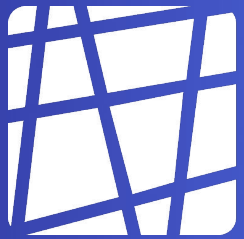
The Build Fellowship

BUILDFELLOWSHIP.COM



Open Avenues

Weekly Updates

- Please provide a quick update on either:
  - Something you did/saw this week that you thought was interesting
  - What you're looking forward to about this week's workshop

(Reminder - please have your cameras on if possible)

The Build Fellowship

OpenVenues

The Build Fellowship

Workshop 4 Featurization & Baseline Modeling

Recap

Sessions Overview

- Workshop 1 – Project Introduction & Setup
- Workshop 2 – Genomic Data (A2 Assignment)
- Workshop 3 – Data Analysis & Visualization (A3 Assignment)
- **Workshop 4 – Featurization & Baseline Modeling (A4 Assignment)**
- Workshop 5 – Model Training Approaches (Final Assignment Set)
- Workshop 6 – Model Tuning
- Workshop 7 – Performance Evaluation (Final Assignment Code/Testing Due)
- Workshop 8 – Results Presentation & Wrap up (Final Presentation Due)

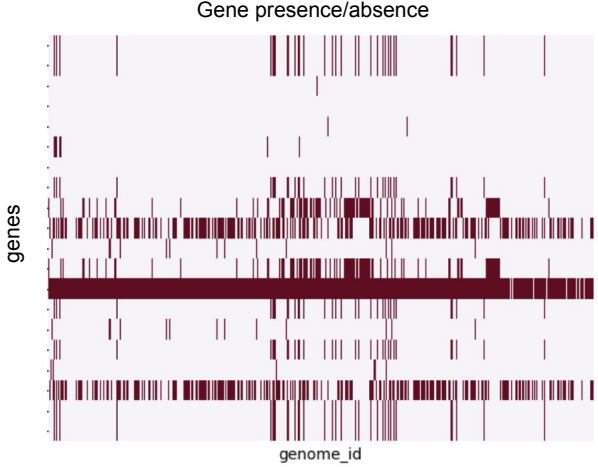
Exploratory Data Analysis

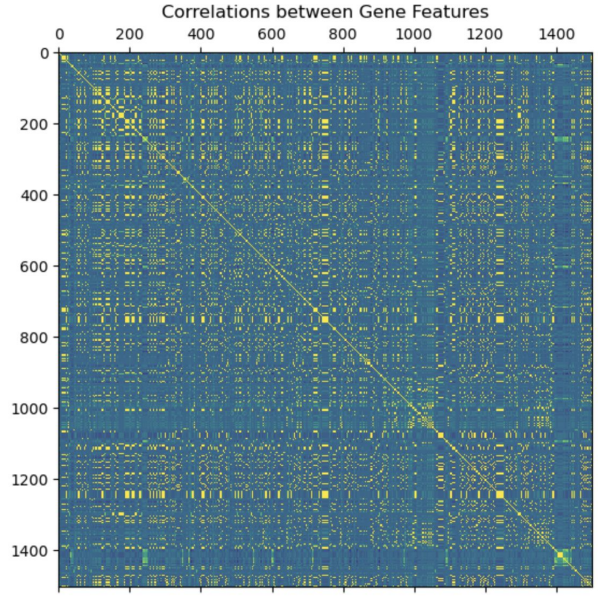
What did we achieve last week?

1. Took a look through our datasets, finding key information
2. Processed our targets (AMR) into clean binary S vs R
3. Identified a few suspicious E coli genomes
4. Reviewed gene presence/absence across samples
5. Started reviewing the nucleotide sequences and identified some mutations

Assignment (Optional):

