Test KEN3450-2022-NormalPeriod

Test ID: 62574
Folder: /Preview
Version: 1.7
Randomised: No
Last modified: Wednesday, 9 march 2022 11:27:00
Number of questons: 24
Blocks: Fixed
Display questions once: No
Tools: None
Test time: 120 minutes
Maximum score: 50 pt.
Chance score: 7.37 pt. / 15%
In test set with: -

Warm-up questions

Question order: Fixed
This block contains 5 questions. For questions 1-4, just pick the right(s) answers. For question 5 you have to match each model with a figure.

Question 1 − MC−loss1 − 103715.2.0

Which of the following functions are NOT reasonable loss functions? Note that $\hat{y}$ is the prediction and y is the actual value. Assume we can find the optimal parameters for each loss function.

Hint: Think about how we use the loss function in the modeling process to find the optimal model parameters.

A $y - \hat{y}$

B $\frac{1}{y - \hat{y}}$

C $(y - \hat{y})^2$

D $e^{-(y-\hat{y})^2}$

Since we minimize the loss function to find the optimal model parameters, a loss function must output higher values when the prediction yˆ is far away from y and lower values when yˆ is close to y.

Option D does not have this property, so we exclude it.

Choice A also is not reasonable, because the optimal y^ is infinite, since we are minimizing the loss function

Choice B is more complicated. The optimal yˆ would be very slightly above y, as it would lead to a large negative loss. But if yˆ is ever-so-slightly less than y, we have very large positive loss, and if y = yˆ, which would be a perfect prediction, the loss is undefined, so this is an unreasonable loss function.

Question 2 − Sampling1 − 103770.2.0

A recent study by Women In Data Science (WiDS) estimates that 80% of the world's data scientists identify as male, 15% identify as female, and 5% who do not identify as either. Terrible, right! Jerry wanted to check how gender distribution in one imaginary class (let's call it Introduction to Data) compares to that of the study conducted by WiDS.

Per enrollment statistics, this class has 250 students. During the first lecture Jerry asked all 220 students who attended whether they identified as female. Jerry received a 97% response rate, and found that 43% of respondents were female, and 57% were not female.

What is the population of interest? All students in Introduction to Data

What is the sampling frame? All students who attended the first lecture

Suppose we believe this sample is representative of the population of students taking Introduction to Data. In other words, we expect around 43% of all students in Introduction to Data to be female. Select one of the following reasons why our assumption may, or may not, hold true. Assume any individual female is just as likely to attend live lecture as any given non-female.

It will hold true. Our sample size is large enough, so there is likely little variability from our sample estimates in the population.

What type of sampling technique are we using in the problem above? Choose one.

Convenience sampling

[List]

All students in Introduction to Data

All data scientists in the world

All data scientists in the WiDS study

All students who attended the first lecture

[List]

All students who attended the first lecture

All data scientists in the world

All data scientists sampled in the WiDS study

All students in Introduction to Data

[List]

It will hold true. Our sample size is large enough, so there is likely little variability from our sample estimates in the population.

It will not hold true. We may have selection bias, as the people who attend live lectures are self-selecting.

It will not hold true. We have significant sampling bias due to our sample being different from our population.

It will hold true. Our response rate among our sample was high, so the response rate among the Introduction to Pigs population must be similar.

[List]

Convenience sampling

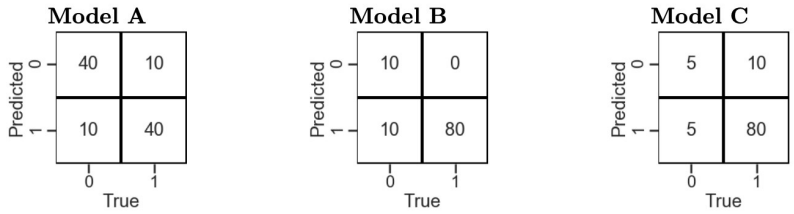
Replacement sampling

Simple random sampling

Quota sampling

Question 3 − Classifier1 − 103761.1.0

Suppose we fit three classifiers which produce the following confusion matrices:



