Multiple Choice Block

This is a block with 7 questions.

Questions 1 through 6 are multiple-answer. You will get full points if you select the answers that are correct, otherwise partial credit applies depending on how many answers are correct (see Instructions).

Question 7 is a matching question. You need to match the figures on the left column with the corresponding figure on the right column.

Question 1 – Multiple response – Question-ID: 131992 (1 point)

Using the following snippet of a dataset describing restaurants (each row should be a record, i.e. a restaurant) select all statements that are true.

```
"business_id","name","address","phone"
10,"TIRAMISU KITCHEN","033 BELDEN PL","+14154217044"
19,"LIFESTYLE CAFE","1200 VAN NESS AVE","+14157763262"
24,"OMNI S.F. HOTEL","    ","9999999999999999"
42,"The "Best", Food!","500 CALIFORNIA ST","+14156211114"
43,"The "Best", Food!","3716 Cesar Chavez","+14156211114"
```

- **A** There are nested records.
- **B** From the available data the `business id` can be used as a unique index for each record in the dataset
- **C** While the data appears to be quoted there may be issues with the quote character.
- **D** There appear to be no missing values

Question 2 – Multiple response – Question-ID: 131994 (1 point)

Which of the following are reasonable motivations for applying a power transformation? Select all that apply.

- **A** bring data distribution closer to random sampling
- **B** to help visualize highly skewed distributions
- **C** remove missing values
- **D** to help straighten relationships between pairs of variables
- **E** reduce the dimension of the data

Question 3 – Multiple response – Question-ID: 131995 (1 point)

From the following list select all statements that are true for Pandas Data Frames.

- **A** You can always index a record by its row number
- **B** Missing values in string columns are always encoded as `NaN`
- **C** All columns must be of the same type
- **D** All data frames must have an index

Question 4 – Multiple response – Question-ID: 131996 (1 point)

Which of the following are advantages to using AdaBoost with *short* trees (e.g. max depth 4) over random forests with an equal number of *tall* trees (i.e. refined until the leaves are pure)? Select all that apply.

- **A** AdaBoost is better at reducing bias than a random forest.
- **B** AdaBoost is better at reducing variance than a random forest.
- **C** AdaBoost is faster to train.
- **D** AdaBoost is more robust against overfitting outliers in the training data.

A: That is the whole design of Adaboost
B: Bagging and randomizing the split choices are designed to reduce variance (in RF), however Adaboost doesn't actively reduce variance.
C: Adaboost uses shorter trees.
D: AdaBoost is expected to be worse at handling outliers because of the boosting steps needed and the fact that it generally reaches 100% training accuracy.

Question 5 – Multiple response – Question-ID: 131997 (1 point)

A: The subsamples will be more alike, which makes trees more alike
B: true, if some features that domnate/are very strong predictors

Which of the following statements are true about bagging (note: not in a random forest; just bagging alone). Select all that apply.

C: Decision trees can have 100% training accuracy!
D: No, always the same algorithm

- **A** Bagging without replacement is more likely to overfit than bagging with replacement
- **B** Even with bagging, sometimes decision trees still end up looking very similar
- **C** Bagging is often used with decision trees because it helps increase their training accuracy
- **D** Bagging involves using different learning algorithms on different subsamples of the training set

---

Question 6 – Multiple response – Question-ID: 132159 (1 point)

