2/23/22 Classification, kNN

Classification

- training set: train a model to learn a rule
- use training set to Label

ex: Model: f: age x tumor size → {yes, no}

apply model to our dataset

Tasks: credit card fraud detection

Techniques: Naive Bayes, Neural Network

kNN

1) compute distance of unseen record
2) identify k nearest neighbour
3) aggregate

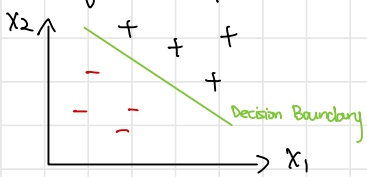
Aggregate method:

- majority rule
- weighted majority based on distance ($w = 1/d^2$)

Choosing k:

if small → sensitive to noise + overfitting (need to be general, not too specific)

if big → include points from other class



Pros: based on similarity, simple, Black Box model, adapts to new attributes

Cons: expensive

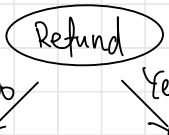
Decision Trees

walk through the tree and make prediction

How to build the tree: Hunt's Algorithm

- split up the dataset to get single class
- Goal: Find the best attribution that majority of one class in each class

example:



split on Refund

class is Yes, No.

C_1	10
C_2	10

C_1	10
C_2	0

We do the splitting recursively

Need to define Base Case first.