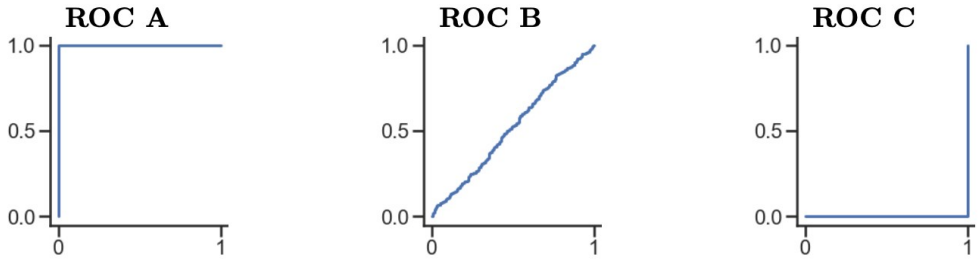
The model with the highest precision is C

The model with the highest recall is B

[List]	[List]
C	**B**
A	A
B	C

Question 4 − Classifier 2 − 103763.2.1

Suppose we fit three more classifiers and plot the ROC curves for each classifier on the test set. The test set contains 100 points: the first 50 points are labeled 0 and the second 50 points are labeled 1. Determine which models produce each ROC curve.



Predicts P(Y = 1|X) using a random number between 0 and 1	ROC B
Assigns P(Y =1|X)=0.3 to the first 50 points and P(Y = 1|X) = 0.4 to the second 50 points.	ROC A
Assigns P(Y =1|X)=0.8 to the first 50 points and P(Y = 1|X) = 0.6 to the second 50 points.	ROC C

Question 5 − timeseries − 103791.1.0

Which of the following statements are true concerning the autocorrelation function (ACF) and partial autocorrelation function (PACF)?

i) The ACF and PACF will always be identical at lag one whatever the model
ii) The PACF for an MA(q) model will in general be non-zero beyond lag q
iii) The PACF for an AR(p) model will be zero beyond lag p
iv) The ACF and PACF will be the same at lag two for an MA(1) model

A (ii) and (iv)

B (i) and (iii)

C (i), (ii) and (iii)

D (i), (ii), (iii) and (iv)

Visualization

Question order: Fixed

In this block you will see 2 visualizations selected from among the top posts on r/datais- beautiful, a subreddit devoted to data visualization. Following each visualization, you will be asked to point out one flaw with the figure, referencing concepts from lecture.

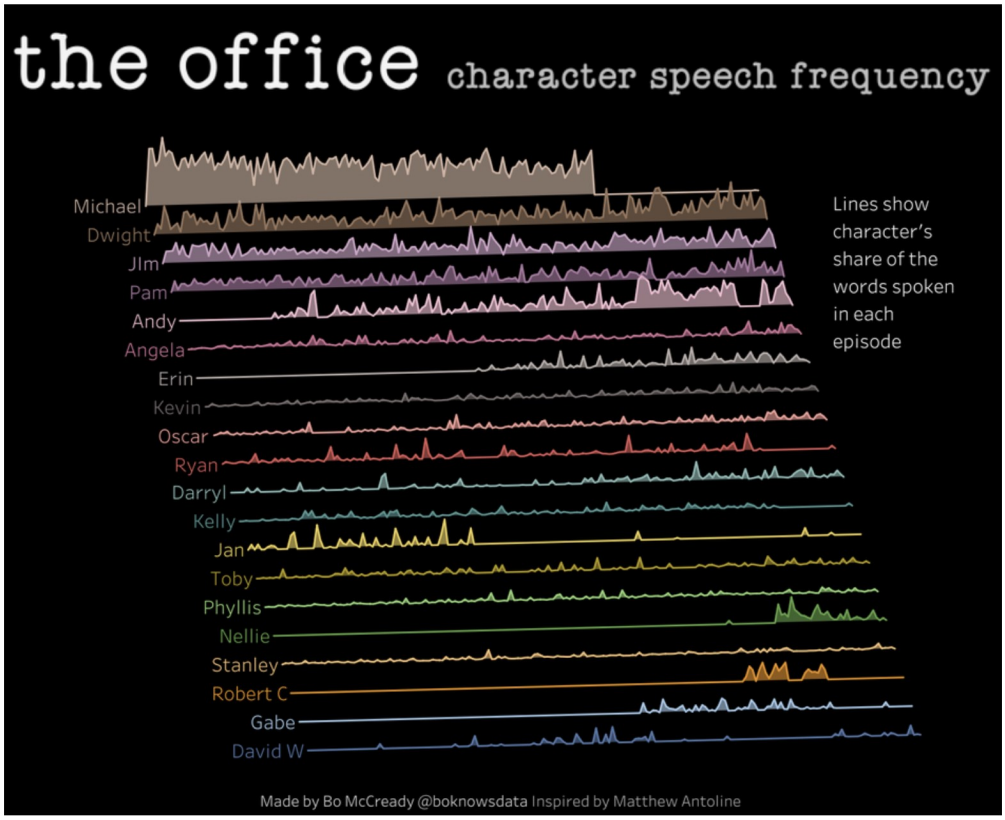
While I have some clear answers in mind, I will be lenient when grading this question. As long as you point out some aspect of the image, and explain why it is a flaw, you will receive full credit. Answers such as ”there is no flaw” or answers about the underlying data (and not the image itself) will receive no credit. Please keep your answers concise.

