Guys, try to give an explanation and maybe the source of your answer

For some of them, put the questions because there are **different versions**

**Guys, label your Part B. Answers, not the same for everyone**

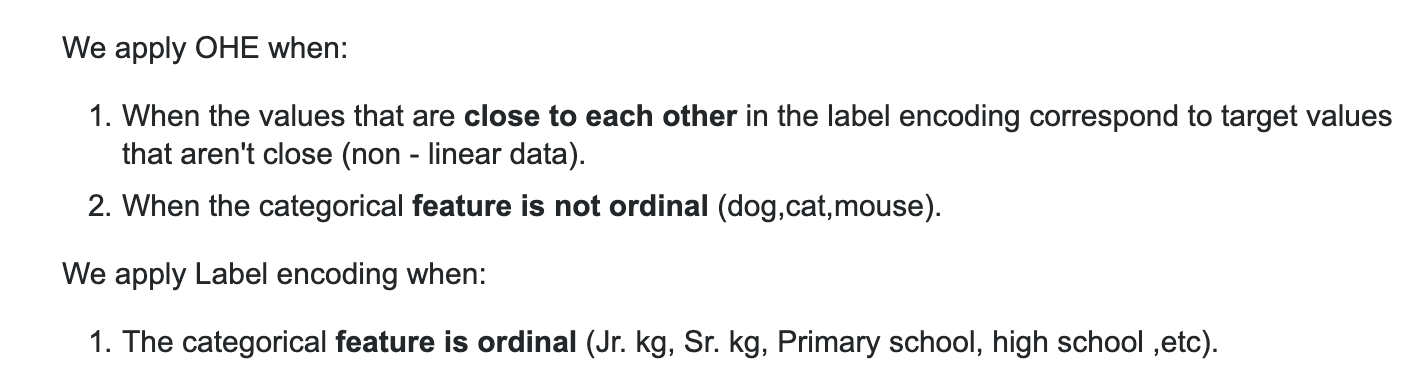
For 3 and 4:

**REVIEW**

Normalizing your data is an essential part of machine learning. You might have an amazing dataset with many great features, but if you forget to normalize, one of those features might completely dominate the others. It’s like you’re throwing away almost all of your information! Normalizing solves this problem. In this article, you learned the following techniques to normalize:

* **Min-max normalization**: **Guarantees** all features will have the exact same scale but does not handle outliers well.
* **Z-score normalization**: Handles outliers, but does not produce normalized data with the *exact* same scale.

A. True or False

Label encoding 1.**T** It’s False

Main Use of Pca 2.**T**

Z-Score is useful 3.**T**

Min-max normalization 4.**F** (yes, since we have “does not guarantee”)(F is right)

Discretization 5. **F**

Common aggregation purpose 6. **False**, Aggregation does not bring in more data for the user (Slide 70)