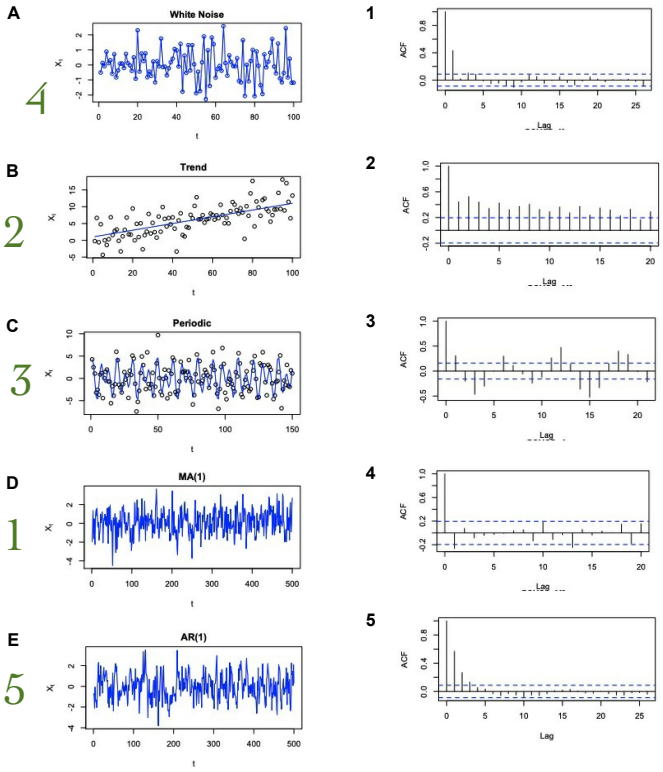
Select all the statements that are true about ROC curves.

- **A** The ROC curve is a better guide for choosing a threshold (separating negative from positive classifications) on real-world data than the threshold suggested by the model (the default way)
- **B** A ROC curve closer to the diagonal line y = x implies that your classifier's loss (risk) is closer to the optimal one
- **C** The horizontal axis represents posterior probability thresholds and the vertical axis represents test set accuracy
- **D** There are (at least) two points on a ROC curve that are not affected by changes in the model. (Note: We are not counting the specific choice of threshold between positive and negative as part of the model)

A: obviously true
B: quite the opposite. being close to y=x suggests random performance
C: Both axes show classification rates
D: true: points (0,0) and (1,1) are always present, since they arise when we only classify as one class

Question 7 – Match (even) – Question-ID: 131998 (5 points)

On the left column, you are shown some timeseries models, namely (a) "white noise", (b) a model with a clear linear trend, (c) a model with a specific periodic pattern in the data, (d) an AR(1) model and (e) a MA(1) model respectively. You are asked to match these models with expected autocorrelation plots (ACF), as shown in the right column. To answer this question drag the ACF plot on the timeseries plot it matches (once done you will see that the selected ACF plot vanishes from the right column and cannot be selected anymore).



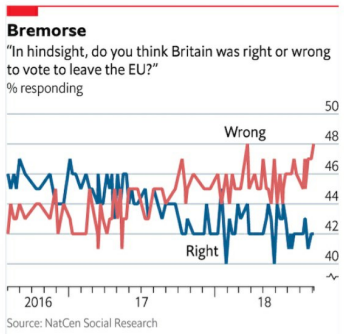
Visualization block

Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Two of their ill-thought-out plots are presented in the next 2 questions (Questions 8 and 9). Consider what aspects of the visualizations are misleading, even faulty, and think of ways in which you can remedy them.

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.

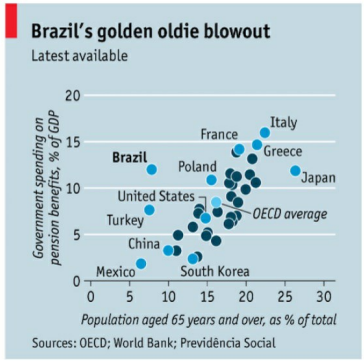
Question 8 – Open-ended – Question-ID: 132000 (1 point)



The continuous line does not give a good impression of the evolution of the response. It would be better to smooth it

Mention at least one thing that is conceptually wrong and/or misleading with the visualization above.

Question 9 – Open-ended – Question-ID: 132001 (1 point)



The coloring of the nodes is not properly explained. Also, some countries are shown but not all.

Mention at least one thing that is conceptually wrong and/or misleading with the visualization above

Principal Component Appliances

The big box electronics store, Good Buy, needs your help in applying Principal Components Analysis (PCA) to their appliance sales data. You are provided records of monthly appliances sales (in thousands of units) for 100 different store locations worldwide. A few rows of the data are shown in the figure below.

location	monitors	televisions	computers
Bakersfield	5	35	75
Berkeley	4	40	50
Singapore	11	22	40
...	...	...	...
Paris	15	8	20
Capetown	18	12	20
SF 4th Street	20	10	5

Suppose you perform PCA as follows. First, you standardize the 3 numeric features of the dataset (i.e., transform to zero mean and unit variance). Then, you store these standardized features into X and use singular value decomposition to compute $X = U\Sigma V^T$

Answer the questions in this block (Questions 10 through 13) keeping this in mind.