Note that the titles for these plots are given directly above the image—answers such as ”bad title” or ”missing title” will also receive no credit.

Question 6 − Vis1 − 103706.1.1

Name at least one fault with this visualization

”The Office Character Speech Frequency” by u/BoMcCready



Grading instruction

Criterion 1 (Number of points: 2)

Name at least one fault with this visualization

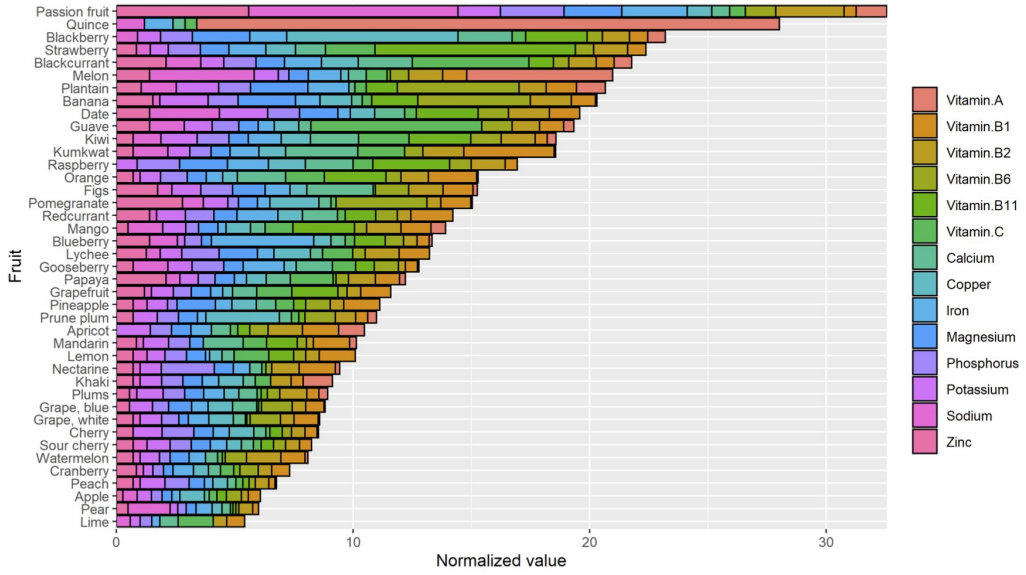
1) The x-axis is not aligned for each character, making it hard to compare different characters' speech frequency at the same point in time.
2) The x-axis is not labeled

other ideas:
3) color not adding any information
4) type of plot not proper

Question 7 − Vis2 − 103716.1.1

Name at least one fault with this visualization.

”An apple a day, might keep the doctor away, but it definitely doesn’t contain the most vitamins and minerals in comparison to other fruits!” by u/Houses_of_Nick



Grading instruction

Criterion 1 (Number of points: 2)

Name at least one falw

1) The baseline is jiggled, so we cannot make a comparison of the difference in levels across fruit for a specific vitamin
2) 40 fruit clustered together makes a plot very cluttered, leading to overplotting
3) We have no idea what the "normalized value" on the x-axis refers to
4) Colors for neighboring vitamins are so similar, making it hard to differentiate between them

Rare Red Rabbits (Regression)

Question order: Fixed

Kermit the Frog and Miss Piggy join a job as data science consultants at a national reserve where there live rare red and blue rabbits. They are tasked with exploring yearly data so that they can develop models to help predict future red and blue rabbit population. They are provided with the following DataFrame, `rabbit`. Specifically, rabbit includes the `location`, `year`, `unique ID`, `name` and `color` of each rabbit.