6. Identified highly correlated genes

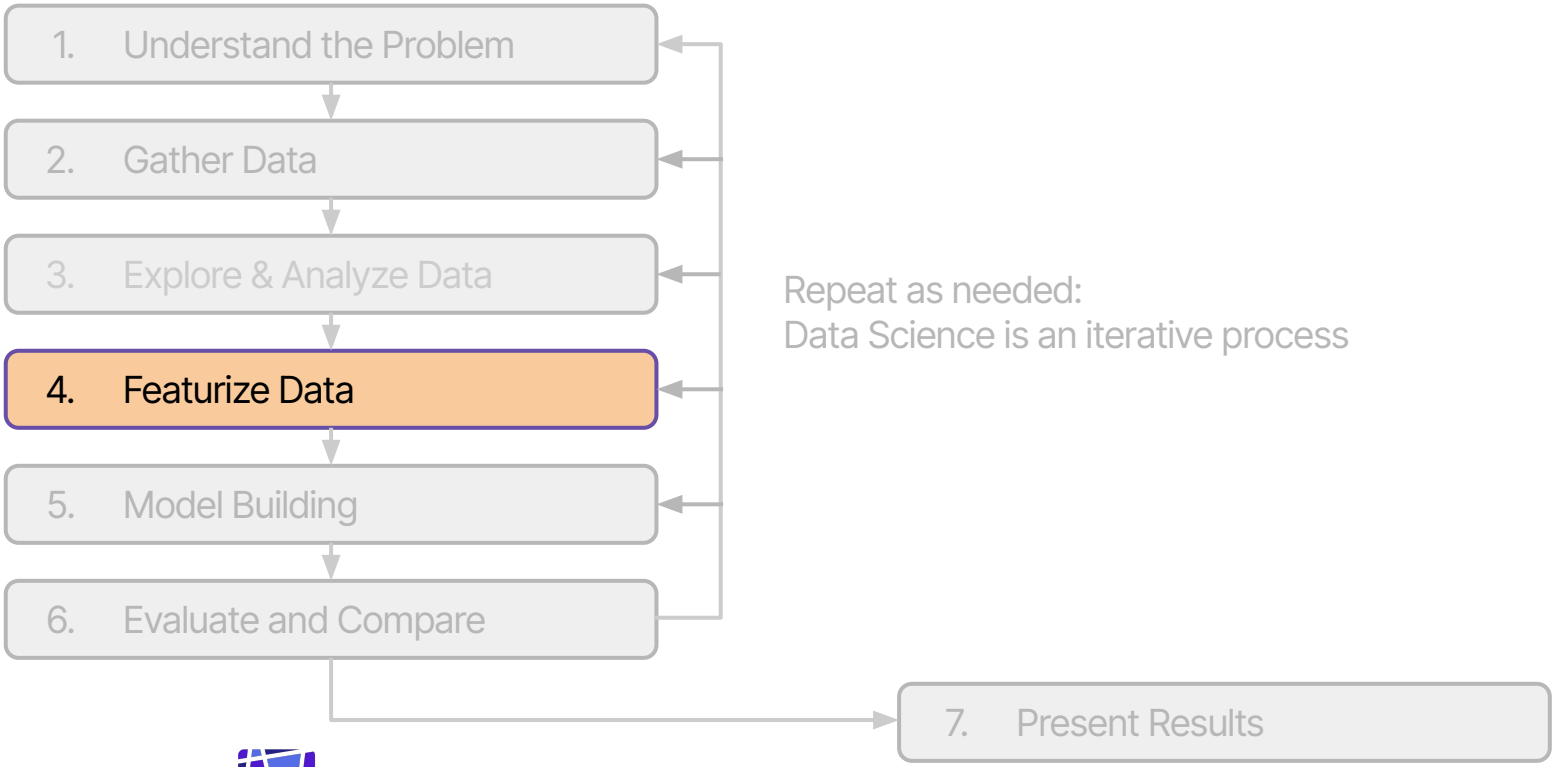




The Build Fellowship

OpenAvenues FOUNDATION

The Data Science Process



The Build Fellowship

OpenAvenues FOUNDATION

What is Featurization?