Question 10 – Multiple response – Question-ID: 132002 (2 points)

Select all the statements that are true about PCA.

A If we select only one direction (a one-dimensional subspace) to represent the data, PCA chooses the eigenvector of the sample covariance matrix that corresponds to the least eigenvalue.

B PCA is a method of dimensionality reduction.

C If we select only one direction (a one-dimensional subspace) to represent the data, the sample variance of the projected points is zero if and only if the original sample points are all identical.

D changing the origin of the coordinate system does not change the predictions or the principal component directions

---

Question 11 – Open-ended – Question-ID: 134480 (1 point)

Why we usually standardize the data before applying SVD (or PCA)?

to have comparable scales for variables that are measured in different units

Question 12 – Multiple response – Question-ID: 134272 (2 points)

Recall that running the SVD in Python should entail executing the following:

```
u, s, vt = np.linalg.svd(X, full_matrices = False)
```

To draw a histogram of the data's distribution along the first principal component of X (our original data), which of the following arrays would you visualize?

Reminder: `A.T` in Python gives us the transpose of matrix A

A `(u * s)[:,0]`

B `X @ u[:,0]`

C `(X @ vt.T)[,0]`

D `X @ u.T[:,0]`

E `X @ vt[0,:]`

C would be correct, if there was the [:,0],
if you selected it, no points were removed

Question 13 – Fill in (multiple) – Question-ID: 132006 (3 points)

Suppose you focus on interpreting the first principal component, PC1. Below is the original data, now with PC1 values, as well as the first row of V^T, which below is called V_1^T

location	monitors	televisions	computers	PC1
Bakersfield	5	35	75	-2.543196
Berkeley	4	40	50	-2.238256
Singapore	11	22	40	-0.268337
...	...	...	...	...
Paris	15	8	20	1.438350
Capetown	18	12	20	1.551479
SF 4th Street	20	10	5	2.277241

$V_1^T = [0.59001398, -0.58165848, -0.55996153]$

Interpret PC1's relationship with each feature xi (monitor sales, television sales and computer sales) below. Positively related means that as PC1 increases, xi increases; negatively related means that as PC1 increases, xi decreases.

Monitor sales [ ] positively related

Television sales [ ] negatively related

Computer sales [ ] negatively related

Imaginary University Rentals

A random sample of 76 apartments is collected near Imaginary University. All of the apartments in the sample have between 1 and 6 bedrooms. The variables recorded for each apartment are **Rent** (in euros) and the number of **Bedrooms**. The regression output is:

The dependent variable is Rent
R squared = 62.0%
R squared (adjusted) = 61.5%
s = 364.4 with 76 - 2 = 74 degrees of freedom

Variable	Coeff	SE(Coeff)	t-ratio	p-value
Constant	357.795	111.6	3.2	0.0020
Bedrooms	400.554	36.42	11.0	< 0.0001

Answer the questions in this block (Questions 14 through 16) keeping this model in mind

Question 14 – Open-ended – Question-ID: 132007 (2 points)

The coefficient value for the number of bedrooms (which is approx. 400.6) appears to be too large. A friend of yours suggests that there is something wrong with the model. Do you agree? Justify your answer shortly.

The value of the coefficient depends on the units of the actual variable. In this case #bedrooms is ranging from 1 to 6 and the output variable is rent price (in euros). Therefore, a large coefficient is needed for getting from the independent variable (1-6) to the dependent variable (which is expected to be in the range of hundreds)

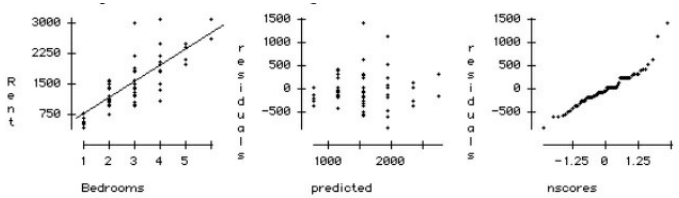
Question 15 – Open-ended – Question-ID: 132008 (2 points)

Explain the meaning of the intercept in the context of this problem and comment on whether you can use it for actual decisions based on the model. Explain shortly your answer.