	Location	Year	ID	Name	Color
0	Site A	2018	A192391839	Alvin	B
1	Site B	2018	B1827333	Kelly	R
2	Site A	2019	M318774443	Peter	R
3	Site C	2017	P8773323	Lapine-Rouge	R
4	Site B	2020	Q382819311	Daisy	B
5	Site A	2019	A192391839	Alvin	B

Additionally, they are provided with another DataFrame, `metadata`. `metadata` contains the year in consideration, carrots eaten in that year, and the number of animal tracks found.

	Year	Carrots Eaten	Tracks Found
0	2018	3817	81273
1	2019	10283	150281
2	2020	30381	300372

Questions 8-12 refer to this problem.

Question 8 – RRR1 – 103860.1.0

Describe conceptually (no code needed) how you would generate a DataFrame where each row corresponds to a year. Each row should also contain the number of animal tracks, carrots eaten, total red rabbit population, and total blue rabbit population for that year. Explain (very shortly) why we prefer this tabularized structure.

Grading instruction

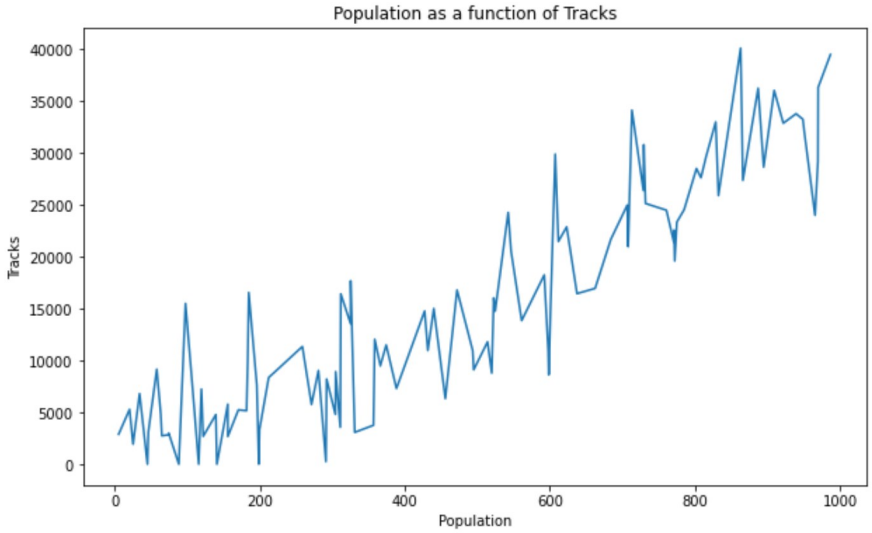
Criterion 1 (Number of points: 1)

We need to "pivot" based on the year (connecting the two dataframes) and then we need to summarize the populations -- different ways to solve this.

We prefer tabularized structure because each row can be a datapoint (instance) and then each column can be a variable (feature/attribute).

Question 9 – Rare Red Rabbits – 103739.2.1

Kermit decides to perform some EDA before diving into some modeling. He decides to focus on the relationship between the population and the number of tracks.



Describe an issue with the visualization above

Grading instruction

Criterion 1 (Number of points: 2)

One fault

Main issue: The choice of plot is incorrect. This should be represented by a scatterplot.

More advanced: The overall trend is masked by the jagged edges

Devin in details: problem with data point when population=0?

Question 10 – RRR2 – 103741.1.1

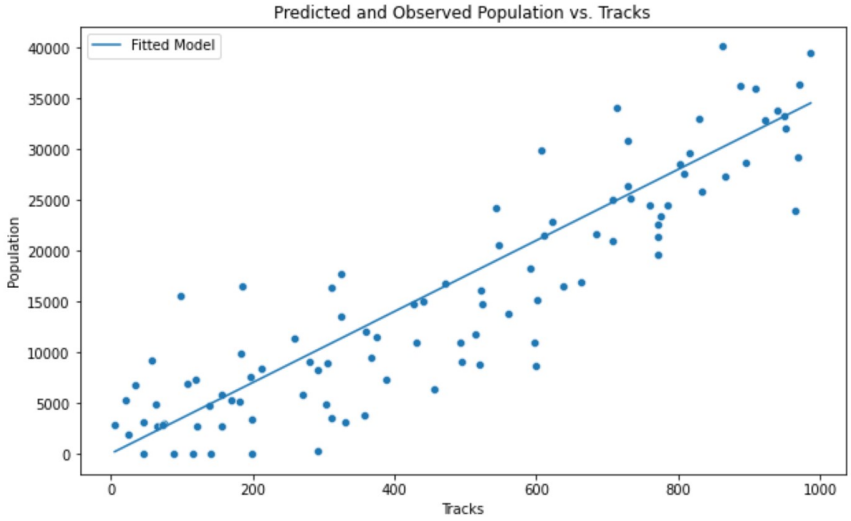
Which of the following is most likely the correlation coefficient (r) between population and number of tracks (refer to the previous figure as well).

