Nama : Fasya Hanifah

NIM : 1103200149

RANGKUMAN PRINCIPAL COMPONENT ANALYSIS (PCA)

Example graph : This graph was drawn from single cell RNA-seq. There were about 10.000 transcribed genes in each cell. Each dot represents a single-cell and its transcription profile. We see that in this graph the blood cellsform one cluster that's different from pluripotent cells which is different from neuronal cells and dermal or epidermal cells. So that question is How does transcription 10.000 genes get compressed to a single dot on a graph?

Here the answer PCA (Principal Componen Analysis) is a method for compressing a lot of data into something that captures the essence of the original data.

An Introduction to Dimensions

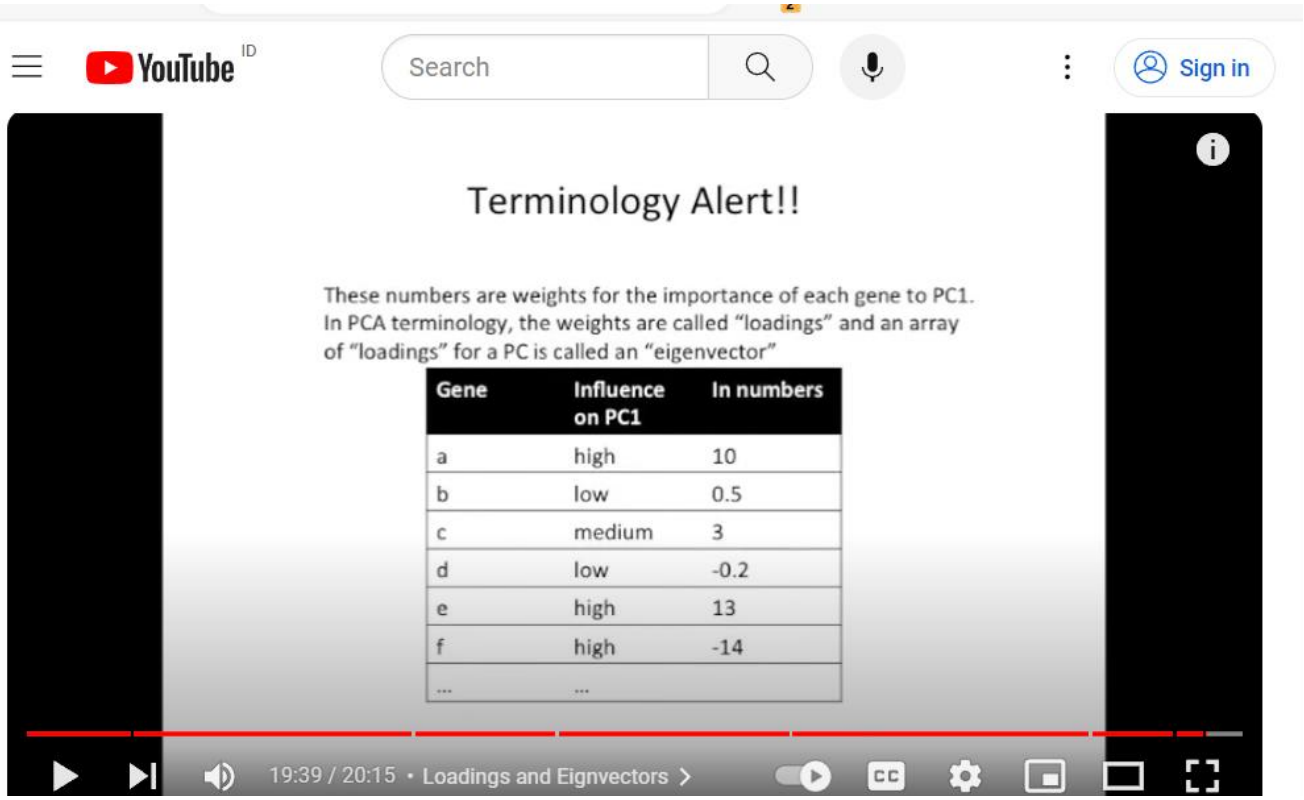
- 1-Dimension (1-D) = a number line, example (A pretend RNA-sew data set for a single cell). There are two output (A uniform distribution and A non uniform distribution) (if we have one data)
- 2-D = a normal graph, in two dimensional graphs we have two axes. We might see two expression in the two cells is correlated and not correlated. If the expression is correlated its mean genes that are highly transcribed in cell one are also highly transcribed in cell two and genes that are lowly transcribe in cell one are also lowly transcribed in self to and if the expression is not correlated its mean gene is highly transcribed in cell one that doesn't tell us anyting about whether its highly or lowly transcribed in cell two. (if we have data from two cells)
- 3-D = a fancy graph that has depth = have three separate axes (if we have data from three cells)

PCA takes a dataset with a lot of dimensions and flattens it to 2 or 3 dimenssions so we can look it. PC1 the direction of the most variation in gene expression, PC 2 the 2nd most variation in gene expression.