The intercept, aka 357.795, would be the price of a house with 0 bedrooms (aka a studio). Given that the dataset we are given mentions that the apartments in the sample have 1-6 bedrooms, we actually have no evidence to do the "interpolation", therefore we cannot use it for making any informed decisions

Question 16 – Open-ended – Question-ID: 132009 (4 points)

On top of the model above, we are also given the following plots (from left to right): (a) the scatterplot with the best line fit, (b) the plot of residuals vs. the predicted values and (c) is the normal probability plot of the residuals. Given the model details (as given in the introduction) and these three plots comment on the fit of the model. Explain your answer shortly.

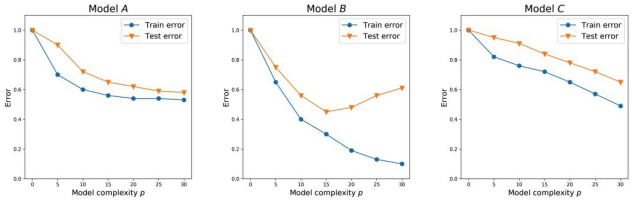


You have the model fit results and 3 plots. Things you could say:

1)R^2 is 62% which is moderate and the coefficient seems significant.
2) Neither the plot of Rent vs. Bedrooms nor the plot of Residuals vs. Predicted show any evidence of curvature.
3) The Residual vs. Predicted plot appears to thicken in the middle, but this is partly due to two large outliers. Ignoring the outliers, the spread is similar throughout the plot.
4) Random residuals: The problem stated that the apartments were a random sample of apartments around the university. Seems okay.
5) The Normal probability plot of the residuals is fairly straight except for the two outliers.

Overfitting

For a specific regression problem, we run 3 different models (A, B and C) and we observe the following train and test errors as a function of model complexity p for three different models.



Answer the questions in this block (Questions 17 through 19) keeping this model in mind

Question 17 – Multiple response – Question-ID: 132011 (2 points)

Mark the values of p and models where the test and train error indicate overfitting.

- **A** model A at p=10
- **B** model A at p=20
- **C** model A at p=30
- **D** model B at p=10
- **E** model B at p=20
- **F** model B at p=30
- **G** model C at p=10
- **H** model C at p=20
- **I** model C at p=30

Question 18 – Multiple response – Question-ID: 132013 (1 point)

Which models, if any, appear to be underfit for all settings of p?

- **A** A
- **B** B
- **C** C

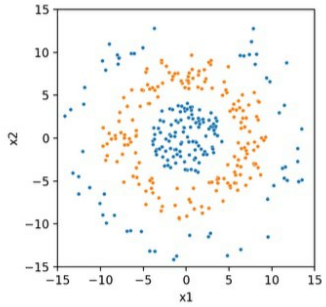
Question 19 – Open-ended – Question-ID: 132012 (2 points)

Which of the models (A, B or C) has the lowest possible bias? Explain your answer shortly.

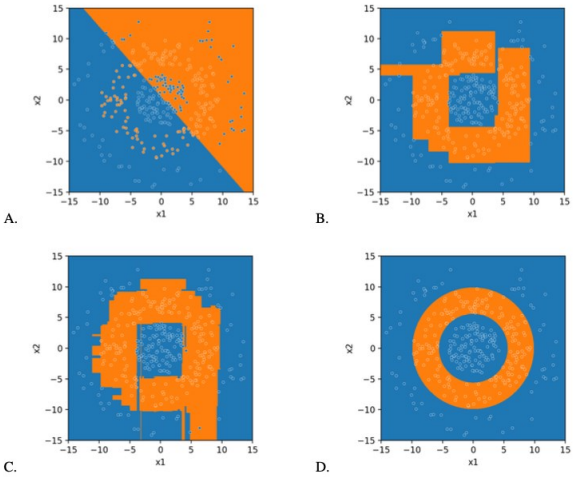
The train error is a good indicator of the bias. B has the lowest possible among the three models, and A and C have around the same.