A	-1
B	-0.9
C	-0.3
D	0
E	0.3
F	0.9
G	1

This displays a positive correlation, so any non-positive correlations are not possible (-1, -0.9, -0.3 and 0). It isn't perfectly correlated, so a correlation of 1 is impossible. Visually, it is clear that the correlation is closer to 0.9 than to 0.3 since a correlation coefficient of 0.3 would display much more scatter (and essentially indicates that there is a very weak correlation between the two variables). The answer is therefore 0.9.

Question 11 − RRR3 − 103742.3.0

Using his findings from the previous questions, Kermit wishes to predict the total rabbit population, *p*, using the number of animal tracks, *t*, found in the year. He decides to use Ridge regression without an intercept term, and with regularization hyperparameter λ . After fitting his model, he plots his predictions against his observations as shown below. Which of the following are true? Select all that apply.



A The fit is inaccurate due to the large amounts of scatter around the line, which suggests that linear regression is inappropriate.

B There is a curvature in the relationship, which suggests that linear regression is inappropriate.

C The training loss would decrease if the regularization hyperparameter λ was 0.

D The plot displays high variance, which suggests that *λ* should be increased to reduce model complexity.

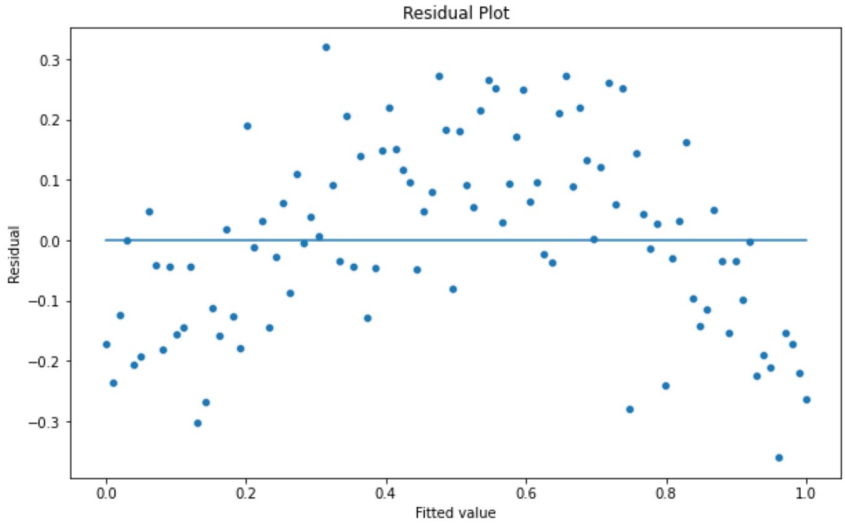
While there is scatter around the line, the general trend suggests that linear regression is appropriate since there is no curvature in the relationship. This rules out the first two options.

Since λ restricts the magnitude of our parameters, removing regularization by setting λ = 0 would improve our fit and therefore reduce training loss.

The plot does display high variance in the observations themselves, not the predictions necessarily. The variance in the observations is not linked to model complexity or the regularization hyperparameter. Further, λ being increased would lead to a poorer fit.

Question 12 − RRR4 − 103743.1.1

Miss Piggy decided to proceed with a linear model (since that's always a good model). She did use MAE (Minimum Absolut Error) as an optimization function. She generates predictions on the test set and create a residual plot as shown below. What does this visualization indicate?



The residual plot should display uniformly random scatter, but this residual plot displays clear curvature in the form of a parabola. The outliers in the residual plot just indicate that our model performed poorly on those points, but it doesn't suggest that linear regression is inappropriate. The residual plot should have no relationship with the fitted value.

A The residual plot displays a roughly normal distribution, which suggests that linear regression is appropriate.

B The residual plot displays roughly uniformly random scatter, which suggests that linear regression is appropriate.

C The residual plot has many outliers, which suggests that linear regression is inappropriate.

