TJML x TJ Bioinformatics

Beginner Group

2/23/22

Welcome!

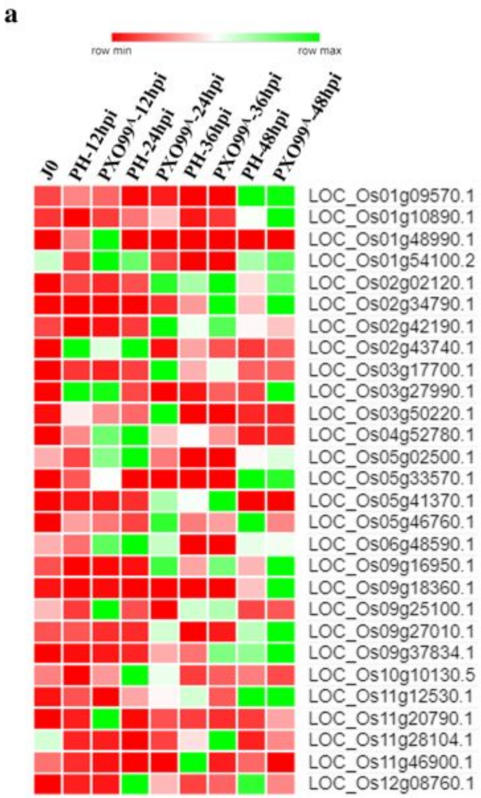
- Brief lecture relevant to what we’re doing today
- Guided workshop!
  - Essentially doing a research project with us!
  - Get help and advice, but you will be doing a good part of the coding

TJ Bioinformatics Society

- **Wednesday 8B**
- Cover topics relating to bioinformatics
  - DNA, RNA, protein, etc. and how we can use this with CS
- What we’ve done:
  - Bioinformatics (ML) **workshops** and **lectures**
  - Coding **competitions and games** with **monetary prizes**
  - Beginner-friendly!
- **Guest lectures** (from you all) have opened!
  - Join our Facebook group (search for it)
  - tjbioinfo.netlify.app

Project

- Using gene expression data from cancer patients with one of 5 types of tumors (RNA-Seq)
- https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq
- Train an ML model to predict tumor type from expression data
- **Omics**
  - Data relating to gen<u>omics</u>, prote<u>omics</u>, etc.
  - This is an example of omics data

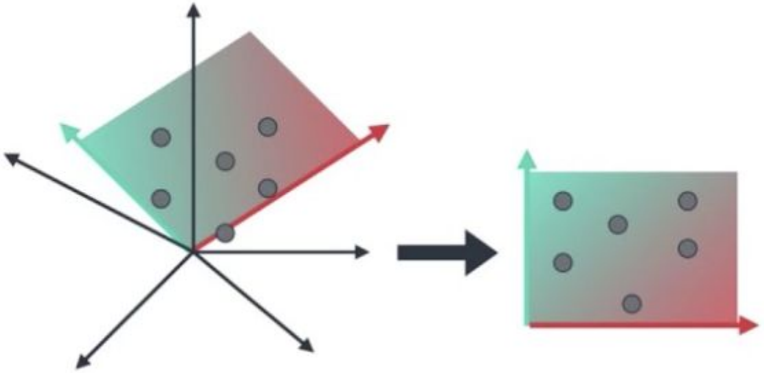


Colab

- https://colab.research.google.com/drive/1Awq3vkTUOvayvrVf588q5ZHOA0ypP-jt?usp=sharing
- ^ Shell code (make a copy for yourself)
- Download data from links and upload to Colab session

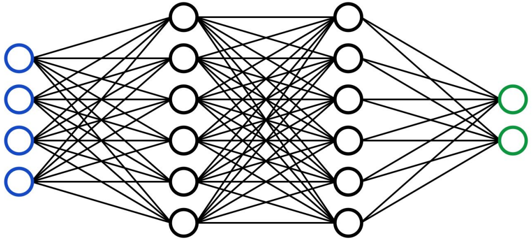
Principal Component Analysis (PCA)

- Dataset contains info from 20,000+ genes; how do we condense this?
- PCA is a form of dimensionality reduction (less features)
  - Still want to retain information, just more condense
  - Ex. 20,000+ → 500
- Finds the best way to project the original data onto a plane with less dimensions (minimizes information loss)



Neural Networks

- Collection of nodes (mathematical functions) organized in layers and connections between nodes of different layers
  - Nonlinearity function applied to weight matrix * input + bias
- Learns relationship between input and output data, useful for prediction
  - Input = expression data (condensed w/ PCA), output = predicted type of tumor (one of 5)
- Able to learn more effectively and efficiently than most other ML models



$$\rightarrow f\left(b+\sum_{i=1}^{n} x_i w_i\right)$$

Experiment Time!

- Try out different network architectures, aim for highest accuracy
- Hint - try adding:
  - A Dense layer, XXX number of layers, activation = “relu”

Thank You!

- We hope you enjoyed today’s workshop and learned a thing or two!
- Come to TJML and TJ Bioinformatics Society!!