Test KEN3450-2024-Normal

Maastricht University

Test instruction

Instructions to students:

• This is an open-book and open-notes exam. You are allowed to have any books as well as your notes (either in printed form or handwritten notes).
• The exam consists of 30 (+1) questions split in 7(+1) blocks. Some of the questions are fast to answer. Questions within each block might be related, so you may have to refer to previous questions in the same block or in the introductory text of each block.
• This exam sums to 50(+1) points and accounts for 69% of your final grade. The other 30% comes from the practical part and 1% is the participation grade.
• Answer every question in the online form. Use the sheets provided for brainstorming and scratch space.
• Ensure that you properly motivate your answers when asked (except for multiple choice, true/false, etc.).
• Provide clear explanations. Answers that cannot be easily understood may lower your grade.
• You are allowed to use a simple calculator (from DACS approved list).
• You are not allowed to have a communication device within your reach, nor to wear or use a watch.
• You have to return all pages of the exam (and the chromebook as well). You are not allowed to take any sheets, even blank, home.
• For multiple answer questions, the points of that question are split over the possible answers. You get full points if you select all answers that are correct. If you select wrong answers, then we subtract half points. You cannot get negative grade for a question. For example, if a question gives you 2 points, has 4 possible answers and 2 are correct, if you select all 4 answers then you will get 0.5+0.5 (from selecting the two correct ones) and -(0.25+0.25) (from the wrong ones) so in the end you get 0.5 points out of 2.
• If you think a question is ambiguous, or even erroneous, and you cannot ask during the exam to clarify this, explain this in detail in the space reserved for the answer to the question. If the question is of closed form, use one of the open-ended questions for this purpose.
• If you have not registered for the exam, your answers will not be graded, and thus handled as invalid.
• Success! Break a keyboard!

Cleaning and EDA

Question order: Fixed

Let's take a look at the California Air Quality Index (AQI) for 2017. The following cells and outputs are for your reference.

```
> aq = pd.read_csv("./air_quality_final.csv", index_col=0)
> aq.head()
```

	Date	AQI	COUNTY_CODE	COUNTY	LAT	LON
0	01/01/2017	24.0	1	Alameda	37.687526	-121.784217
1	01/02/2017	19.0	1	Alameda	37.687526	-121.784217
2	01/03/2017	NaN	1	Alameda	37.687526	-121.784217
3	01/04/2017	15.0	1	Alameda	0.000000	0.000000
4	01/05/2017	20.0	1	NaN	37.687526	-121.784217

```
> aq.iloc[49437:49442]
```

	Date	AQI	COUNTY_CODE	COUNTY	LAT	LON
49437	01/01/2017	NaN	113	Yolo	38.534450	-121.773400
49438	01/02/2017	15.0	113	Yolo	38.534450	-121.773400
49439	01/03/2017	36.0	113	Yolo	38.534450	-121.773400
49440	01/04/2017	18.0	113	Yolo	37.995239	-121.756812
49441	01/05/2017	16.0	113	NaN	38.534450	-121.773400

```
> aq.describe()
```

	AQI	COUNTY_CODE	LAT	LON
count	49810.000000	49812.000000	49812.000000	49812.000000
mean	38.270167	56.169678	36.294056	-119.859815
std	24.750558	30.486150	2.235560	2.099002
min	0.000000	1.000000	0.000000	-124.203470
25%	21.000000	29.000000	34.144350	-121.618549
50%	35.000000	65.000000	36.487823	-119.828400
75%	52.000000	77.000000	37.960400	-118.147294
max	537.000000	113.000000	41.756130	0.000000

```
> print(aq['COUNTY'].nunique())
51
```

Question 1 – EDA1 – 164465.1.6

Select all that apply. You can assume that the records shown in the block introduction are indicative of the complete dataset.

A Supposing that there is a one to one mapping from COUNTY CODE to COUNTY, we can extrapolate the value of COUNTY for index 4 (NB: index here refers to the dataframe index)

B Grouping by COUNTY is equivalent to grouping by LAT, LON

C DATE can be used to uniquely identify every record in the dataset

D LAT/LON values seem to be correct and have no errors

For B: there are different latitude/longitudes for a country (e.g. look at row with index 3)
For C: dates are not unique
For D: the zeros are indicative of faulty entries

Question 2 − EDA2 − 164466.1.2

Select the best plot to visualize the AQI for Los Angeles, San Diego, San Francisco, Humboldt, and Inyo counties over the first 7 days of January 2017.

- **A** Stacked bar plot
- **B** Side by side line plot
- **C** KDE (kernel density estimation) plot
- **D** Side by side violin plot

Question 3 − EDA3 − 164468.1.2

Select the best plot to visualize the distribution of site locations by latitude and longitude.