Gene Alignment to Sequence Input

CGTATAACTGA GGAACAGCGGT

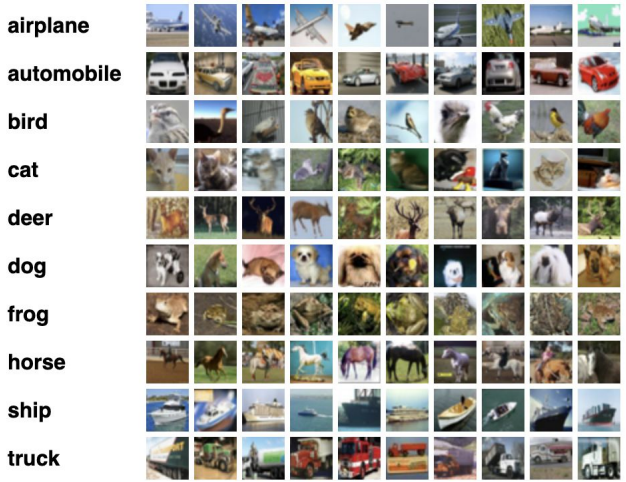
AACGTATAACTGATCGGAACAGCGGTA

- Tightly coupled with model building
- How do we go from raw data to predictions?
- Different ML models expect different data types
  - Tabular (Matrix)
  - Sequences
  - Images
  - Graphs
- Data formats & shapes need to be **consistent**
- We can't expect to pass raw data directly into our model
- Usually need to Encode our data into a numeric format

CGTATAACTGA + GGAACAGCGGT

= CGTATAACTGAGGAACAGCGGT

Cifar10 - Common ML Image Dataset



Model Options

The Build Fellowship

OpenAvenues FOUNDATION

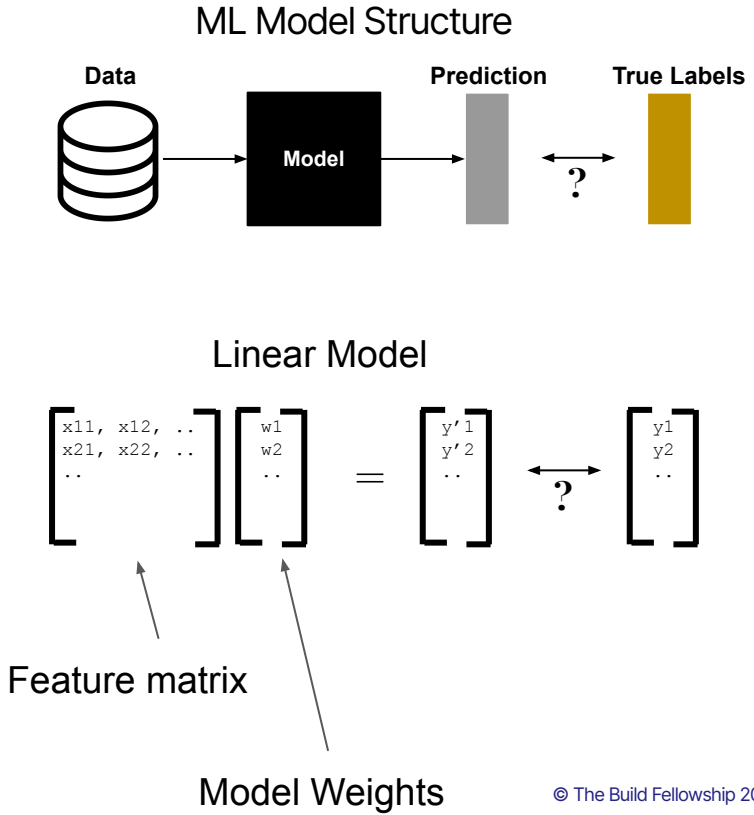
What model types can we use?

What sort of models can we leverage?

- Linear models (Tabular)
- Tree based models (Tabular)
- Neural networks (Tabular + Sequence)

Tabular
- Resistance genes
- Kmers

Sequence
- Gene sequences (mutations)



What model types can we use?

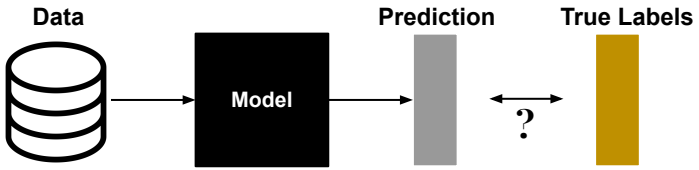
What sort of models can we leverage?