D The residual plot has a weak relationship with the fitted value, which suggests that linear regression is inappropriate.

E The residual plot has curvature, which suggests that linear regression is inappropriate.

Rare Red Rabbits Revisited (PCA)

Question order: Fixed

Kermit wants to apply PCA to the rare rabbit dataset from the previous question to understand patterns in rabbit population per location as a function of the year. Provided is a Pandas DataFrame, rabbit pop (shown below), which contains the rabbit population for every particular year and location. Note that not every year and location is shown here.

	2017	2018	2019	2020
Site A	8789	29372	49271	101822
Site B	18573	38317	102847	192742
Site C	402	3928	20212	80272
Site D	4392	28172	93172	203082

Kermit needs to preprocess his current dataset in order to use PCA.

Questions 13-16 refer to this problem.

Question 13 – PCA1 – 103744.1.1

Select all appropriate preprocessing steps used for PCA.

A	Transform each row to have a magnitude of 1 (Normalization)	
B	Transform each column to have a mean of 0 (Centering)	
C	Transform each column to have a mean of 0 and a standard deviation of 1 (Standardization)	
D	None of the above	We can use standardization or centering for PCA. We cannot compute the covariance matrix correctly if the data is not centered with mean 0.

Question 14 – PCA2 – 103745.1.1

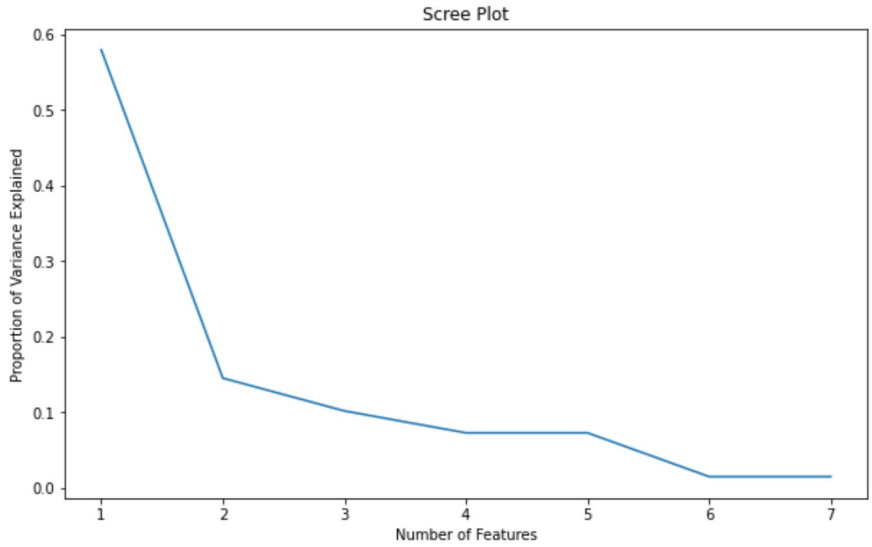
Kermit wishes to apply a transformation to the rabbit population at each site for each year so that he can apply PCA more effectively. What transformation function $f(p)$ would be most effective to apply to the rabbit population p for using PCA?

Hint: Notice the population at each site appears to grow exponentially every year.

A	$\log(p)$	
B	$exp(p)$	Since the population is exponential, we use the log function to linearize the relationship for PCA.
C	p^2	
D	$\sqrt{p}$	

Question 15 − PCA3 − 103746.1.1

Kermit successfully applies PCA and makes a scree plot that is displayed below. How many principal components should Kermit use to capture at least 80% of the variance in the rabbit population data?



A 1

B 2

C 3

D 4

E 5

F 6

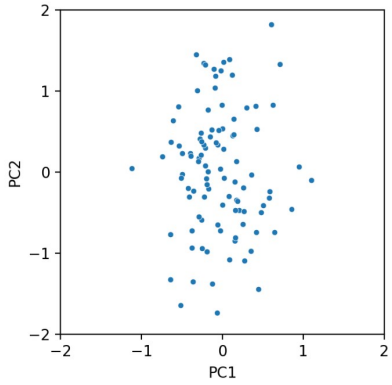
G 7

The first principal component captures around 58% of the variance, the second captures 15%, and the third component captures around 11%. This sums to around 84%, so we need 3 features.