The Donut problem

Below is a dataset from which we want to create a classifier. We have two features, x1 and x2, and two classes. Assume the orange points are in class 1 and the blue points are in class 0. The points displayed here are training data.



Below are 4 different possible decision boundaries (a.k.a. classifiers) we can generate by letting each model overfit on the training data. The orange regions are areas where new points would be classified as class 1, and the blue regions are areas where new points would be classified as class 0.



Answer the questions in this block (Questions 20 through 24) keeping these figures in mind

Question 20 – Multiple response – Question-ID: 132160 (1 point)

Which of the classifiers (A, B, C, D) has perfect accuracy on the training set? Select all that apply.

Note: Do not try to distinguish borderline points—you may assume points right on the boundary are classified correctly.

A A
B B
C C
D D

Question 21 – Multiple response – Question-ID: 132161 (1 point)

Which of the following models could have generated boundary A?

A Logistic Regression
B Decision Tree
C Random Forest
D None of the above

Question 22 – Multiple response – Question-ID: 132162 (1 point)

Which of the following models could have generated boundary B?

A Logistic Regression
B Decision Tree
C Random Forest
D None of the above

Question 23 – Multiple response – Question-ID: 132163 (1 point)

Which of the following models could have generated boundary C?

A Logistic Regression
B Decision Tree
C Random Forest
D None of the above

Question 24 – Multiple response – Question-ID: 132164 (1 point)

Which of the following models could have generated boundary D?

- **A** Logistic Regression
- **B** Decision Tree
- **C** Random Forest
- **D** None of the above

Exponential Regularization

Jerry is experimenting with regularization functions, and he wants to explore an exponential regularization function defined as follows:

$$Pow(\theta) = \lambda \sum_{i=1}^{p} 10^{\theta_i}$$

where p is the number of predictors, λ is the regularization parameter and θ_i are the coefficients/parameters of the model.

Answer the questions in this block (Questions 25 through 28) keeping this model in mind

Question 25 – Multiple response – Question-ID: 132154 (2 points)

Suppose Jerry fits a multiple linear regression model with L2 loss and the exponential regularization function specified in the introduction, so that the objective function minimized is:

$$\frac{1}{n}\sum_{i=1}^{n}(Y - X\theta)^2 + \lambda \sum_{i=1}^{p} 10^{\theta_i}$$

.

In this function n is the number of data points, p is the number of predictors and Y, X, θ are the real output targets, the input data and the coefficients/parameters of the model respectively, given in vector format.

Which of the following are true if we compare our exponential regularization with L1 (Lasso) and L2 (Ridge) regularization if we use the same λ for all three? Select all that apply. In the next question, justify your answer shortly.

- **A** Exponential regularization always penalizes large positive coefficients more than L2 regularization.
- **B** Exponential regularization always penalizes large negative coefficients more than L2 regularization.
- **C** L1 regularization often performs feature selection by setting parameters corresponding to non-contributing features to 0, but exponential regulariza- tion cannot be used for feature selection.
- **D** L2 regularization is more effective than exponential regularization when we want parameters close to zero.

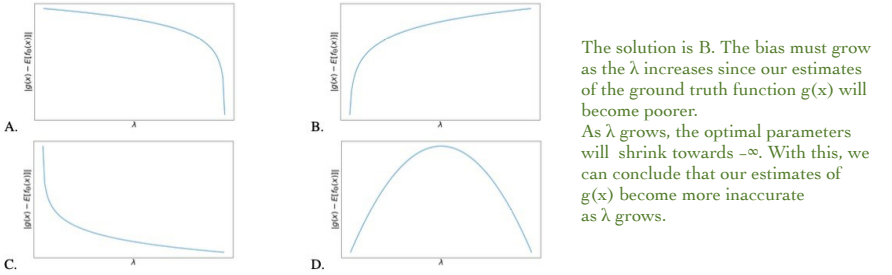
Question 26 – Open-ended – Question-ID: 132155 (3 points)

Justify your answer to the previous question (Question 26) shortly.