- **A** histogram
- **B** scatterplot
- **C** barplot
- **D** 1-dimensional KDE plot

Jerry Reed's Visualization Trawler

Question order: Fixed

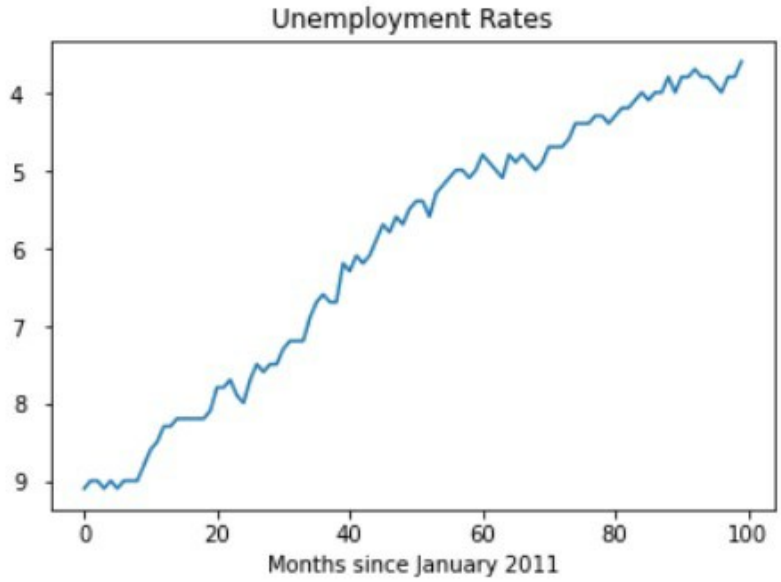
You have just bought some visualizations from the shady visualization seller, Jerry Reed. However, there are some flaws in the visualizations he's sold you. In the next 2 questions you are asked to find issues with the visualizations.

While I have some clear answers in mind, I will be lenient when grading this question. As long as you correctly identify some flawed aspect of the image and explain your answer thoroughly, you will receive full credit. Answers such as "there is no flaw" or answers about the underlying data (and not the plot itself) will receive no credit. Please keep your answers concise—nothing more than a couple of sentences.

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 4 – Plot1 – 164415.1.1

List two aspects of this plot that are incorrect or misleading.



Possible answers:

flipped y axis

missing axis labels

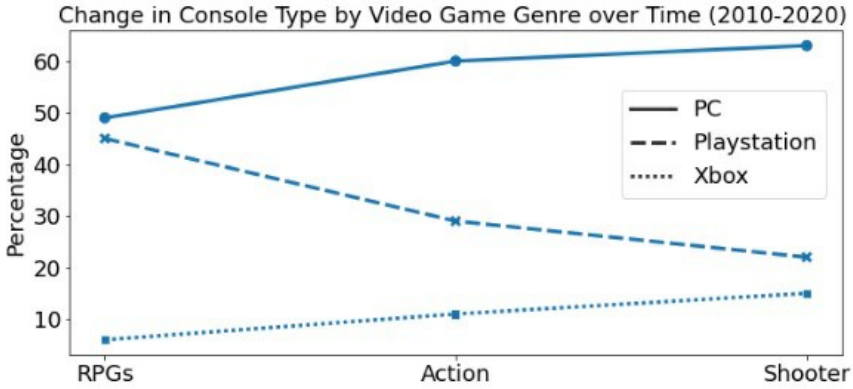
notes after grading: the lineplot is not a problem here

Grading instruction

2 aspects (Number of points: 1)

Question 5 – Plot2 – 164416.1.2

Below, "Console Type" refers to PC, Playstation and Xbox; and "Video Game Genre" refers to RPGs, Action, and Shooter. Describe two flaws with this plot.



Possible answers:

plots categorical data as lineplot (a barplot is better)

mentions time in title, but time is not shown

Grading instruction

2 aspects (Number of points: 1)

Dungeons and Data

Question order: Fixed

Snog the smog owns a drugstore near the village of Grimstone. One day, in a local dungeon, he discovers an ancient recipe for a health potion.