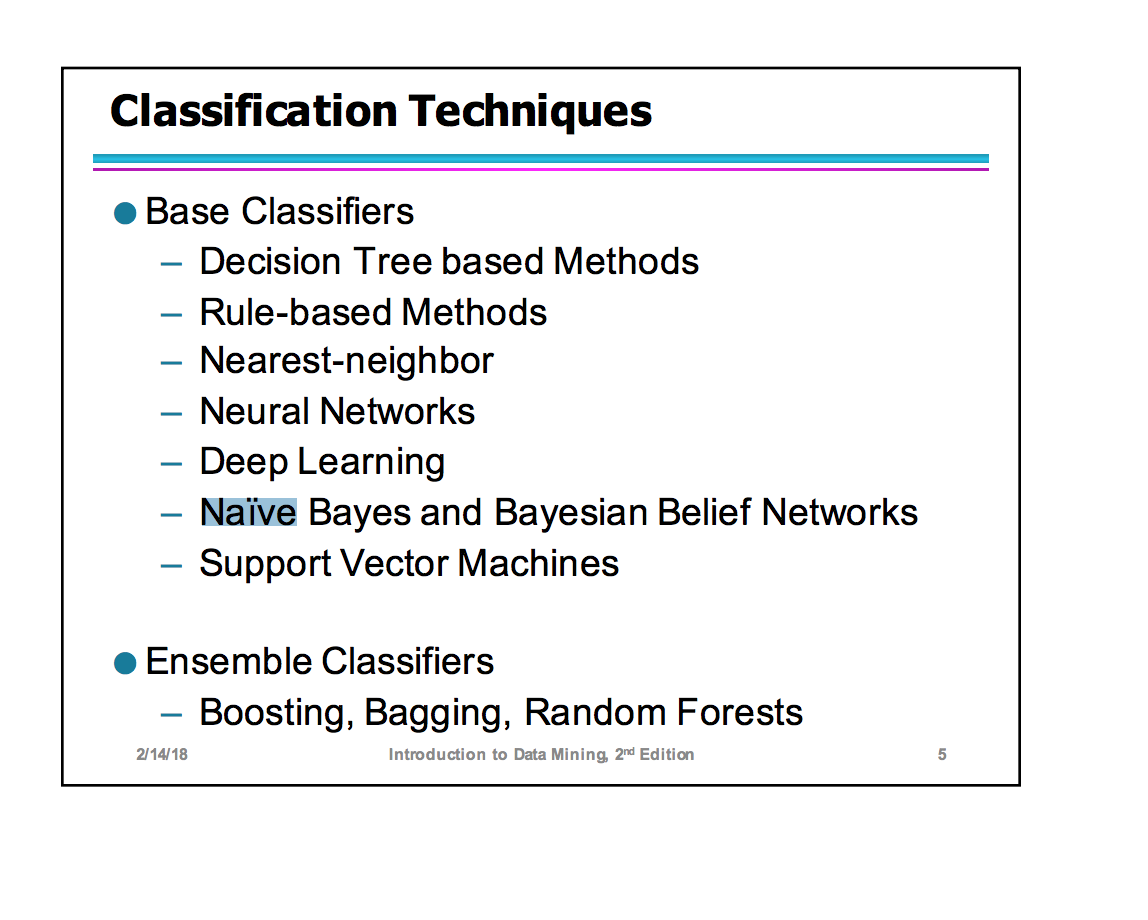
Normalization 7. **T** (yes True, <https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029> )

Support vectors 8. **T**

K-Means uses 9. **F**

Scatter Plot 10. **T**

**B. Matching**

1. PCA -> **J,** PCA is choosing the correct features aka feature selection
2. Z- Score - > IS THIS B? Yes its B
3. KNN -> **Nvm, I think it’s (I),** it uses a distance measure.
4. Barplot ->**E,** [**https://www.datamentor.io/r-programming/bar-plot/**](https://www.datamentor.io/r-programming/bar-plot/)
5. Naive Bayes -> **H** Naive bayes is classification technique)
6. Logistic -> **G**(pg 284 on the PDF) and also Quiz 3, question B1
7. Single link hierarchical -> C
8. Missing values -> **D** We replace missing values with mean and median, not sure about mode. I think we can do mode as well . I agree with D
9. Backward elimination -> **F,** feature elimination In backward feature elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.,
10. One hot encoding -> **E**, (not sure though; it is in tutorial #4; **encode\_text\_dummy** - Encode text fields as numeric, such as the iris species as a single field for each class. Three classes would become "0,0,1" "0,1,0" and "1,0,0". Encode non-target features this way. used when the data is part of input ***(one hot encoding)***) It’s definetely not A.. one hot encoding is changing a categorical attribute to a number