- Linear models (Tabular)
- Tree based models (Tabular)
- Neural networks (Tabular + Sequence)

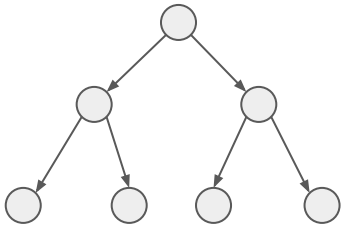
Tabular
- Resistance genes
- Kmers

Sequence
- Gene sequences (mutations)

ML Model Structure



Tree Based Models



What model types can we use?

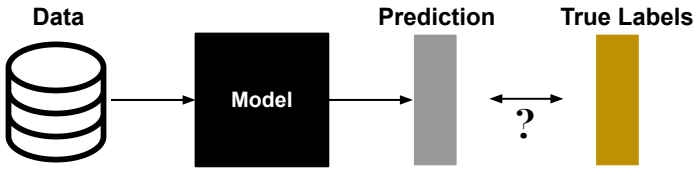
What sort of models can we leverage?

- Linear models (Tabular)
- Tree based models (Tabular)
- Neural networks (Tabular + Sequence)

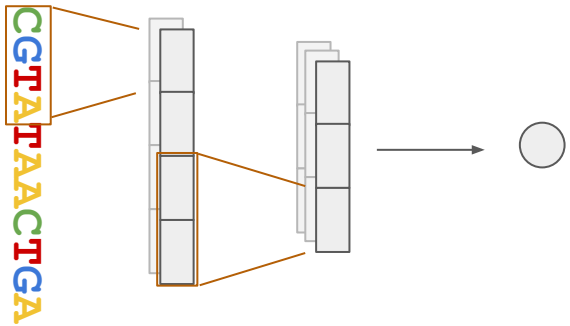
Tabular
- Resistance genes
- Kmers

Sequence
- Gene sequences (mutations)

ML Model Structure



Sequence Based Model



The Build Fellowship

OpenAvenues FOUNDATION

QUIZ TIME !?

How can we featurize DNA sequences?

a) Convert sequences to tabular counts
b) Don’t bother, just take it as is
c) Stitch all the sequences together into a single length
d) Consult experts to take only the important information

The Build Fellowship

OpenAvenues

Feature Options

The Build Fellowship

OpenAvenues FOUNDATION

Gene Presence Absence

- Simplest and most approachable feature
- Minimal processing required
- Saw an example of this last week

Tabular approach
- Standardize the data into a fixed shape
- Need to have a consistent size per sample

Binary
- Outcomes will be present vs absent for every possible gene

Input Data - Gene Alignment

genome_id	contig	res_gene	match_start	match_end	match_qual
562.11346	FLKS01000064	gb|U00096.3|-|3324062-3324911|ARO:3003386|Ecol...	96835	97684	849M
562.11346	FLKS01000044	gb|AP009048.1|+|3760295-3762710|ARO:3003303|Ec...	61096	63511	2415M
562.11346	FLKS01000070	gb|BA000007.3|+|4990267-4994296|ARO:3003288|Ec...	22038	26067	4029M
562.11346	FLKS01000068	gb|U00096.3|-|2336792-2339420|ARO:3003294|Ecol...	755765	758393	2628M
562.11346	FLKS01000064	gb|AP009048.1|-|3172159-3174052|ARO:3003316|Ec...	241841	243734	1893M

query_str	ref_gene_str
ATGAAACTCTTTGCCCAGGGTACTTCACTGGACCTTAGCCATCCTC...	ATGAAACTCTTTGCCCAGGGTACTTCACTGGACCTTAGCCATCCTC...
ATGTCGAATTCTTATGACTCCTCCAGTATCAAAGTCCTGAAAGGGC...	ATGTCGAATTCTTATGACTCCTCCAGTATCAAAGTCCTGAAAGGGC...
TTACTCGTCTTCCAGTTCGATGTTGATACCCAGCGAACGAATCTCT...	TTACTCGTCTTCCAGTTCGATGTTGATACCCAGCGAACGAATCTCT...
TTATTCTTCTTCTGGCTCGTCGTCAACGTCCACTTCCGGAGCGATT...	TTATTCTTCTTCTGGtTCGTCGTCAACaTCCACTTCCGGAGCGATT...
ATGACGCAAACTTATAACGCTGATGCCATTGAGGTACTCACCGGGC...	ATGACGCAAACTTATAACGCTGATGCCATTGAGGTACTCACCGGGC...