To test the recipe, Snog heads into Grimstone to find the first 50 villagers he sees, knocks them out with his club, and drags them back to his lair. Each of them sustains a large amount of damage from the club, as measured in "health points" (HP). Snog gives each of them a free sample of his new potion and observes how many HP are restored. He records various attributes about each study participant, as well as the number of HP that were restored by his potion as shown in the table below.

	age	species	str	dex	con	int	wis	cha	restored
0	22	human	15	14	12	13	9	11	6
1	25	dwarf	19	14	18	8	9	12	13
3	36	human	13	9	13	7	19	17	10
...	...	...	...	...	...	...	...	...	...
48	121	gnome	17	10	25	16	13	21	9
49	372	elf	9	24	12	16	18	12	8

The variables are:

`age`: in years (as an integer)
`species`: dwarf, elf, gnome or human (as a string)
`str`, `dex`, `con`, `int`, `wis`, `cha`: strength, dexterity, constitution, intelligence, wisdom, and charisma. Each is an integer between 1 and 30. For this problem, it doesn't matter what they mean.
`restored`: the amount of HP restored by the potion (response variable, as an integer)

Snog wants to figure out about how many health points are usually restored, and also use regression modeling to figure out whether some types of villagers can expect to gain more from the potion than others.

For the questions in this block you may always refer back to the information provided here.

Question 6 − Snog−Q1 − 164380.1.3

Snog's sampling frame is the set of all villagers in Grimstone. Which of the following is true about the sample that Snog gets?

A it is a convenience sample

B it is a probability sample

C it is a random sample with replacement

D we can be reasonably confident that the sample is representative of the sampling frame

Question 7 – Snog–Q2 – 164381.1.5

Snog's first idea is to use a constant model (no parameter) to predict `restored`. He decides to minimize the MSE. What will be the optimal constant prediction?

- **A** The maximum of the `restored` column
- **B** The minimum of the `restored` column
- **C** The mean of the `restored` column
- **D** The median of the `restored` column
- **E** The mode of the `restored` column

Question 8 – Snog–Q3 – 164384.1.4

What type of variable is the `species` column?

- **A** Categorical, ordinal
- **B** Quantitative, continuous
- **C** Categorical, nominal
- **D** Quantitative, discrete

Question 9 – Snog–Q4 – 164390.1.2

Snog wants to run a multiple linear regression to predict the response, `restored`, from all other variables. He needs to do something about the variable `species`, which is currently represented as a string, instead of a numerical value. He plans to use one-hot encoding. How many dummy variables would Snog need?

- **A** 2
- **B** 3
- **C** 4
- **D** 5

Question 10 – Snog–Q5 – 164395.1.6

Snog writes a Python script to implement linear regression from scratch (using Ordinary Least Squares (OLS)), and uses it to run a regression with the full data set of 50 observations, using all of the other variables to predict the response `restored`. He checks the MSE for his model's predictions, and he finds that it is even higher than the MSE for the constant model from Question 7. Which of the following do you think is the best explanation for what happened?

- **A** He probably overfit the data, and that might be why he got such a high MSE.
- **B** He probably underfit the data, and he needs to add more features to the regression until his MSE goes down.
- **C** This should never happen with a correctly implemented OLS method. He must have a bug in his code.

For A: Snog used the **full data** to train both models and computed the MSE on the same data, so even if he is overfitting, we would not know based on the training MSE
For B: The constant model is the simplest model there is; any OLS LR model would be more complex (and thus less underfitting) than the constant model. This also does not explain why the MSE for the LR model is higher

Question 11 – Snog–Q7 – 164399.1.4

Snog is still a little confused about what happened in the model of Question 10 but he has heard that when you have a lot of predictor variables, you might want to consider using regularization, such as the LASSO or Ridge regression, still using the MSE loss. Which of the following considerations might be a reason to prefer LASSO? Select all that apply.

A LASSO is more likely than ridge to give an answer with most coefficients set to 0, which might make the final prediction equation a little easier to understand.

B Because LASSO uses L1 instead of L2, the LASSO solution will not be affected much by outlier data points, even if they are extreme outliers.

C LASSO regression is better because it optimizes the absolute loss of the data points (compared to squared loss of Ridge).

For A: we discussed it in class (also in the review session). For B: The loss function is still squared, therefore outliers still affect the result. For C: Again, the loss function is squared, the L1 norm is only used for the coefficient values

Question 12 – Snog–Q8 – 164402.1.6

Snog decides to use Ridge regression instead, and wants to use 5-fold cross-validation to select the hyperparameter. He takes his folds to be continuous blocks, using the rows 0 to 9 for the first fold, 10 to 19 for the second, and so on. He notices that his data set is sorted by `age`, the first column, so the villagers in the first fold are much younger than the ones in the last fold. Which of the following ideas would allow him to resolve this problem and keep it from affecting the results? Select all that apply.

A Don't use `age` as a predictor variable in the regression.

B Re-sort the data, in increasing order of the variable `restored`, and then use continuous blocks for his folds.

C Shuffle the rows of the data frame randomly, and then use continuous blocks for his folds.

