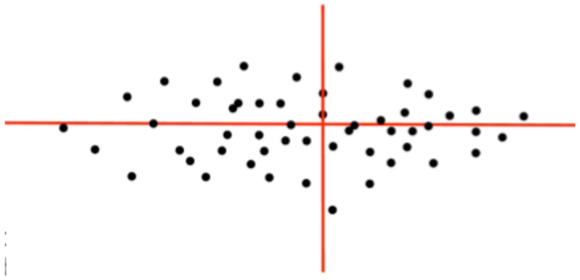
- Extract new features by linear or non-linear combination of original features (e.g. new feature = gene 1 + gene 2)
- New features may NOT have physical interpretation / meaning
- Methods include: PCA, SVD, Isomap, LLE, CCA...

1.2 Principal components analysis (PCA)

- Transform vector space such that 1st dimension captures maximal variance while 2nd dimension (orthogonal to 1st dimension) captures maximal residual variance
- 1st dimension may capture enough information that it is possible to ignore other axis







1.3 Procedures of PCA

1. Normalise each feature to make the average of each feature 0
2. Calculate the covariance matrix of the normalised matrix
3. Find the eigenvectors and eigenvalues of covariance matrix
4. Select some eigenvectors with largest eigenvalues (principal components)
5. Project the normalised data onto the eigenvectors' direction

1.4 Calculating example of PCA

Step 1: Original data matrix

X1	1	1	1
X2	2	2	2
X3	3	3	3

=>

X', normalised data matrix

X1	-1	-1	-1
X2	0	0	0
X3	1	1	1

Step 2: sigma, covariance matrix

$$\Sigma = \frac{1}{n-1} X'^T X' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Step 3: V, eigenvectors; lambda, eigenvalues

$$\Sigma * V = \lambda * V$$

$$|\Sigma - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)^3 + 1 + 1 - (1-\lambda) - (1-\lambda) - (1-\lambda) = 0$$

$$\lambda = 3 \; or \; \lambda = 0$$

$$\Sigma * V = \lambda * V$$

$$(\Sigma - \lambda I) * V = 0$$

$$\begin{bmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{bmatrix} * \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0$$

$$\lambda_1 = 3 \qquad \lambda_{2,3} = 0$$

$$V_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix} \qquad V_{2,3} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Step 4: P, principal components (V1, V2)

$$\begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix}$$

Step 5: X hat, projection of X' on P

$$\hat{X} = X'P = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} = \begin{bmatrix} -\sqrt{3} & 0 \\ 0 & 0 \\ \sqrt{3} & 0 \end{bmatrix}$$

Final PCA result (each column represents new coordinate on each axis)

X1	1	1	1
X2	2	2	2
X3	3	3	3

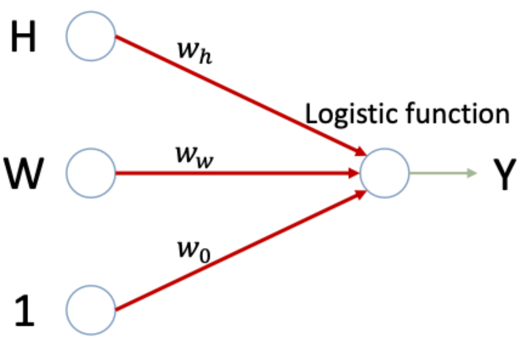
=>

X1	$-\sqrt{3}$	0
X2	0	0
X3	$\sqrt{3}$	0

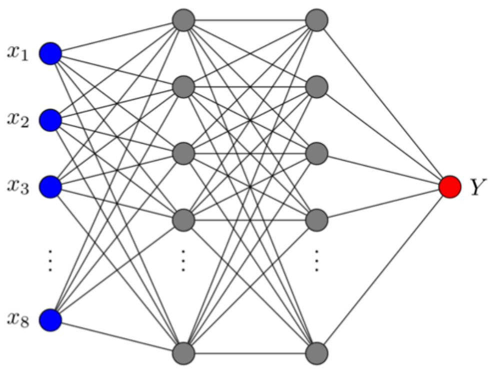
2.1 Neural networks: Evolution of simple logistic regression

- Relationships among real-life data can be much more complex than linear combination
- Using simple logistic regression may cause
  - Underfitting
  - Low model capacity
- Deep neural networks can be regarded as a collection a logistic regression by
  - Increase the number of nodes
  - Increase the number of layers
  - Add non-linear function
- Fully-connected layers is a general function approximator, according to universal approximation theorem

Simple logistic regression



Fully-connected deep neural networks



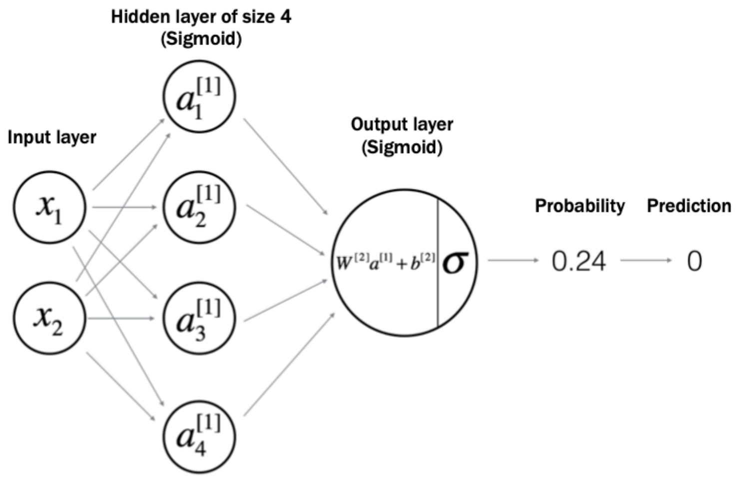
2.2 Composition of functions

- Each internal node
  - Represent a new extracted feature by linear or non-linear combination
  - May NOT have physical interpretation / meaning
  - Act as a function
    - Receive and integrate input from nodes in previous layer
    - Fire output to the nodes in next layer
- Stacking multiple layers of internal nodes result in functions of functions
- Analogy: dog = f(head, hoof, shape...), hoof = f(pixel color, pixel value...), ...

2.3 Number of parameters

- For fully-connected deep neural networks, every node is connected to all nodes (and bias) in previous layer and in next layer
- One parameter is trained for every edge in the graph

An example with 17 parameters to be trained



$$a_1 = \frac{1}{1+e^{-(w_{11}*x_1+w_{21}*x_2+b_1)}}$$
$$a_2 = \frac{1}{1+e^{-(w_{12}*x_1+w_{22}*x_2+b_2)}}$$
$$a_3 = \frac{1}{1+e^{-(w_{13}*x_1+w_{23}*x_2+b_3)}}$$
$$a_4 = \frac{1}{1+e^{-(w_{14}*x_1+w_{24}*x_2+b_4)}}$$
$$Y = \frac{1}{1+e^{-(w_1*a_1+w_2*a_2+w_3*a_3+w_4*a_4+b)}}$$

Potential project

Project title: *Data preprocessing for the gene expression matrix*

- Data collecting and merging (if needed)
- Exploration
- Visualisation
- Data cleaning
- Dimension reduction
- Get distance matrix
- Perform classification/clustering
- Performance evaluation

Python package for feature selection and dimension reduction

Scikit-learn

- https://scikit-learn.org/stable/modules/feature_selection.html
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Resource

- *Machine learning: A probabilistic perspective*: Chapter 1.4.3, 12.2-12.5

Uncovered topics

- The curse of dimensionality
- How to get the PCA algorithm
- SVD
- PCA VS SVD VS ICA