# WEEK 1

8/22/22

* Kinsman’s rules of his class
  + Keep yourself alive and (relatively) healthy
  + August center is good, use it if need be
* *Take picture with a licensed counselor on campus*
* *Update MyCourses with a picture of yourself*
* Believe it or not, taking notes and paying attention is good.
* Cheat sheets are good and are a graded assignment.
* Mathematical duality is an opposite way of thinking about it;   
  i.e: left handed vs non-right handed  
  i.e. maximizing the amount of gas before
* Convex optimization can be used to either maximize or minimize a problem
* “Eggs-Here’s some eggs” - Prof. Kinsman
* “Eggs are really important in computer science” - Prof. Kinsman
* Tri State logic - like binary, but something can be neither on nor off; somewhere in the middle.
* Crisp and fuzzy mixture models are exactly what it sounds like
  + Crisp is easier to separate, think a black and white glazed cookie.
  + Fuzzy is “a little of this and a little of that everywhere, like sprinkles”

8/24/22

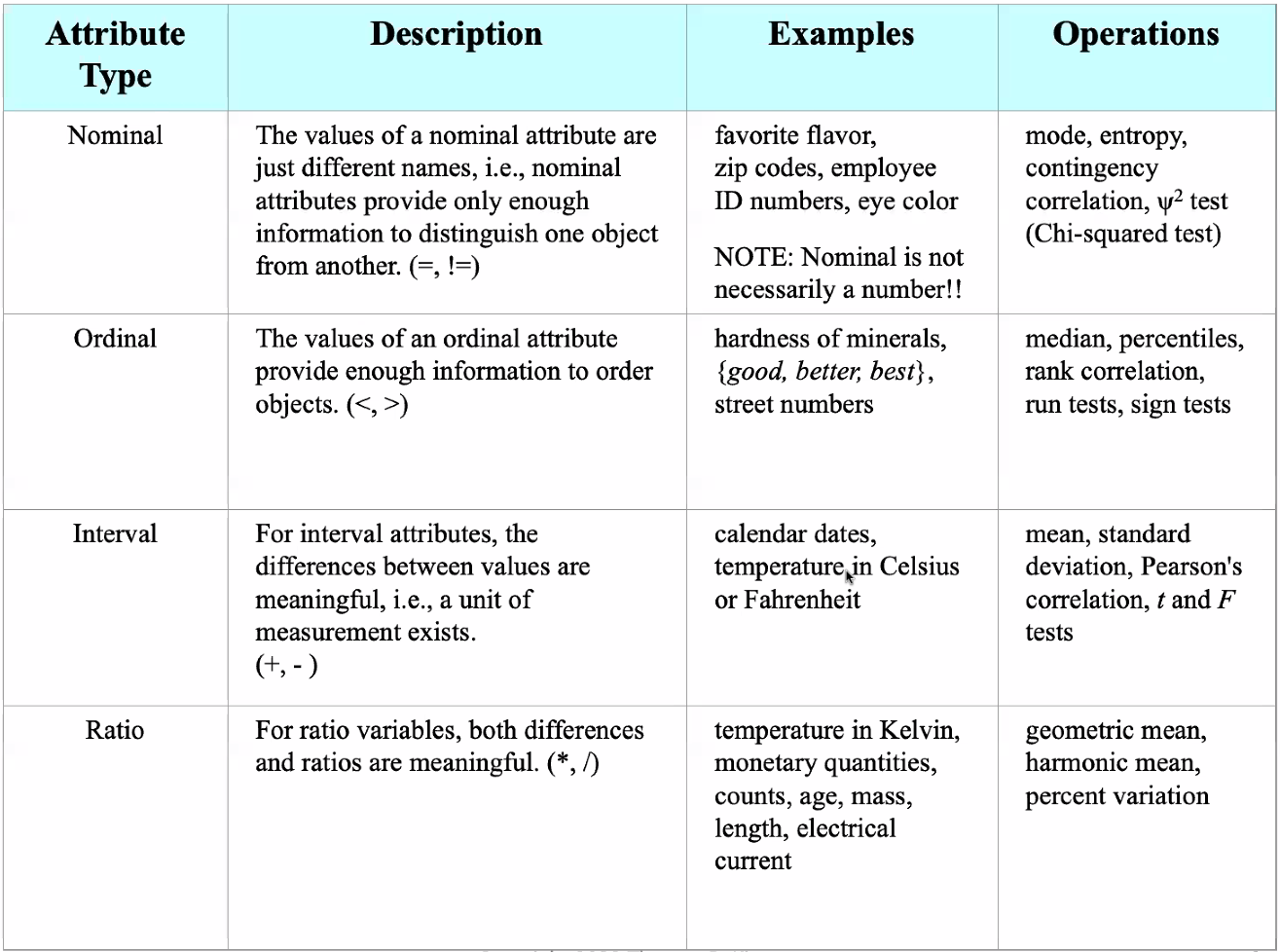
* Talking about data mining vocabulary
* “Which of these is the best pen?” *points to a picture of a lot of pens and pencils*The best one was clear view sharpies
* “Try them all and test”
* computing log\_2 -> log\_2(x) = log(x) / log(2)
* Expect to have problems with data; incorrect, missing, and redundant.
* Association analysis and driving to work; “If you are going into work between 6:30 and 8:30, you can expect cars to drive quickly”
* Self organizing happens automatically but it is part of how the data works; Pine trees give off an acid that kills off other trees, so pine trees will grow nicely while killing the trees around them.
* The data gap is the gap between the amount of data scientists vs the total number of TB of data; the solution to this is machine learning so we can automate data processing.