D There is nothing he can do to fix it, he just needs to start over and collect a new data set.

check also one of the class notebookes, where the CV option of ski-kit learn is assessed

Question 13 – Snog–Q9 – 164406.1.6

Assume Snog has fixed the issue from the previous part to his satisfaction and he now has 5 folds that he is happy with. He uses 5-fold cross-validation to decide on the hyperparameter Q to use as the maximum squared L2 norm for the parameters. That is, any particular choice of Q means minimizing MSE in respect to the parameters θ for all `d` variables,

such that $\sum_{j=1}^{d} \theta_j^2 \leq Q$

Snog does 5-fold cross-validation to evaluate a grid of 100 different values for Q ranging all the way from tiny values to enormous values of Q. He is surprised to find that, as Q gets larger and larger, the cross-validation MSE stops changing, so the top 13 values of Q all give exactly the same cross-validation estimate of MSE. What do you think could best explain this?

A Above a certain value of Q, he is overfitting so much that he is making an exactly correct prediction for every single training point, and once he reaches this point he can't overfit any more.

B Above a certain value of Q, his answer is very close to the constant model, so the predictions are virtually identical.

C Above a certain value of Q, his predictions will coincide exactly with the predictions for OLS linear regression.

D Above a certain value of Q, his optimization algorithm is not able to converge to the optimal parameters.

E Above a certain value of Q, the optimal parameters are so large that numerical overflow will make the calculation fail.

Recall from class, that Q is the "budget" or the "radius of the ball" in the example we used, that we restrict our coefficients in. It is inversely related to the regulatization parameter (λ) we use in the usual formulation.
As Q gets larger, the ball gets bigger and the restictions gets looser. At one point the ball will include the OLS solution and we will have no regularization at all

Beaver Bedtimes

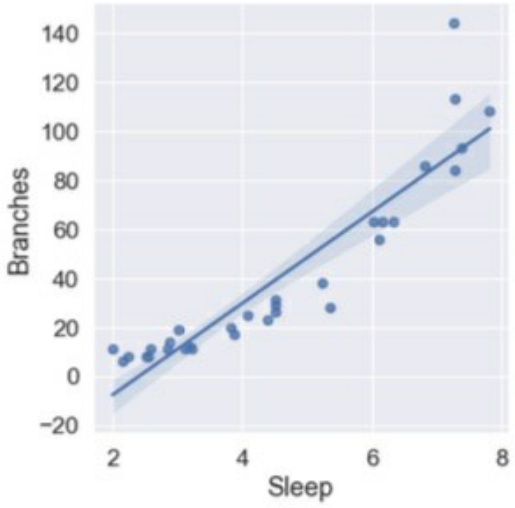
Question order: Fixed

Brenda Beaver is building a dam, and she has been waking up earlier and earlier in the morning trying to build it faster. However, her friend Olivia Otter thinks Brenda is overworking herself, and that her dam production is suffering as a result.

To convince Brenda to get more sleep, Olivia created a DataFrame called `production`, shown below, with one row for each of the 31 days in the last month. For each day, Olivia recorded the amount of sleep Brenda got (in hours) and how many branches Brenda was able to add to her dam on that day. The first five rows of the DataFrame look like the left figure below, while on the right figure we can see the linear regression line fit based on `Sleep` being the predictor variable (x) and `Branches` being the dependent variable (y).

	Sleep	Branches
0	4.5	29
1	6.3	63
2	2.0	11
3	3.8	20
4	2.9	14

production



For the questions in this block you may always refer back to the information provided here.

Question 14 – beaver–Q1 – 164409.1.1

Olivia wants to know how the linear model performs on the dataset. Looking at the regression line (as presented in the block introduction), which of the following are correct? Select all that apply.

A sleep and branches seem to be linearly correlated, therefore linear regression is the right tool to model their association

B for small values of sleep (less than 3.8), our linear model tends to underpredict branches

C for small values of sleep (less than 3.8), our linear model tends to overpredict branches

D for values of sleep (between 4-6), our linear model tends to underpredict branches

E for values of sleep (between 4-6), our linear model tends to overpredict branches