**C. Multiple choice q’s**

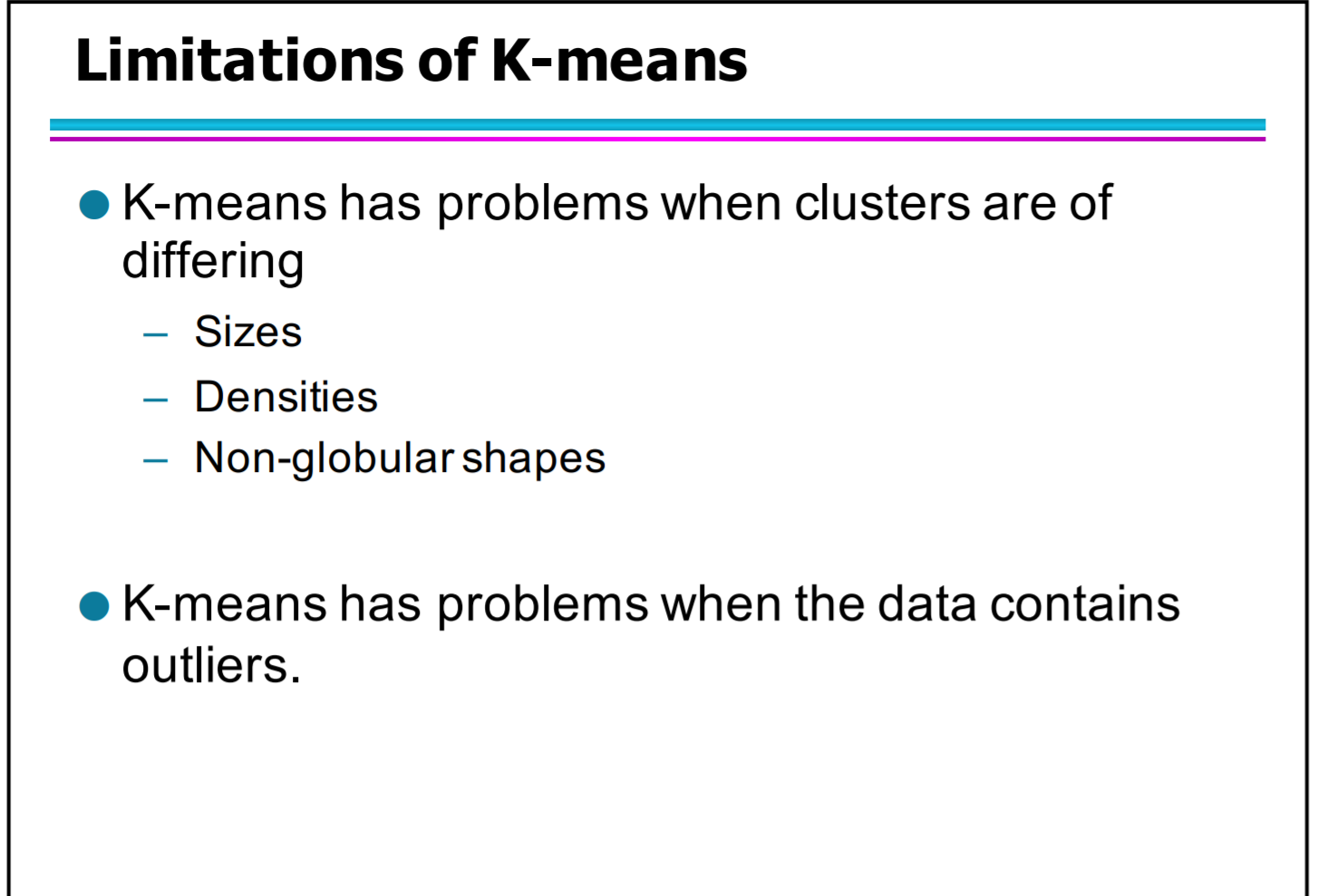
1: C, because normalizing will make the columns comparable in this problem. Changing them to a similar unit of measurement will be wasting time and code.

2: A, im guessing A too. Slide says to estimate…. I agree with A, because if this is a classification problem, we do not want to screw around with the mean, therefore, putting it as the mean would be the best option

3: **D?** boxplots and z-score can be used for outlier detection, so it has to be D