- If we had 2 cell = PC1 captures the direction where most of the variation is, PC2 captures the direction with the 2nd most variation.
- If we had 3 cells = PC1 span the direction of the most variation, PC2 span the direction 2nd most variation, PC3 span the direction of the 3rd most variation
- If we had 4 cells = PC1 span the direction of the most variation, PC2 span the direction 2nd most variation, PC3 span the direction of the 3rd most variation, PC4 = span the direction of the 4th most variation

HOW WE PLOT CELL? The length and direction of PC1 is mostly determined by the circled genes. Genes with little influence on PC1 get value close to zero, and genes with more influence get numbers further from zero.

For identify key genes we want to find out which genes had a big influence in putting dermal cells on the left neural cells on the right, we could look at the influence scores in PC1 and if we wanted to find put which genes help distinguish blood cells from neural and dermal cells, we could look at the influence score in PC2.



Terminology Alert!!

These numbers are weights for the importance of each gene to PC1. In PCA terminology, the weights are called "loadings" and an array of "loadings" for a PC is called an "eigenvector"

| Gene | Influence on PC1 | In numbers |
|---|---|---|
| a | high | 10 |
| b | low | 0.5 |
| c | medium | 3 |
| d | low | -0.2 |
| e | high | 13 |
| f | high | -14 |
| ... | ... | |

Nama : Fasya Hanifah

NIM : 1103200149

STATQUEST DECISION AND CLASSIFICATION TREES

Decision trees are part of the foundation Machine Learning because Decision tree are quite simple, very flexible, and pop up in a very wide variety of situations. In general Decision Tree makes a statement and decision based on whether or not that statement is true or false. If Decision Tree classifies things into categories it's called a Classification Tree and if Decision Tree predicts numeric values it's call regression tree.

In the classification tree it combines numeric data with yes/no data its okay to mix data types in the same tree. Numeric thresholds can be different for the same data. Sometime in classification tree you see True and False labels, sometimes you don't. It's not big deal. Terminlogy in classification Tree :

- The top of tree = Root node
- Below the top = Internal Nodes or Branches
- Below the branches = Leaf nodes, leaves have arrows pointing to them but no arrows pointing away from them

One of the most popular methods is call Gini Impurity, but there are also fancy sounding methods like Entropy and Information gain. Overfit data if we hard to have confidence that it will do a great job making predictions with future data.

Nama : Fasya Hanifah

NIM : 1103200149

STATQUEST THE K NEAREST NEIGHBORS ALGORITHM

THE K NEAREST NEIGHBORS ALGHORITHM is a super simple way classify data. If you already had a lot of data that defined these cell types. If we want to decide which tyoe of cell this cells ex Stem Cells, Blood Vessel Cells, and Fat Cells. There is step by step to decide it :

Step 1 : Start with a dataset with known categories

Step 2 : Add a new cell, with unknown category to the PCA plot

Step 3 : We classify the new cell by looking at the nearest annotated cells

HEATMAPS = drawn with the same data and clustered using hierarchical clustering

If K=1, we just look at the nearest cell, and that cell is light blue, if K = 5 we look at the 5 nearest cells, which also light blue, if K = 11 (7 nearest neighbors are light blue, 4 are light green) so still light blue. If the new cell is right between two categories (K is odd we can avoid a lot of ties, if still tied vote we can flip coin or decide not to assign the cell a category)

Machine Learning/Data Mining Terminology = The data used for the initial clustering (data where know the categories in advance) is called “training data”

A few thoughts on picking a value for “K” =

- You have to try a few values before settling on one do this by pretending part of the training data is unknown
- Low values for K (like K=1 or K=2) can be nosy and subject to the effects of outliers
- Large values for K smooth over things, but you don’t want K to be so large that a category wit only a few samples in it will always be out voted by other categories.