* Remote sensors of a satellite and scientific simulations of gravity happen at RIT
* Data lag is the time between when an event happens and when that data becomes usable and actionable.
* Clustering: starting with unlabeled data, “finding the lumps in the data”
  + Finding the structure inherent in the data
  + Describing the data simply
* Classification : starting with labeled data, “ i hear something walking in the house at night, it sounds like a person, versus scurrying and thinking it is a chipmunk”
  + Use measures to predict values of hidden states
  + Try to predict a target value
* Prediction/Regression
  + Estimation (Parzen estimation/density estimation)
  + Regression (linear)
  + Fitting a function for the data
  + Sometimes used for classification
* “Try them all” programming pattern
* Prof. Kinsman has a lot of pottery tools
* In the toolbox of pottery tools, there was a lot of dust/dirt in the box, this is considered noise. There was also a chop-stick in there, which is an anomaly.
* Give each cluster a name, this can be called a pseudo-name or nickname
* Data sent back is called telemetry data
* Outlier vs anomaly
  + outlier - something that is in the distribution, but is on the edges of that distribution, it is actually meant to be there
    - considering people’s heights: a very short person and a very tall person
  + anomaly - something that isn't meant to be in the distribution
    - considering people’s heights: a person with negative height
* To handle outliers, you can reserve judgment, take them out of consideration, give extra attention, find a different data set, and/or some combination of these.
* Latent variable is a variable that is hard to observe