4: A, for visualization. Helps compare different attributes

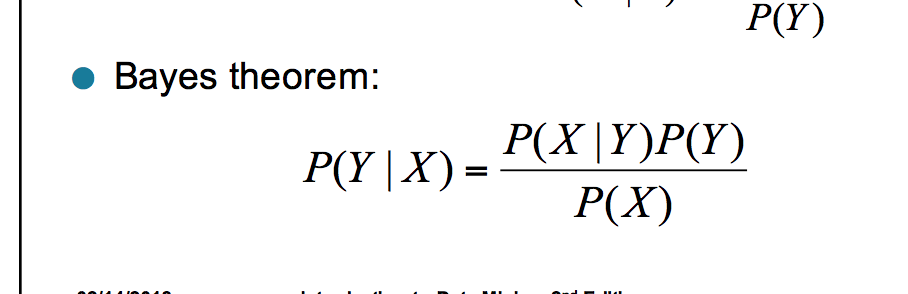
5: **A? K-Means or B? SVM**, because this relies on the mean of the data set.. I’m second guessing this, because the question calls for the MOST sensitive to outliers. SVM is also sensitive], **See below, I think it’s K-means**



6: D

7: **B** - Unsupervised learning…. Google search?

8: **C** - Conditional Probability… pic below (Shoudlnt it be C?) yeah my bad

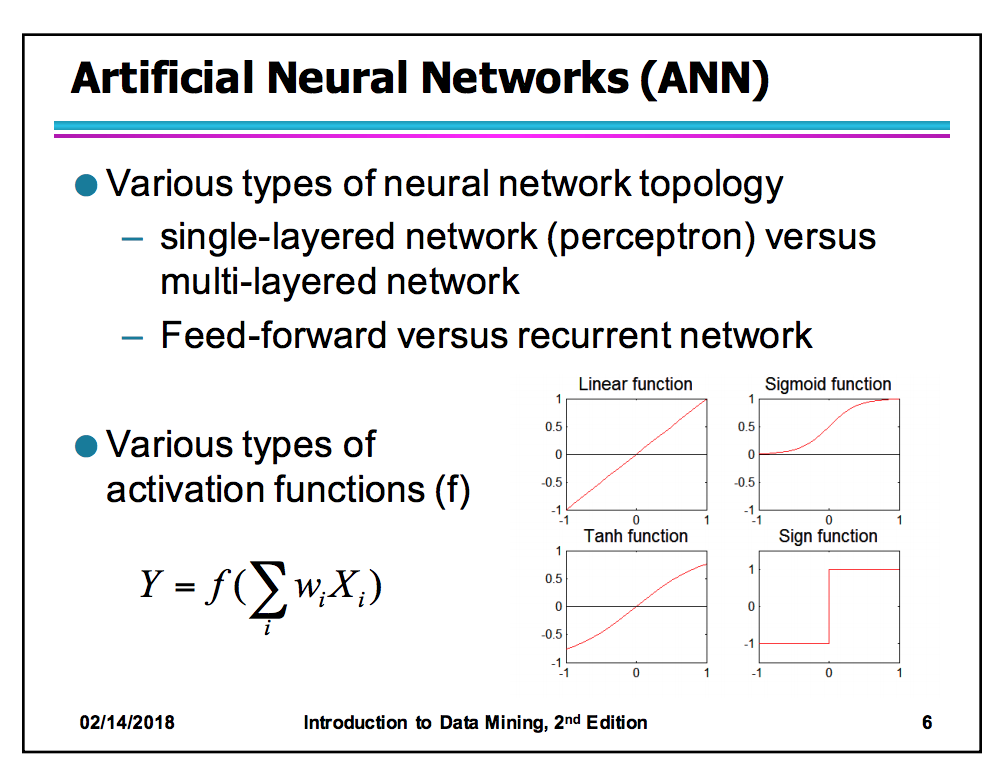


9: **C**? ….from slides: Activation'function'f'must'be'differentiable

10: **D?**: sigmoid and softmax are both activation functions

-Softmax is an activation function, but is it associated with ANN? I’m not hella sure

- yes it is..



Part D:

1. Because we use one hot encoder to perform “binarization” of the category and include it as a feature to train the model.

2. If one of the conditional probabilities is zero, then the entire expression becomes zeroNeed to use other estimates of conditional probabilities than simple fractions (slide 18)

3. noise'refers'to'modification' of'original' values. Role is to make the model conform too closely to slightly inaccurate data can infect the model with substantial errors and reduce its predictive power.

4.the ratios between the majority and minority classes are imbalanced.

