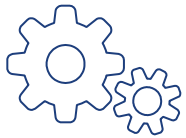
FAES BIOF509

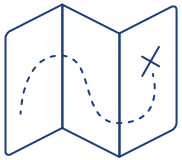
# Genetic Variants and Cancer



## Machine Learning Final Project

Mary B. Makarious 16.05.2019

# BACKGROUND



**Want to Follow Along?**

**Kaggle:** https://bit.ly/2LHDjSd

**Notebook:** https://bit.ly/2LMLPj9

## WHERE?

The data is from a Kaggle competition that was held ~2 years ago

I opted to use this specific dataset because:

- Interesting topic
- Easy to download
- Discussion on different methods used

## WHAT?

A lot of the ideas that went into this project came from:

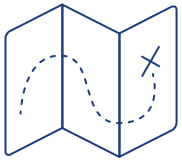
- Class materials
- Google
- Kaggle Discussions
- Medium articles
- GitHub Snippets/Tutorials
- What actually worked...

## DATA + GOAL

Data from Kaggle:

- Information about the genetic variants: `training_variants.csv` and `test_variants.csv`
- Clinical evidence (in text form) that was used to manually classify the variants: `training_text.csv` and `test_text.csv`

**Goal:** To classify each variant into 1 of 9 mutation classes (unknown to you)



**Want to Follow Along?**

**Kaggle:** https://bit.ly/2LHDjSd

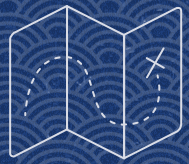
**Notebook:** https://bit.ly/2LMLPj9

# WORKFLOW

- Data Input
- Preprocessing
- Natural Language Processing
- 4-Layer Neural Net
- Developing the Model
- Training the Model
- Predictions on Test Data
- Reformatting for Output

- Loss per Iteration Plot



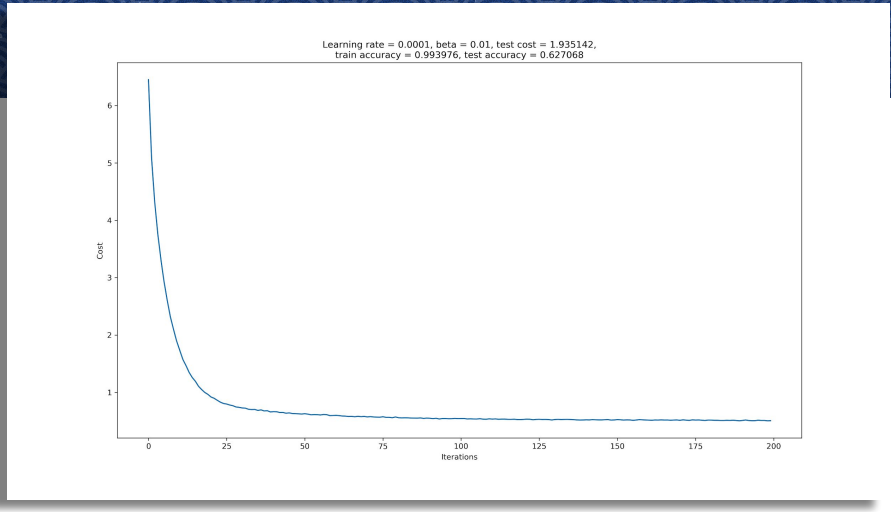


Want to Follow Along?

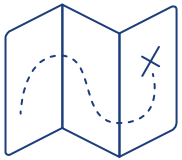
Kaggle: https://bit.ly/2LHDjSd

Notebook: https://bit.ly/2LMLPj9

# MODEL TRAINING + PLOT





**Want to Follow Along?**

**Kaggle:** https://bit.ly/2LHDjSd

**Notebook:** https://bit.ly/2LMLPj9

**Personal Interpretation**: Both the NLP and the 4L-NN were successful. The test accuracy is well above chance (~63%).

Would be interesting to see how to improve this (would adding another layer help?)

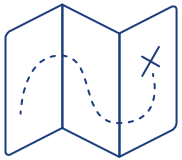
# OUTPUT EXAMPLE

| | class1 | class2 | class3 | class4 | class5 | class6 | class7 | class8 | class9 | id |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.000546 | 0.000325 | 0.015337 | 0.536189 | 0.019209 | 0.119735 | 0.306442 | 0.000516 | 0.001702 | 1 |
| 1 | 0.664214 | 0.183878 | 0.006075 | 0.017869 | 0.034264 | 0.047821 | 0.003931 | 0.022043 | 0.019904 | 2 |
| 2 | 0.366961 | 0.131212 | 0.024599 | 0.136363 | 0.072051 | 0.159974 | 0.036710 | 0.030545 | 0.041585 | 3 |
| 3 | 0.366961 | 0.131212 | 0.024599 | 0.136363 | 0.072051 | 0.159974 | 0.036710 | 0.030545 | 0.041585 | 4 |
| 4 | 0.094935 | 0.111473 | 0.039871 | 0.185180 | 0.093909 | 0.193979 | 0.201982 | 0.031785 | 0.046886 | 5 |



**Personal Interpretation**: Each variant is reported with the distances between each cluster following NLP+NN

(What way would be best to visualize this?)



**Want to Follow Along?**

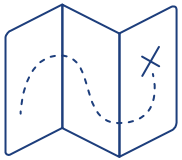
**Kaggle:** https://bit.ly/2LHDjSd

**Notebook:** https://bit.ly/2LMLPj9

# LESSONS LEARNED

- Machine Learning/Deep Learning can be tough
- There are multiple ways to do something
- This was my first attempt at both NLP and NN
- There are plenty of resources online explaining ways to further improve your model





Question for **you**: Ideas on how to best visualize this data?

**Want to Follow Along?**

**Kaggle:** https://bit.ly/2LHDjSd

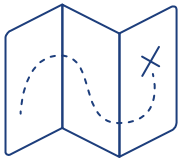
**Notebook:** https://bit.ly/2LMLPj9

## CREDITS

This project would not have been possible without:

- Martin and Alex
- Kaggle
- Google
- Countless hours of frustration
- Kind people who spend hours cultivating scripts, articles, and tutorials online
- Developers behind all the packages

Any Questions?



**Want to Follow Along?**

**Kaggle:** https://bit.ly/2LHDjSd

**Notebook:** https://bit.ly/2LMLPj9

**GitHub: m-makarious/ML_GeneticVariants_in_Cancer**