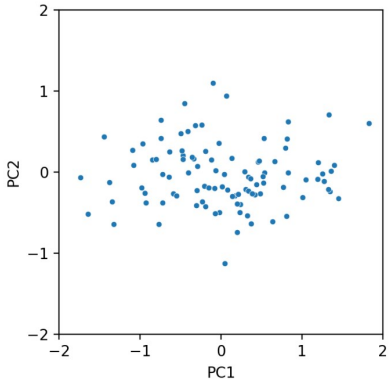
Question 16 − PCA4 − 103747.3.1

We now wish to display the first two principal components in a scatterplot. Which of the following plots could potentially display the first two principal components?
Hint: The scree plot (of the previous question) may be helpful.

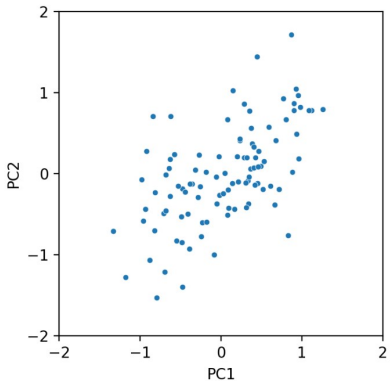
A.



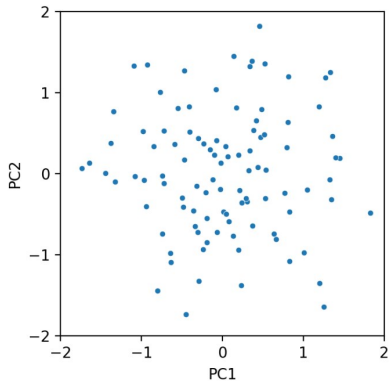
B.



C.



D.



A	A
B	B
C	C
D	D

The first principal component PC1 must capture more variance than PC2 by the properties of PCA. The scree plot from the previous question tells us that PC1 captures nearly 4 times as much variance. Therefore, options A, C, and D don't work because PC1 doesn't have 4 times as much variance in any of them. Additionally, the principal components must be axis-aligned, which is another mark against C. Therefore, the answer is B.

Trees and Forests

Question order: Fixed
Recall that a random forest is created from a number of decision trees, with each decision tree created from a bootstrapped version of the original training set. One hyperparameter of a random forest is the number of decision trees we train to create the random forest.

We define *T* to be the number of decision trees used to create the random forest. Let's say we have two candidate values for *T*: A and B. We want to perform K - fold cross-validation to determine the optimal value of *T*. Assume A, B, and K are integers.

Questions 17-20 refer to this particular problem.
Question 21 refers to trees in general.

For all questions, justify your answers shortly

Question 17 – RF1 – 103748.1.0

In this cross-validation process, how many random forests will we train? Your answer should be in terms of A, B, and/or K and should be an integer.

2 * K

Grading instruction

Criterion 1 (Number of points: 2)

Question 18 – RF2 – 103749.1.0

In this cross-validation process, how many decision trees will we train? Your answer should be in terms of A, B, and/or K and should be an integer.

Grading instruction

Criterion 1 (Number of points: 2) (A+B) * K

Question 19 – RF3 – 103750.1.0

Let's say we pick three hyperparameters to tune with cross-validation. We have 5 candidate values for hyperparameter 1, 6 candidate values for hyperparameter 2, and 7 candidate values for hyperparameter 3. We perform 4-fold cross validation to find the optimal combination of hyperparameters, across all possible combinations.

In this cross-validation process, how many random forests will we train? Your answer can be left as a product of multiple integers, e.g. "1 * 2 * 3", or simplified to a single integer, e.g. "6". (These are not the correct answers to the problem).

Grading instruction

Criterion 1 (Number of points: 2) 4*5*6*7 = 840

Question 20 – RF4 – 103751.2.0

Here is some code that attempts to implement the cross-validation procedure described above. However, it is buggy. In one sentence, describe the bug below.