# WEEK 2

8/29/22

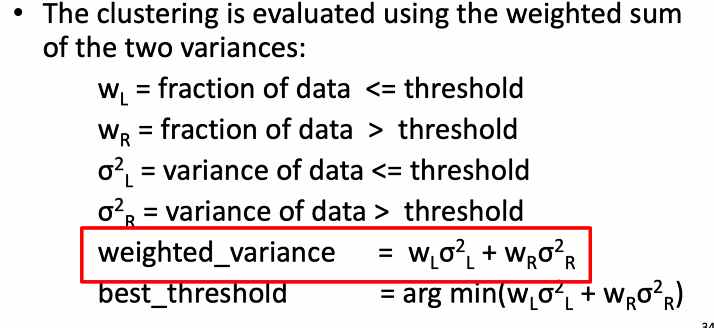
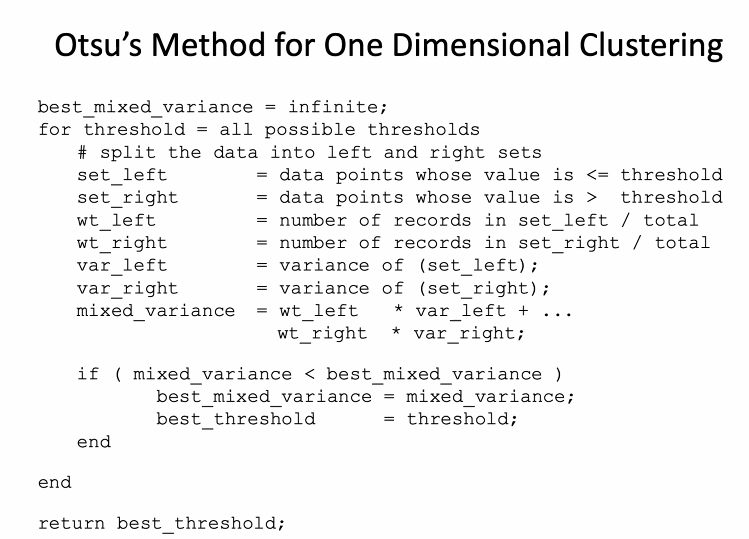
* Decision trees are glorified if/else statements
* Any classifier can suffer from overfitting
* Overfitting is when a classifier strictly works for one dataset, but not for any in the future.
* Law of unintended consequences; any time a law is created, there are always unintended consequences.
* 5 basic business transaction phases
  + pricing
  + request
  + order
  + sale
  + return
* Master Data Management; redundancy breeds inconsistency
* AAA for data
  + Authentication - validating that you are who you say you are
  + Authorization - validating that you should have access to this data
  + Accounting - recording the transactions that occurred
* *Read chapter two, NOIR (French for dark); Nominal Ordinal Interval Ratio*

8/31/22

* Kinsman’s litmus test of ethics is “Can someone take you to court with it?”
* Data governance
  + How data is stored
  + How long is it stored for
  + When it expires
  + Information needs to expire to protect the company
* toxic data - data which if released
* Discrete attribute date
  + has only a finite set of values
    - like zip codes, counmts, or set of words
  + often represented as integer variables
  + binary attributes are a special case
* Continuous attribute
  + has real numbers as attribute values
    - like temperature, weight, or height
  + practically, real values can only be measured and respresnted using a finite number of digits
  + typically represented as floating point variables
* Taxonomy of types of attributes
  + Categorial
    - Nominal, like a name, can test if they are equal or not equal
    - Ordinal, like a shirt size, has a comparable order
  + Numeric
    - Interval, like shoe sizes, like a temperate(F or C)
    - Ratio, must have an absolute zero, like Kelvin
* 
* The 3 types of data
  + Record data
    - Records the current status
  + Graph based data
    - Records relationships
    - Contains graphics and chemistry
    - Social Network Analysis
  + Ordered data
    - Event data that contains a time component
    - Happens in some sort of sequence
    - Not all ordered data has a time component, but all event data does
    - Once you click a link, you cannot unclick it