5.To separate the points of XOR, you'll have to use at least two lines (or any other shaped functions). This will require two separate perceptrons.

6.Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

7. Define a centroid = average of all the points in a cluster (slide 12)

8.Slides says: OR Google says **Cross**-**validation** is primarily **used** in applied machine learning to estimate the skill of a machine learning model on unseen data

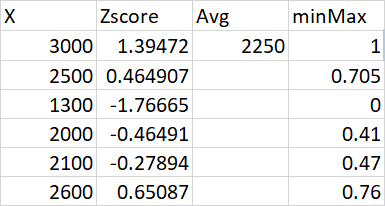
9. Naive Bayes is Naive because it assumes the data is independent of one another:

<https://towardsdatascience.com/whats-so-naive-about-naive-bayes-58166a6a9eba>

10. Why are decision trees considered Greedy? **Uses Hunt’s algorithm which is greedy-- makes the optimal decision each step of the way; (Nodes'with'purer class'distribution'are' preferred)**

E.)  
 1.) Given dataset and std. Dev, (537.75) and mean(2250), apply min-max and z-score normalization techniques

(check this out, but not sure: <https://www.socscistatistics.com/tests/ztest/zscorecalculator.aspx>)



**z score = 1.39, 0.46, -1.77, -0.46, -0.28, 0.65**

E2. How to do this??? <https://people.revoledu.com/kardi/tutorial/KNN/HowTo_KNN.html>

ANYBODY GOT THE WORK FOR THIS???!!

LMAO

?

**F.) Analytical Thinking**

Below are 3 different linear models with r^2 and rmse given

Model 1, (line under dots) RMSE : 0.63 R^2 : 0.47

Model 2, (line close to dots) RMSE: 0.02 R^2: 0.85

Model 3, (line through dots) RMSE: 0.11, R^2: 0.81

**A.) Which model seems to underfit data and explain your answer?** Model 1

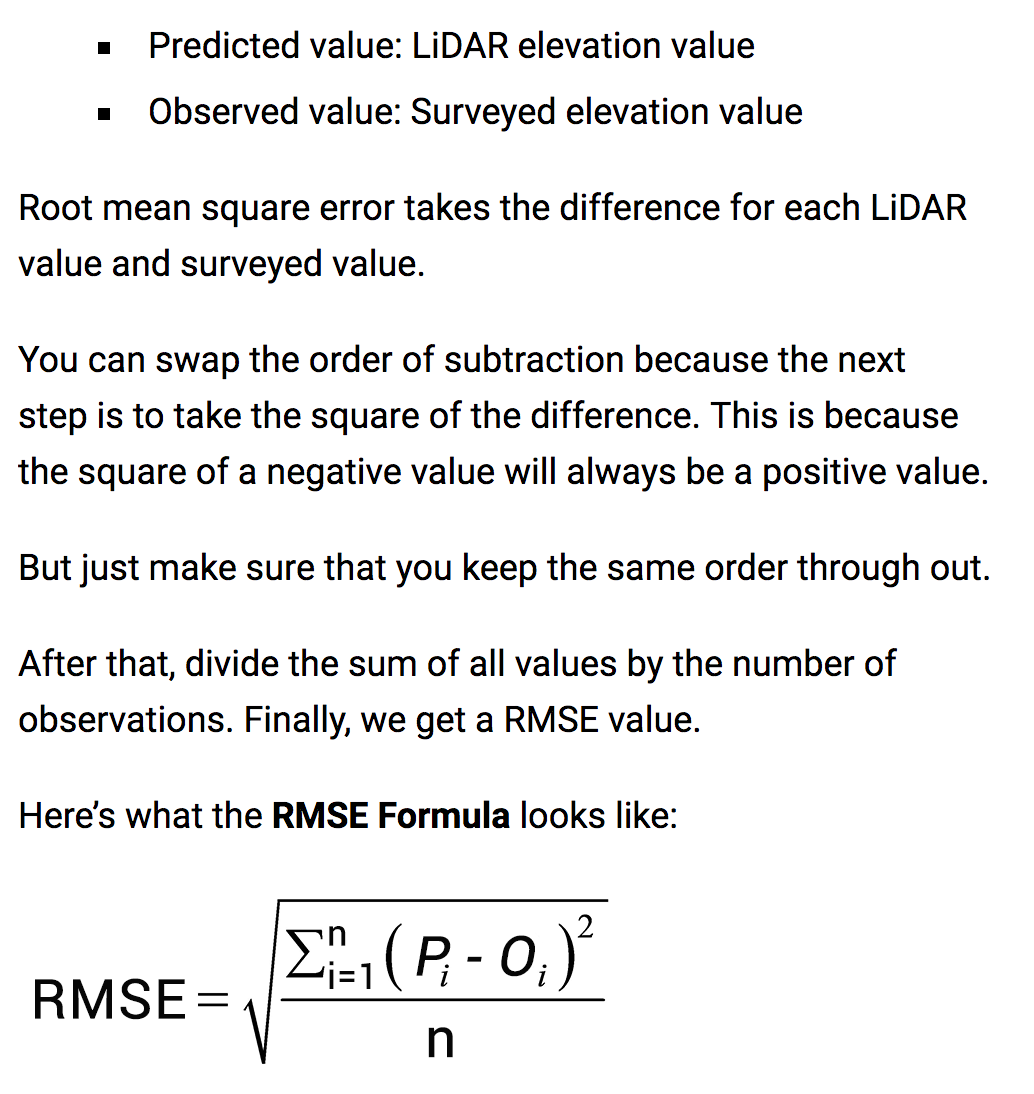
**B.) Which model seems to overfit data and explain your answer?** Model 2