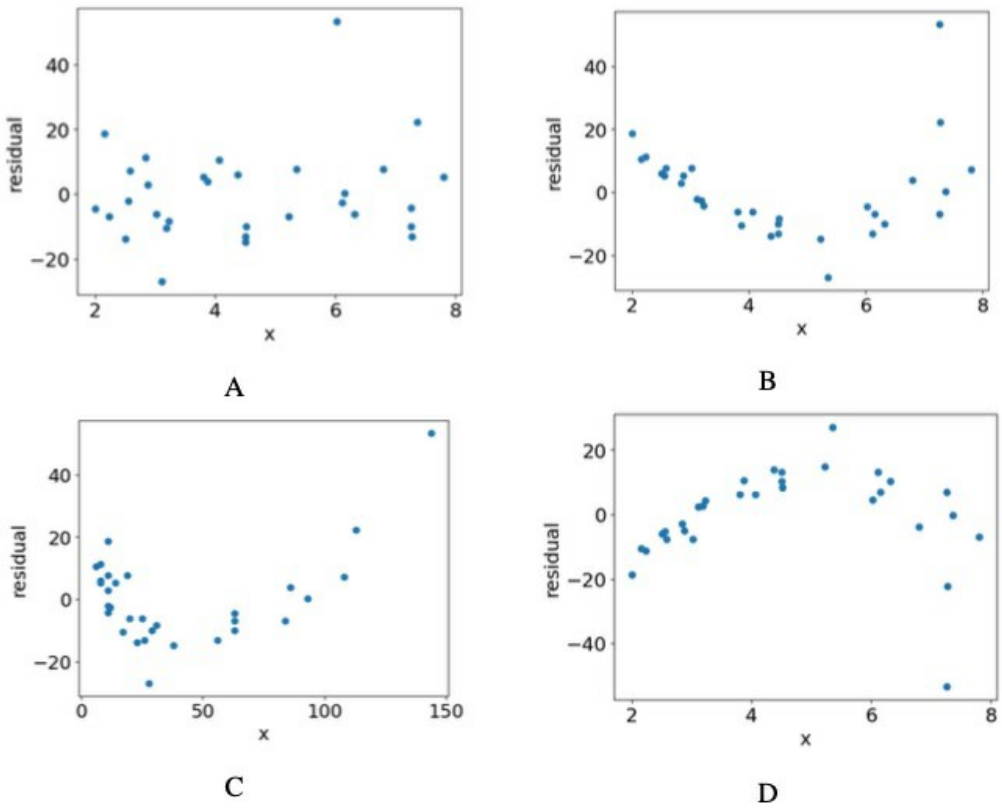
For A: Despite the outlier, the relationship is linear and strong

For the rest:
if the points are above the regression line, then the residual (error) is negative, i.e. the prediction (which is on the line) is lower than the true value, i.e. the model under-predicts

similarly, if the points are below the line, the residual is positive, i.e. the prediction is higher than the true value, i.e. the model over-predicts

Question 15 – Beaver–Q2 – 164410.1.0

Olivia makes some more visualizations to evaluate the linear model. She makes a plot of the residuals $y - \hat{y}$ (y=branches) on the vertical axis against the predictor x(=sleep) on the horizontal axis. Which of the plots below (A-D) best matches her plot?



A A

B B

C C

D D

based on the plot and the previous answer, should be obvious

Question 16 – Beaver–Q3 – 164411.1.1

Olivia is thinking about using a transformation to make her model fit better. Which one transformation, from the following options, do you think is most likely to improve the fit?

A Change the response variable to `log(Branches)`

B Change the predictor variable to `log(Sleep)`

C Change the response variable to $\texttt{Branches}^2$

D Change the predictor variable to `60 × Sleep` (so sleep is measured in minutes)

You can either think as a data scientist or refer to the bottom-right part of the Tukey-Mosteller Bulge Diagram. We can apply any of the transformations there (sqrt(y), log(y), x^2, x^3), but the only one that fits is the log(y), aka A

Question 17 − Beaver−Q4 − 164413.3.3

Regardless of what you answered before, assume that Olivia comes to believe that the logarithms of x (`Sleep`) and y (`Branches`) are related, so she should actually use the prediction function

$$f(x) = \theta_0 + \theta_1 \log(x)$$

where $f(x)$ is the predicted value for $\log(y)$ (both logarithms are base e). What is the prediction for y as a function of x, in terms of θ_0 and θ_1?

