Block introduction

For the following questions select the right answer (they are either multiple-choice or multiple-answer).

Question 1 − Fill in (multiple) − Question-ID: 109200 (4 points)

The Municipality of Maastricht wants to hear from its homeowners on issues related to housing.

(For the purposes of this question, homeowners are individuals who own their home, instead of leasing or renting from someone else.)

One method of surveying would be to have city workers come to UM campus and ask passing by students and faculty members for their thoughts. Suppose for now that the question “Are you a homeowner?” is not asked.

What type of sample is this? convenience sample

Many students and faculty members aren’t homeowners, but will be surveyed anyways.

What form of bias or error is this? selection bias

Maastricht municipality has a list of all the homeowners’ email addresses. Instead of the previous surveying technique, now suppose they take the list of all homeowners’ email addresses, shuffle it, and send a survey to every other email address. That is, from the shuffled list, they email the first, third, fifth, seventh, and so on.

In this new sampling technique, the sampling frame is equal to the population of interest.

In this new sampling technique, some homeowners may see the survey and choose not to respond. The only form of bias or error in this new surveying technique is non-response bias. True of False? FALSE

Question 2 − Multiple choice − Question-ID: 109171 (2 points)

Tom and Jerry built a model on their data with two regularization hyperparameters a and b. They have 4 good candidate values for a and 3 possible values for b, and they are wondering which (a,b) pair will be the best choice. If they were to perform five-fold cross-validation, how many validation errors would they need to calculate?

A 3

B 4

C 5

D 12

E 35

[✓ **F**] 60 3*4*5 errors

Question 3 − Multiple response − Question-ID: 109176 (3 points)

Suppose you are working with a partner to train a model with one hyperparameter λ. Together, you and your partner run 5-fold cross validation and compute mean squared errors for each fold and value of λ from a set of 4 candidate values for λ. However, your partner forgets to send you the results for the last two folds! The table below contains the mean squared errors for the first three of five total folds.

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	64.2	60.1	77.7	79.2	70.3
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0

Your partner uses the full table containing data for all five folds to create a final model to use on test data. Given the information above, what can you conclude about the final model? Select all that apply.

A Our final model should use $\lambda = 0.4$

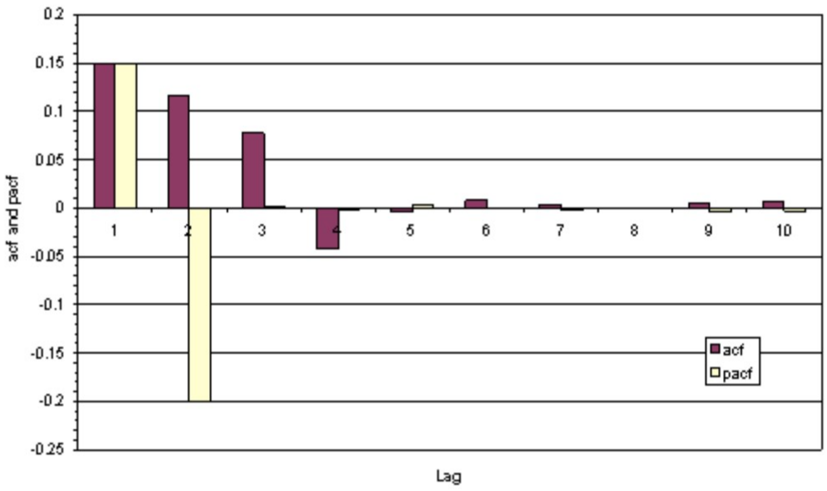
B Our final model should be trained on fold 2, since it achieves the highest row average.

[✓ C] We would need more information

D Our final model should be trained on fold 1, since it achieves the lowest row average.