**C.) Which model seems to have more accuracy and why?** Model 3

**D.) Which model has line that explains most variance in data?**

Model 2, because it has the largest r squared value

**E.) How is RMSE calculated?**



**2.) Describe preprocessing with right ordering of steps**

Where Can i find this??

● Aggregation ● Sampling ● Dimensionality Reduction ● Feature subset selection ● Feature creation ● Discretization and Binarization ● Attribute Transformation

Tutorial 4?

Slide 69

· Data Quality Issues: Checked for missing values which were marked with “?” and changed them to “NaN”

· Calculated median value of that certain attribute and replaced the Nan with the median value

· Missing Values: Dropped the missed values on rows

· Outliers:

1. Converted that certain attribute to numerical and plotted the graph which depicts that the “incident-id” is different from other plots which is due to the high numbers of incidents.

2. Derived the z score by standardization process.

3. Discarded the columns that were out of interval

· Duplicate Data: Checked the numbers of duplicated rows and removed them

· Shuffling Data frames: Shuffling the data frame randomly

· Sorting Data frames: Sorted the data frame by using the “State” attribute

· Saving Data frames: Saved the result of data frame

· Feature Normalization: Applied normalization and calculated z-score to set the numbers in a standard form

· Missing Values: Dropped the rows with NA values and replaced them with median values.

· Concatenating: Concatenated the rows and columns to create a new data frame

· X/Y TensorFlow Function:

1. Read the file “gun-violence.csv”

2. Applied label encoding converted the text fields to numeric through one of the attributes

3. Applied one hot encoding converted categorical variables to a form that can provide machine learning algorithms to do a better prediction.

4. Joined the columns in a new data frame

· Training and Test Split: The data is split into 2 different data such as training and test which contains a known output and a model in order to be generalized to the validated data.

· Aggregation: combined multiple objects into a single object by the following:

1. Reducing the size of data

2. Changing the granularity of analysis

3. Improving the stability of the data

· Sampling:

1. Initially, displayed first few rows from the dataset.

2. Then we selected a sample of size n randomly from the dataset without replacement.

3. Selected a sample of percentage n randomly from the dataset without replacement.

4. We finally performed sampling with the replacement where we observe some duplicate instances.

· Discretization: We used functions to distribute the attribute values to plot in a histogram.

· Principal component analysis:

1. Read each image file to a feature values in order to convert PCA to an image dataset.

2. We are storing all the image files in a data frame object called projected.

3. Considering all the projected values obtained in the previous step, we then plot all the values in a graph.