Gene Presence Absence

Format
- Reshape to a clean binary matrix
- Predicting AMR from genes directly
- Loss of a lot of information (sequences, counts)

Correlations
- Genes can be very similar
- Some genes can appear in all samples

Clustering
- Naive subsetting (throw away genes)
- Clustering can be used to group genes hierarchically

Presence/Absence Feature Matrix

	Gene A	Gene B	Gene C	Gene D	…
Sample A	1	1	1	0	
Sample B	0	1	0	1	
Sample C	1	0	1	0	
Sample D	0	1	0	0	
Sample E	1	0	1	1	
	↑	…	↑		

Identical Presence Absence

The Build Fellowship

OpenAvenues FOUNDATION

Gene Presence Absence

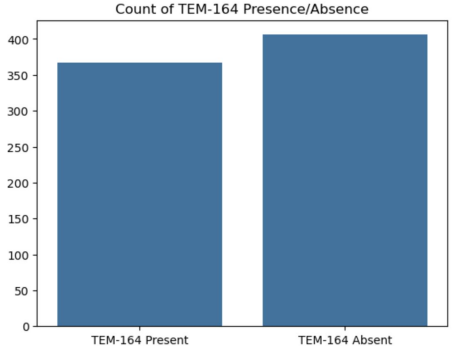
Why might Gene presence absence not always work?

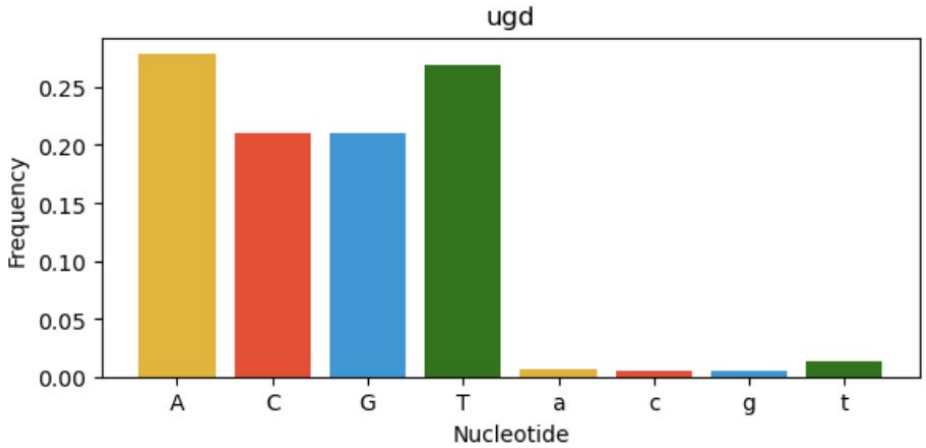
- Weak correlations
- Genes can be always present (mutations matter)
- Losing information

Query vs Ref Gene String

- Raw data has both query & reference
- Query = CARD sequencing
- Reference = Sample genomes
- Lower case = difference in nucleotides

Can we capture these sequence differences?





The Build Fellowship

OpenAvenues FOUNDATION

Kmers

Kmers is one of the most common approaches for featurizing genomic data

Core concepts:

1. K = Fixed length representation (parameter)
2. Wish to capture the nucleotide information
3. Convert from arbitrary sequences to a fixed shape representation

High K = huge feature space, high chance of uniqueness

Low K = small feature space, high chance of repetition within genomes

K is a parameter

2-mers: AA AC AT AG CA CC CT CG
TA TC TT TG GA GC GT GG

16 unique combinations

5-mers: AAAAA AAAAC AAAAT AAAAG
AAACA AAACC AAACT AAACG

1,024 unique combinations

10-mers AAAAAAAAAA AAAAAAAAAG
AAAAAAAAAC AAAAAAAAAT

1,048,576 unique combinations

The Build Fellowship

OpenAvenues FOUNDATION

Kmers

How to generate Kmers?