# WEEK 3

9/7/22

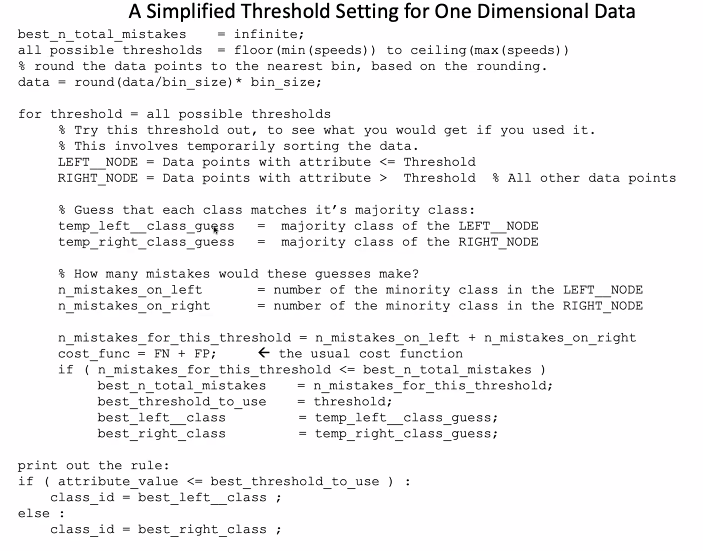
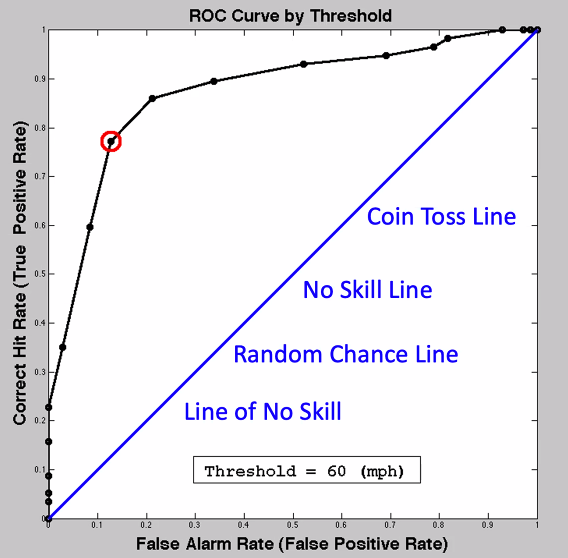
* Outliers throw off the computations
* Inconsistent values
  + Zip codes as house numbers
  + Product code vs order number
* Signs of racing
  + Taking little time on surveys
  + Picking only the first or last option
* De-duplication was the first money maker for data mining
  + it reduced redundant exact duplicate entries
* Dis-ambiguation
  + When the unique key isn’t really unique
  + To fix this you can either
    - Merge the info in the records
    - Delete the oldest/least reliable record
* A jutter is a random noise that was added to the data set to differentiate data
* Clustering vs Classification
  + Clustering throws everything into groups
  + Classifiers have a trained set of data
* Occam’s razor - usually choose the easier explanation.
* Otsu’s method - splits data into two clusters with the min average variance between them
* Inter cluster - space between different clusters in minimized
* Intra cluster - space within a cluster is minimized
* Training set - using a set of labeled data where you know the underlying group identification to try to assign future instances to a group
* Testing set - error rates are calculated via this set, not used in designing the model
* Validation set - only used by the customer to judge the classifier
* Cross set - use a training set, cut it up, then use part of it as a validation set
* 
* 

# WEEK 4

9/12/22

* The cost function is the objective + regularization
* The objection function is something you absolutely want to be 0 (not crashing while driving)
* The regularization function is stuff that would be nice if it was 0 (time it takes to go from A to B)
* Kinsmans regularization recipe
  + Decompose the problem into
    - terms you really need to be 0, and normalize with a normalization function so that it is ~1
    - terms you want to be 0, and normalize with a normalization function so that it is ~1
* data visualization - common technique is to plot every variable vs every other variable