A $\hat{y} = e^{\theta_0} + x^{\theta_1}$

B $\hat{y} = e^{\theta_0} \cdot x^{\theta_1}$

C $\hat{y} = e^{\theta_0} \cdot \theta_1 x$

D $\hat{y} = e^{\theta_0} \cdot (\theta_1)^x$

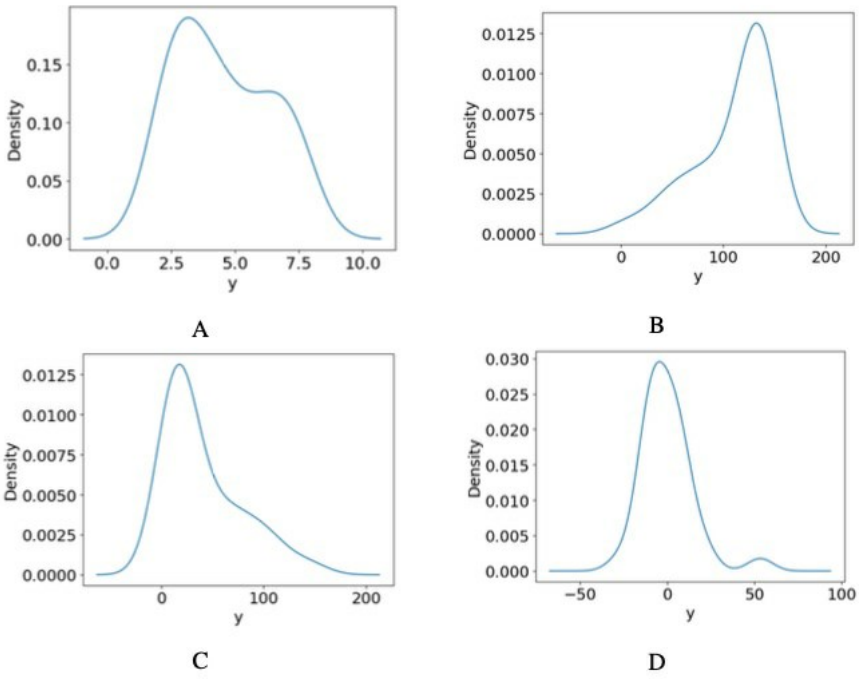
E $\hat{y} = e^{\theta_0} + (\theta_1)^x$

f(x) = log(y), so by raising e to the power of both sides (and our basic algebra knowledge) we have:

$$e^{\log(y)} = e^{\theta_0 + \theta_1 \log(x)}$$
$$y = e^{\theta_0} \cdot e^{\theta_1 \log(x)}$$
$$y = e^{\theta_0} \cdot e^{\log(x^{\theta_1})}$$
$$y = e^{\theta_0} \cdot x^{\theta_1}.$$

Question 18 − Beaver−Q5 − 164414.1.2

Olivia wants to take a closer look at the distribution of y, the branches column. Which of the Kernel Density Plots (A-D) best matches the distribution of y?



A A

B B

C C

D D

Modeling preferences

Question order: Fixed

Dominic has some survey data of his friends and he wants to better understand their interests, preferences and habits. In this case, Dominic wants to predict whether students live in the United States based on their favourite animated cartoon character (`pref_ cartoon_character`) and number of movies watched per year (`n_movies_per_yr`). The following training data is sampled from the dataset.

	pref_cartoon_character	n_movies_per_yr	lives_in_united_states
0	Mickey Mouse	6	1
1	Homer Simpson	2	0
2	Fred Flintstone	4	1
3	Mickey Mouse	2	0
4	Franklin	9	0

For the questions in this block you may always refer back to the information provided here.

Question 19 – mickey1 – 164432.1.2

Using the Logistic Regression model in scikit-learn with no regularization, we obtain 100% accuracy on the training set by training a model on one-hot encoded features specifying the cartoon character and the number of movies watched per year. What does this suggest about the relation between the features and the dependent variable?

Grading instruction

Criterion 1 (Number of points: 2)

The features are linearly separable based on the unique values of the response variable.

Note after grading: Mentioning strong association gave partial credit.

Question 20 – mickey3 – 164434.1.5

Instead of using one-hot encoding, suppose we encode the `pref_cartoon_character` with a 1 if the character originated from the United States and 0 if it did not. We train a logistic regression model on this feature along with the number of movies per year and an intercept.

Suppose the optimal parameters obtained for our optimal logistic regression model trained on these three features (including the intercept, listed as the first element of the vector) are $\theta = [2, 1, 0.25]$. Unfortunately, we find that an outlier has strongly affected our cross-entropy loss and our model parameters!

Calculate the cross-entropy loss (using log base e) on the outlier data point $x_{outlier} = [1, 1, 4]$ if the corresponding $y_{outlier} = 0$.

It's fine if you just present formulas (in any format) and not a final result, but please clearly show your work.

Grading instruction

correct equations (Number of points: 2)

details (Number of points: 1)

Note after grading: The loss is always positive, probabilities cannot be more than 1, logarithms can be computed for positive numbers

$p(Y=1|x) = \sigma(\theta x) = \sigma(2x2 + 1x1 + 0.25x4) = \sigma(4) = 0.982$

Cross entropy loss will be simplified (y=0) to:

$-\log(P(Y=0|x) = -\ln(1-0.982) = 4.017$

Question 21 – mickey4 – 164435.1.1

After some tinkering, Dominic is able to train a logistic regression model. Suppose our test set consists of the following 6 points, with the corresponding predictions shown below. Calculate the recall of the model on the test set.

y	$\hat{y}$
1	1
1	0
1	1
0	1
0	0
0	1

We find TP=2, FP=2, TN=1 and FN=1

Recall = 2 /(2+1) = 2/3

Grading instruction

Criterion 1 (Number of points: 1)

Question 22 – mickey6 – 164452.2.7

Next Dominic is trying to use an ensemble of decision trees for a classification problem with bagging. He notices that the decision trees look too similar. He would like to build a more diverse set of learners. For this reason he is thinking of the following options:

A. Increase the size of each random subsample
B. Decrease the size of each random subsample
C. Sample without replacement

Which one would you suggest and why? Justify your answer shortly.