- Want unique occurrences of each kmer
- High K's = millions/billions of sequences
- Sliding window approach
- Scan the whole sequence one step at a time
- Keep track of counts

Using 5-mers:

Slide a window of size K across the genome

Kmers

How to generate Kmers?

- Want unique occurrences of each kmer
- High K's = millions/billions of sequences
- Sliding window approach
- Scan the whole sequence one step at a time
- Keep track of counts

Using 5-mers:

CGTAT = 1
GTATA = 1

Count each occurrence

Kmers

How to generate Kmers?

- Want unique occurrences of each kmer
- High K's = millions/billions of sequences
- Sliding window approach
- Scan the whole sequence one step at a time
- Keep track of counts

Using 5-mers:

CGTAT = 1
GTATA = 1
TATAA = 1

Count each occurrence

Kmers

How to generate Kmers?

- Want unique occurrences of each kmer
- High K's = millions/billions of sequences
- Sliding window approach
- Scan the whole sequence one step at a time
- Keep track of counts

Using 5-mers:

CGTAT = 1
GTATA = 1
TATAA = 1
ATAAC = 1

Count each occurrence

The Build Fellowship

OpenAvenues FOUNDATION

Kmers

Kmer Count Matrix

	CGTAT	GTATA	TATAA	ATAAC
Sample A	10	1	34	0
Sample B	0	3	30	0
Sample C	3	1	54	0
Sample D	8	2	21	0
Sample E	0	1	24	1

…

…

Using 5-mers:

CGTAT = 1
GTATA = 1
TATAA = 1
ATAAC = 1
…

Count each occurrence

The Build Fellowship

OpenAvenues FOUNDATION

Sequences

- Sequences are the most raw format of data
- Most difficult to work with
- Convolutional/Recurrent Neural Networks
- Few proven working examples in literature

Challenges:

- Data size (full sequences are millions of characters)
- Complexity vs Training data

Working with Sequence models is an **optional extension**

- Computationally complex
- Possible option for final project
- See paper for reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8722762/

Sample A

```
GeneA: GTATCTGAGGAAC
GeneB: GAGGAGATTCGCT
GeneC: TGAAACGTATGCC
```

Sample B

```
GeneA: GTATCTGAGGAAC
GeneC: TGAAACGTATGCC
GeneD: GAGGAGATTCGCT
```

The Build Fellowship

OpenAvenues FOUNDATION

Sequences

Simple sequence featurization scheme:

- Randomly concatenate genes
- Find all genes and stack together
- Require a consistent length
- Truncate or Pad ends
- One hot encode to numeric

Sample A

```
GeneA: GTATC
GeneB: GAGGA
GeneC: TGAA
```

↓

```
GTATC-GAGGA-TGAA
```

Sample B

```
GeneA: GTATC
GeneC: TGAAA
GeneD: TATGCA
```

↓

```
GTATC-TGAAA-TATGCA
```

The Build Fellowship

OpenAvenues FOUNDATION

Sequences

A=1 C=2 G=3 T=4 Pad=0

Simple sequence featurization scheme:

- Randomly concatenate genes
- Find all genes and stack together
- Require a consistent length
- Truncate or Pad ends
- One hot encode to numeric