9/14/22

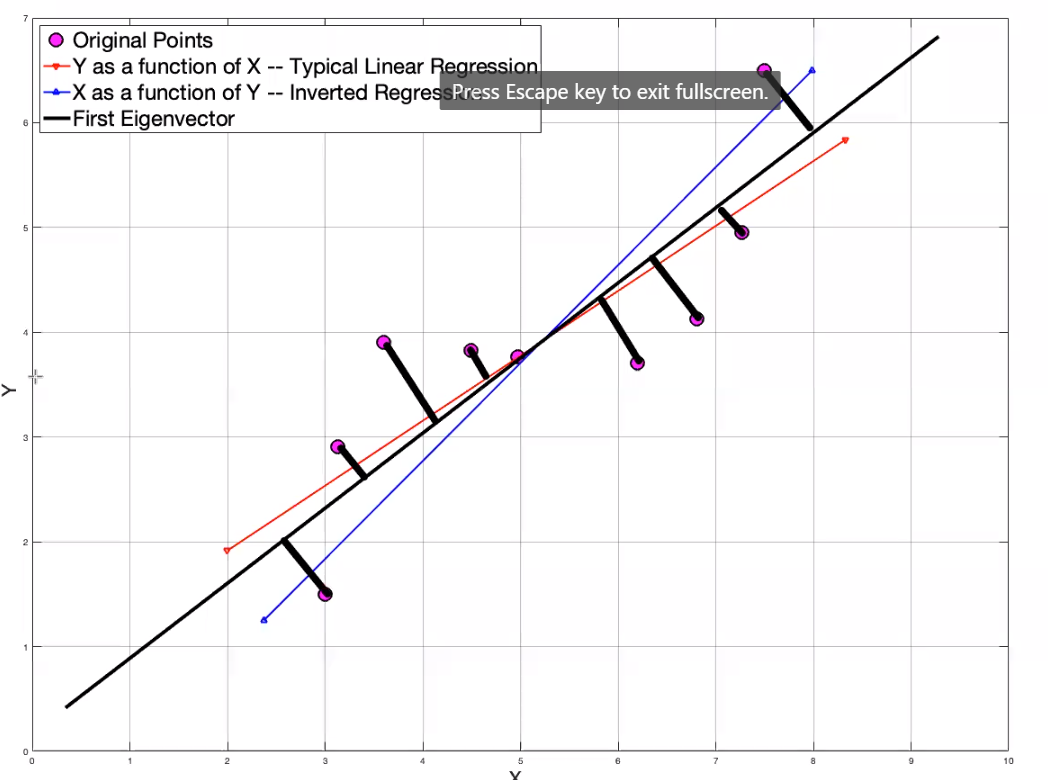
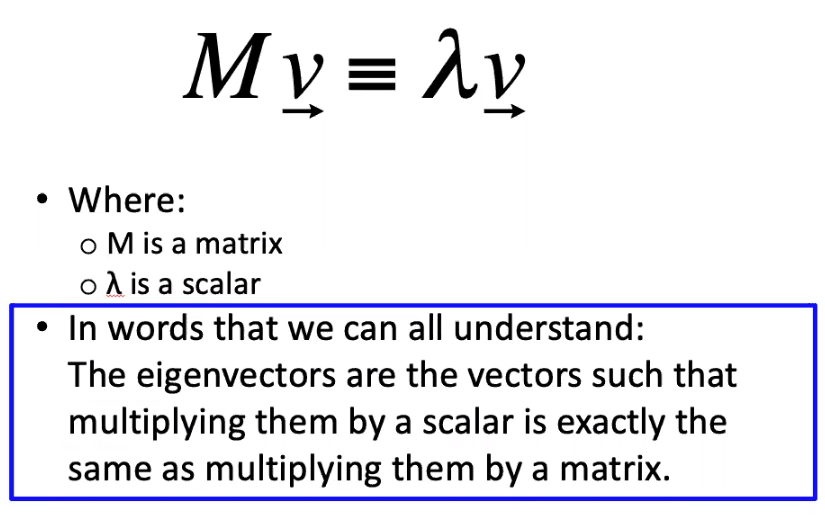
* False positive and false negatives are bad
* However having no false positives is ideal, if you have humans involved you need to try to not bug them with excess false alarms.
* Having no false negatives is more ideal if you absolutely NEED to catch things that can go wrong, like in health care.
  + In medical screenings, it is better to have a false positive because you will treat yourself like you have the disease versus having a false negative and walking around like you are perfectly healthy when you are indeed sick
* 
* ROC Curves - Receiver Operator Characteristic Curves
* 