C Sampling without replacement increases the diversity of the subsamples

B can also be correct under circumstances

Having too big samples increases chances of similarities between them

Grading instruction

ChoiceA (Number of points: 1)

Justification (Number of points: 1)

Note after grading: Formally, C would cancel the definition of bagging, however if your argumentation was correct, it was taken as correct

PCA and Timeseries potpouri

Question order: Fixed

Questions in this block refer to PCA and timeseries. They can be answered independently from each other.

Question 23 – PCA1 – 164471.1.4

Consider the matrix X below.

$$X = \begin{bmatrix} 0 & 2 & -1 \\ 0 & 2 & -2 \\ 1 & 1 & -3 \\ 1 & 1 & -4 \\ 2 & 0 & -5 \end{bmatrix}$$

Suppose we decompose X using PCA into $X = U\Sigma V^T$. Let $r \times c$ be the dimensions of V^T .

What is r and c?

- **A** r=2, c=2
- **B** r=2, c=3
- **C** r=3, c=3
- **D** r=5, c=2
- **E** r=5, c=3
- **F** r=5, c=5

U is to be 5x3
Σ is to be 3x3 (diagonal)
V is to be 3x3

Question 24 – PCA2 – 164474.1.4

Let P be the principal component matrix of X . That is, $P = U\Sigma$. Suppose we now decompose the principal component matrix P into its principal components, that is $P = U_P \Sigma_P V_P^T$. What are the dimensions and values of V_P^T ?

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

One can arrive at this answer by equating UΣ = Up Σp VpT and noticing that this impliesVpT =I. While correct, is not very interesting or useful, except that you remember algebra

Instead, we can see that VT is a rotation matrix that rotates our data X so that it is axis-aligned. In other words, P=UΣ is the data rotated such that the greatest variance occurs along the x axis. If we perform PCA again, we're basically trying to rotate P so that it is axis-aligned. However, since it is already axis-aligned, so the rotation will do nothing. That is why is the identify matrix, which does not transformation on the data. The dimensions are 3x3 for the same reasons as in the previous question.

Grading instruction

some justification (Number of points: 1)

detailed one (Number of points: 1)

Question 25 – PCA3 – 164476.1.2

When we created 2D PCA scatter plots in class, we were usually plotting the first 2 columns of which matrix ($U,\ \Sigma,\ V^T,\ U\Sigma,\ \Sigma V^T\ or\ U\Sigma\Sigma V^T$)? Justify your answer shortly (showing code is okay but not necessary). Keep your answer short.

Grading instruction

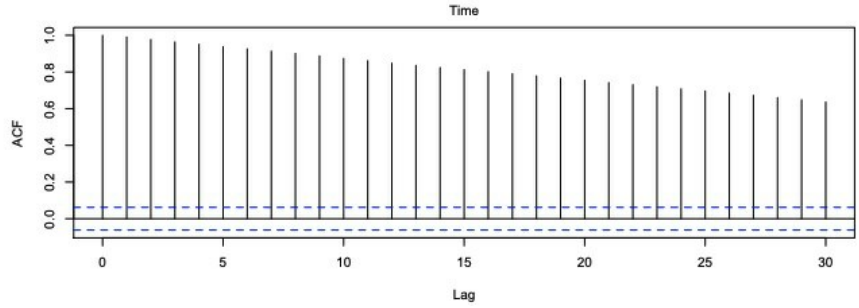
Short answer (Number of points: 1)

Should be easy to answer that we plot the 2 columns of UΣ (which holds the principal components)

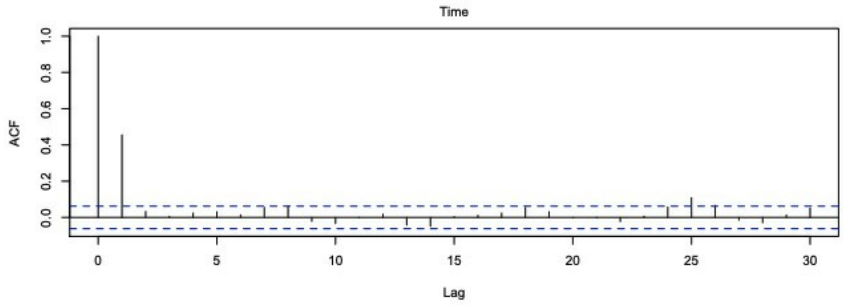
Question 26 − time − 164486.2.1

After getting the Autocorrelation Plots (ACF) for two different timeseries, you get the following two figures. What do these suggest for the timeseries and their stationarity? If you identify issues with stationarity, mention briefly how you can correct them. Keep your answers short.