Sample A

GTATCGAGGATGAA

341423133143100

Sample B

GTATCTGAAATATGCA

341424311141432 1

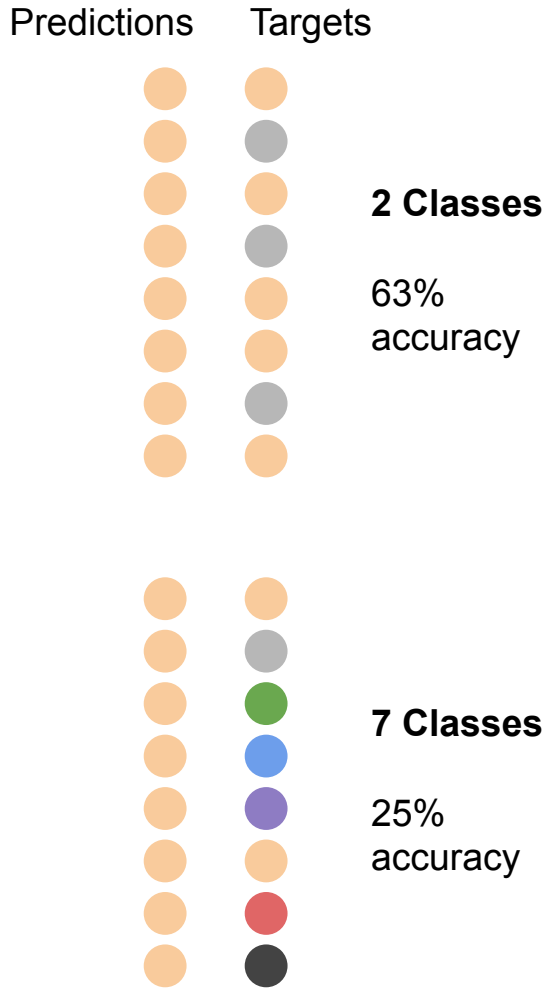
Baseline Models

The Build Fellowship

OpenAvenues FOUNDATION

What is a Baseline Model?

- In any ML task: set ourselves a target to beat
- Understand how well we're performing
- Performance statistics out of context can be misleading
- Is 80% accuracy good? What about 20% accuracy?
- If we have two targets? If we have 50 targets?
- Baselines should be **interpretable**



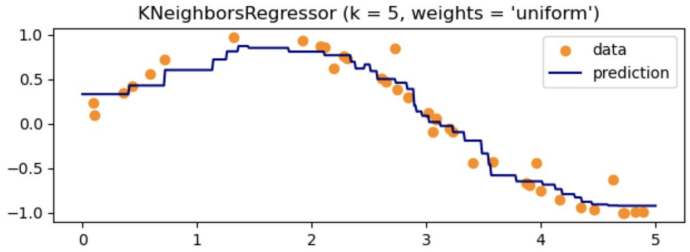
The Build Fellowship

OpenAvenues FOUNDATION

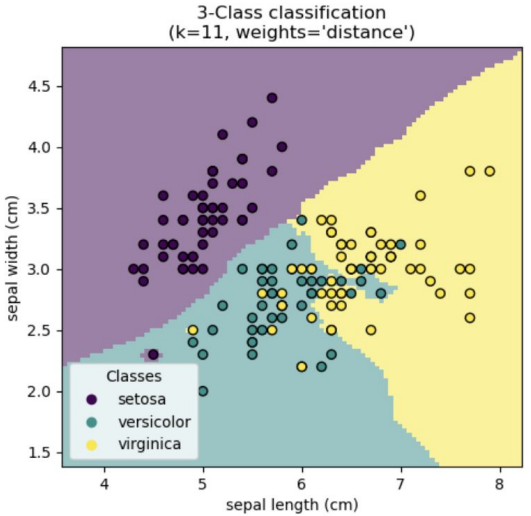
Common Baselines

- Completely Random
  - Nonsense worst case model
  - Useful to flag if data processing has failed
  - E.g. accidentally mixed up targets
- Predict the majority class
  - True “baseline”
  - Simplest way to get maximum performance
- KNN (K-nearest neighbor)
  - Doesn’t require training
  - Directly learn based on neighbors
  - Simple and interpretable

Examples: KNN Regression & Classification



[1]



[2]

The Build Fellowship

OpenAvenues FOUNDATION

Workshop 4 Featurization & Baseline Modeling

The Build Fellowship

OpenAvenues FOUNDATION

References

[1]:https://scikit-learn.org/stable/auto_examples/neighbors/plot_regression.html#sphx-glr-auto-examples-neighbors-plot-regression-py

[2]:https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py