# WEEK 5

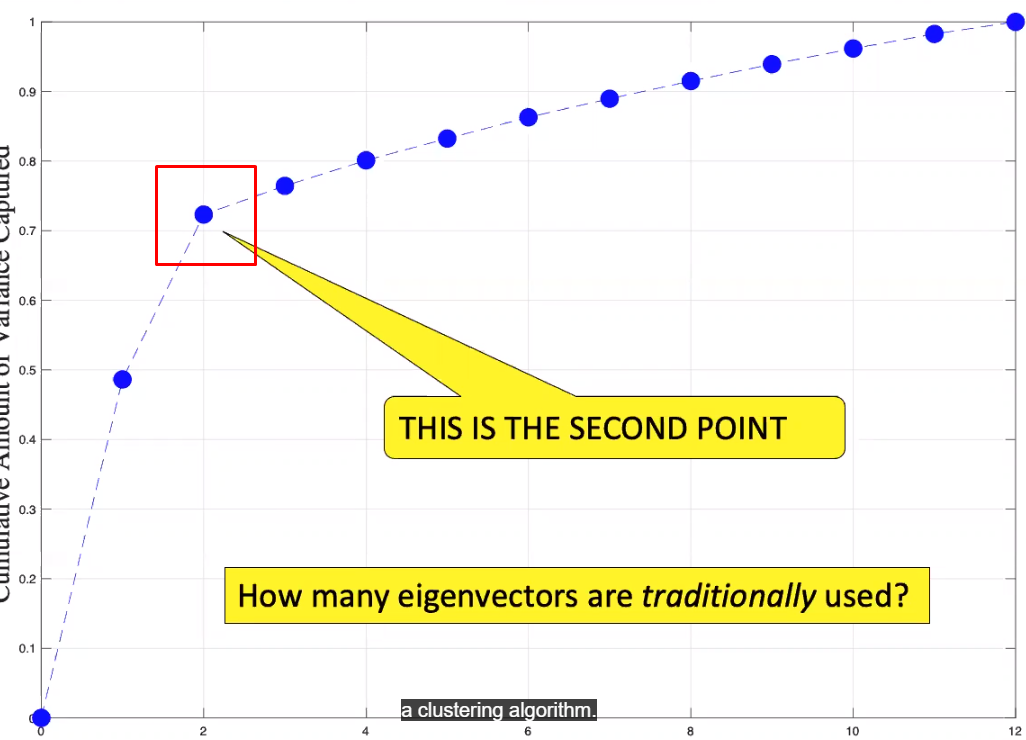
9/19/22

* Talking about the obtuse quiz, there were some interesting submissions
* conditional probability - a chance of something happening due to something else happening
* semantic gap - jump between what is said and what was really meant (using sarcasm)
* for decision trees, the answer to one question is another question
* google is an ensemble classifier

November 2, 2022

* PCA first eigenvector (aka the first principal component) tells you the maximum direction of variation
* For PCA you just get the distance from the line to each point
* The second eigenvector is always perpendicular to the first eigenvector
* kmeans is for clustering and k-nn is for classification
* GPUs are made to do vector math
* Projecting N, M dimensional data points onto M eigenvectors, you have M-dimensional data still
* 
* When trying to find the eigenvectors of the covariance matrix of the data, M is the covariant matrix and lambda is a scalar variable
* Each eigenvalue has an associated eigenvector, and every set of eigenvectors form an orthogonal set of basic sectors
* The eigenvalues relate to how much variation the associated eigenvector contains
* Each of the eigenvalues represent how much of the variation is explained by the associated eigenvector, you do this until 95% of the variance has been explained.
* Eigenvalues can help you guess the shape of the data
  + one big EV, and two nearly the same “ means you have a hotdog”
  + two big EVs, and one small one “means you have a pancake”
  + First three are all about the same size “means you have a meatball”

November 7, 2022

* A priori rule- if something is an important set in the whole data set, then it must be important in one of the subsets
* TED talk about homophily, we are bound to be around people who appear to be similar to us.
* the only two times a shopper changes habits is when they get married or a child
* we project for dimensional reduction and feature reduction
* inflection points- saw one in the knee point in kmeans, second one is
* The essence of PCA is to remove irrelevant attributes
* The one rule gives you the most important attribute and you need to plan on how to break ties (if there is one, refer to the zero rule)
* the zero rule is based on looking at zero attributes
* target variable is always on top

November 9, 2022

* what is the difference between PCA and FLD(Fisher Linear Discriminant),
  + PCA is a pre-processing technique that helps us rotate data.
  + FLD is always used for us to make a classifier
* LDA(Linear discriminant analysis) is for classification.