A:



B:



Grading instruction

Figure 1 (Number of points: 1)

Figure 2 (Number of points: 1)

Explanations (Number of points: 1)

1) Strong linear trend, differencing might be a first good approach
def. not stationary

2) More difficult to say. However, given that only 2 lags are outside the bands, we could say that the timeseries is staionary

Note after grading: If you argued differently (e.g. some said that we probably have a MA(1) model) and your answer still makes sense full points were given

Classification

Question order: Fixed

Questions in this block refer to different issues around classification and they can be answered independently. The last one is more challenging, so I suggest to leave it last.

Question 27 – Classification1 – 164451.1.6

Let classifier A be a random forest and classifier B be an ensemble of decision trees with bagging, identical to the random forest except that (as we know from lectureS) we do not limit the splits in each treenode to a subset of features like random forests do; that means that at every treenode, the very best split among all d features are chosen.

Do you agree/disagree with the following propositions and why? Justify your answers shortly.

Conclusion 1: Classifier B will tend to have higher variance than Classifier A
Conclusion 2: Classifier B will tend to have higher training accuracy and Classifier A

Grading instruction

Conclusion 1 (Number of points: 2) Correct: Randomly limiting the splits, as RF does, tends to decrease the variance

Conclusion 2 (Number of points: 2) Correct: Again, limiting the splits decreases bias, therefore we expect better training accuracy

Question 28 – Classification2 – 164453.1.1

Your friend argues that AdaBoost algorithm can benefit from a learner (base model) that has only 5% training accuracy. Do you agree or not? Justify your answer shortly.

Grading instruction

Yes (Number of points: 1)

justification (Number of points: 2)

True! Think the way AdaBoost works. If a learner has 5% training accuracy, then AdaBoost will treat this by flipping its output so the accuracy is 95% - Remember AdaBoost is iterative and learns from the errors. Therefore, it can benefit very much.

Remember, in boosting the learners are weak since we combine them (with weighting)

Question 29 – Classification3 – 164454.1.1

In the following statements, "bias" and "variance" refer to the bias-variance trade-off when it comes to models. Which ones are true or false? Justify your answers shortly.

A. A model trained with n training points is likely to have lower variance than a model trained with $2n$ training points
B. If my model is underfitting, it is more likely to have high bias than high variance
C. Increasing the number of parameters in a model usually improves the test set accuracy
D. Adding L2-regularization usually reduces variance in linear regression

Grading instruction

A (Number of points: 1)

B (Number of points: 1)

C (Number of points: 1)

D (Number of points: 1)

A: False: Increasing the # of samples, generally reduces variance
B: True: You underfit if you have insufficient model capacity to even perform well on the training set
C: False: Increasing the number of parameters can lead to overfitting (therefore reduce the test set accuracy)
D: True: Regularization effectively reduces model complexity

Note after grading: The exercise covers the "general case". However, It was quite nice to see some of you argue for special cases, e.g. for C: many of you said that "it depends". Again, if your argument was correct, no points were reduced

Question 30 – Classification4 – 164456.1.3

This is a tough question that requires some heavy thinking. Leave it for last. You are running logistic regression to classify two-dimensional sample points X_1 and X_2 into 2 classes $y = \{0, 1\}$. Unfortunately, regular logistic regression isn't fitting the data very well. To remedy this, you try appending an extra feature, $X_1^2 + X_2^2$. After you run logistic regression with the new feature, the decision boundary could be:

A. a line
B. a circle
C. an ellipse
D. an S-shaped logistic curve

Select the correct one(s) and justify your answer shortly.

The decision boundary would be determined by:

w1x1 + x2x2 + w3 (x1^2+x2^2) + w0= 0

This equation can express arbitrary circles and lines. It cannot express ellipses because there is only one quadratic feature and it causes x1^2 and x2^2 to always have the same coefficient. For sure we can't have S-shaped curves.

Grading instruction

some effort done (Number of points: 2)

full effort (Number of points: 3)

Bonus Question

Note after grading: I knew this is not easy, however many of you argued in the right direction and some even got full points here! Congrats!
Partial credit was given to answers that argued based on the form of the model, the fact that LR is linear etc.

Question order: Fixed
This block contains the bonus question

Question 31 – qb–2024 – 168030.1.1

According to the course's Discord, how many ways are there to make yourself cry and let out your feelings?

A 0
B 5+
C 10+
D 15+
E 20+