You may assume the following:

- `X_train` is a `pd.DataFrame` that contains our design matrix, and `Y_train` is a `pd.Series` that contains our response variable, both for the full training set.
- Assume `ensemble.RandomForestClassifier(**args)` creates a random forest with the appropriate hyperparameter values. The bug is not on this line.
- The candidate values for each hyperparameter have been loaded into the lists cands1, cands2, and cands3, respectively.

```
1: from sklearn.model_selection import KFold
2: from sklearn import ensemble
3: import numpy as np
4: import pandas as pd

6: kf = KFold(n_splits = 4)

7: cv_scores = []
8: for cand1 in cands1:
9:     for cand2 in cands2:
10:        for cand3 in cands3:
11:            validation_accuracies = []
12:            for train_idx, valid_idx in kf.split(X_train):
13:                split_X_train, split_X_valid = X_train.iloc[train_idx], X_train.iloc[valid_idx]
14:                split_Y_train, split_Y_valid = Y_train.iloc[train_idx], Y_train.iloc[valid_idx]

16:                model = ensemble.RandomForestClassifier(**args)
17:                model.fit(X_train, Y_train)

18:                accuracy = np.mean(model.predict(split_X_valid) == split_Y_valid)
19:                validation_accuracies.append(accuracy)
20:            cv_scores.append(np.mean(validation_accuracies))
```

Grading instruction

Criterion 1 (Number of points: 2)

Each iteration of the algorithm trains a random forest on the entire training set, as opposed to the part of the training set that is not reserved for validation.

Question 21 – RF5 – 103754.1.2

Kermit and Piggy are each training their own decision tree for a classification task. Kermit decides to limit the depth of his decision tree to depth 3 while Piggy decides to not set a limit on the depth of his decision tree. When plotting the training error, Kermit's error seems to be much higher than Piggy error. However, when plotting the validation error, Piggy's error seems be much higher than his training error as well as Kermit's error. Andrew is confused and surmises that there must be a bug in his code that is causing this to happen. What happened? Explain. What can he do to improve it? Name at least 3 things he can do to improve his error. Please limit your repsonse to 2 sentences per reason.

Grading instruction

Explanattion (Number of points: 2)

Andrew == Kermit. Most of you interpreted this correctly. If you were confused, it was taken into account in the final grading soup.

3 things (Number of points: 3)

He is not correct. Kermit's high validation error and low training error is due to overfitting. Miss Piggy did not run into this error because she limited her depth to 3. For Kermit to improve his validation error, he should try to limit the depth of his tree, try pruning his decision tree, preventing splits that have less than 1% of the samples, or using Random forests.

Regularization

Question order: Fixed

This last block contains 2(+1) questions on regularization. For question 22 you just need to select the right(s) answer(s). For question 23 you need to select the right(s) answer(s) and then justify your choice(s) on question 24.

Question 22 – Reg1a – 103782.1.0

Which of the following are indications that you should regularize? Select all that apply.

- **A** Our training loss is 0.
- **B** Our model bias is too high.
- **C** Our model variance is too high.
- **D** Our weights are too large.
- **E** Our model does better on unseen data than training data.
- **F** We have linearly dependent features.

Question 23 – LRR2a – 103783.3.0

Suppose we have a data set which we divide into 3 equally sized parts, A, B, and C. We fit 3 linear regression models with $L2$ regularization (i.e. Ridge regression), X, Y , and Z, all on A. Each model uses the same features and training set, the only difference is the λ used by each model. Select all below that are always true. We also use the notation $Loss\,(m, n)$ to refer to the loss of model m on dataset n .

In the next question, justify your selection(s) shortly.

- **A** Suppose Z has the lowest average loss on B. Model Z will have the lowest average loss when evaluated on C
- **B** If A and B have the same exact mean and variance, the average loss of model Y on B will be exactly equal to the average loss of Y on A.
- **C** If $\lambda = 0$ for model X , then $Loss(X, A) < Loss(Y, A)$ and $Loss(X, A) < Loss(Z, A)$.
- **D** If $\lambda_Y < \lambda_Z$, then $Loss(Y, A) \leq Loss(Z, A)$
- **E** If $\lambda_Y > \lambda_Z$, then $Loss(Y, B) \geq Loss(Z, B)$
- **F** None of the above.

Question 24 – LRR2b – 103784.2.0

Justify your answer to the previous selection.

Grading instruction

Criterion 1 (Number of points: 4)

A: Not guaranteed since we don't know the distributions of B, C.

B: Having the same mean and variance does not imply that the data are the same.

C: Since increasing λ increases bias, the loss of X must be less than or equal to the loss of Y, Z on A.

Test matrix

No matrix defined

D: Since Y and Z were trained on A, and Y is less restricted than Z, the loss of Y on A must be less than the loss of Z on A.

E:Even though Z is a more restricted(i.e.simpler)model, it is possible that the dataset B is slightly better for Z. In other words, minimizing training error with a regularized model does not guarantee minimized error on unseen datasets.