For A, B: Exponential regularization as shown in the equation Pow(θ) will penalize large positive coefficients since 10^k for k > 0 grows quickly, but it will not penalize negative coefficients since 10^-k can be bounded by 1 (i.e. it is always less than 1). As a result, exponential regularization encourages parameters to be negative, and ridge (as we know) encourages parameters to be zero, so D option seems correct. C is also correct since the parameters do not even tend to 0 as λ grows; in general, exponential regularization cannot be used for feature selection.

Question 27 – Open-ended – Question-ID: 132157 (3 points)

Which of the following could plausibly be a plot of the magnitude (i.e., absolute value) of model error (we note this in the y axis as the difference between the ground truth function g and the expected (mean) prediction/output of our model) as a function of λ if we use multiple linear regression with exponential regularization? Assume that values on the x-axis are linear from 0 to asymptotically large λ. Justify your answer shortly.



The solution is B. The bias must grow as the λ increases since our estimates of the ground truth function g(x) will become poorer.
As λ grows, the optimal parameters will shrink towards −∞. With this, we can conclude that our estimates of g(x) become more inaccurate as λ grows.

Question 28 – Open-ended – Question-ID: 132158 (3 points)

For the model described in Question 25 (i.e. multiple linear regression model with L2 loss and the exponential regularization function), which of the following is true about the entries of the optimal vector $\hat{\theta}$ as λ approaches positive infinity (i.e. $\lambda \longrightarrow \infty$)?

A. $\hat{\theta} \to -\infty$

B. $\hat{\theta} \to 0$

C. $\hat{\theta} \to \infty$

Justify your answer shortly.

Your thinking in the previous questions should have helped you here. Based on the previous analysis, A is correct.

There are a few ways of approaching this question. In the next page, a detailed answer is provided, however in no way you were not expected to have such analytical answer during the exam.

see next page

Bonus question

This block contains the bonus question

Question 29 – Multiple choice – Question-ID: 134885 (1 point)

How many #datamadness videos we watched during the last session (exam preparation)?

- **A** 7
- **B** 8
- **C** 9
- **D** 10
- **E** 11

End of test KEN3450 - Data Analysis - actual exam time 120 minutes

Intuitive approach

Similarly to ridge regression, as λ grows, the loss function value starts "mattering" less and less to the minimization of the objective function. In other words, the dominant term is the regularization term, and the optimizer (whether it is gradient descent or some other algorithm) will try its best to focus on minimizing that term.

The same holds in the case of exponential regularization, and in fact, it is even more pronounced. Exponential functions blow up very quickly - and it is even more so a priority for the optimizer as λ grows large, to keep the exponentiated θ values as small as possible. In other words, we ideally want each $10^{\theta_i} \to 0$, which means that $\theta \to -\infty$.

By minimizing this more dominant term as λ grows big, we obtain the optimal $\hat{\theta}$ vector that results

Another detailed approach

Some may not be convinced that the optimal θ vector approaches $-\infty$ since all the options listed have an unboundedly large objective value. What makes one of them more correct than the other?
To answer this more concretely, consider a simple example of a constant model fit on one datapoint (y1 = 0) using exponential regularization. The objective is:

$$\arg\min \theta^2 + \lambda 10^{\theta}$$

Consider λ = 0, where there is no regularization. In this case, the clear optimal answer is θ = 0 as expected.
Consider $\lambda = 10^{100}$. While this function can't be minimized through calculus very easily, we will use intuition to approximate an optimum. For θ = 0, we incur an enormous regularization penalty of 10^{100}. For values of θ between −95 and 0, we incur loss values on the order of magnitude of $(90)^2 \approx 8 \cdot 10^3$ at worst, but our regularization penalty is between $10^{100} \cdot 10^{-95} = 10^5$ and 10^{100}. Clearly, none of these minimize the dominant term once λ becomes large.

In this case, consider that $\hat{\theta}$ would be near -100, which makes sense since the regularization penalty is bounded by fairly small values given $\theta \approx 100$ (around 1). We could apply similar analyses to conclude that for any large $\lambda = 10^k$ for $k > 0$, the optimal $\hat{\theta} \approx -k$.
Note that this is just an illustrative example for a simple constant model!