Question 4 − Multiple choice − Question-ID: 109195 (1 point)

Consider the picture below and suggest the model from the following list that best characterises the temporal process:



[✓ A] ARMA(2,1)

B AR(1)

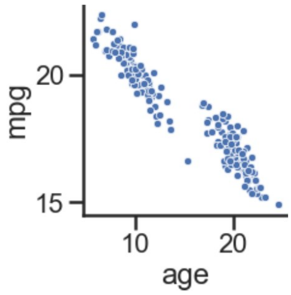
C MA(2)

D AR(2)

Block Plots + PCA

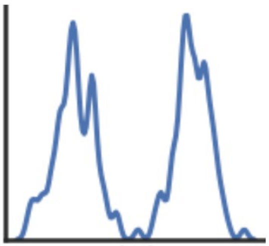
In this block there is a single question about plots and PCA. Study carefully the plots before answering the questions

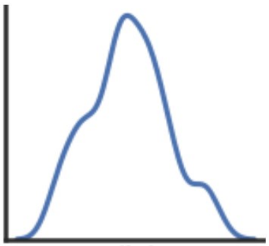
Question 5 − Fill in (multiple) − Question-ID: 109165 (8 points)

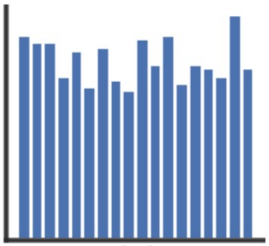


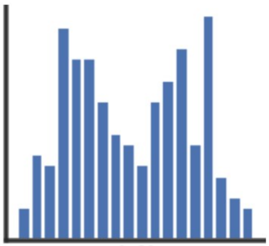
During a data analysis on car attributes, Tom created several plots. However, he has lost the axis labels for all of his plots except for the scatter plot shown on the left.

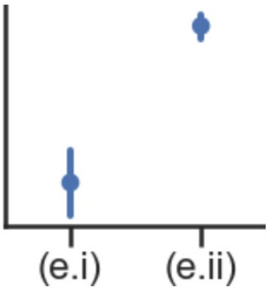
Assume that: The KDE plots use the same bandwidth, the histograms use the same number of bins, and point plots show the means of two columns and 95% confidence intervals. The axis limits for each plot were automatically chosen to display all plotted marks.

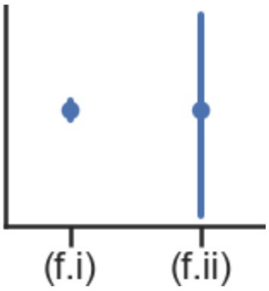

(a)

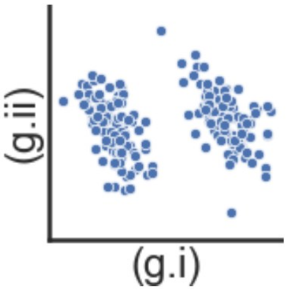

(b)

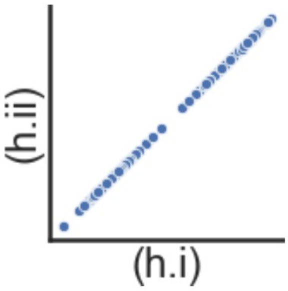

(c)


(d)









Determine whether the plots below (marked from (a) till (i)) were generated from the same data. If so, mark the axis label that makes each plot consistent with the data in the scatter plot.

Plot (a): age
Plot (b): neither
Plot (c): neither
Plot (d): mpg
Plot (e.i): age
Plot (e.ii): mpg
Plot (f.i): neither
Plot (f.ii): neither
Plot (g.i): neither
Plot (g.ii): neither
Plot (h.i): age or mpg
Plot (h.ii): age or mpg

(a-d): Notice that both age and mpg are bimodal, but the data have a gap in age and not mpg.
(e-g): The average age is lower than the average mpg.
(h): Either age for both axes or mpg for both axes is correct. The same variable plotted on both x and y axes will give a line with slope 1.

After conducting PCA, Tom projected each point onto the two principal component axes. He stored the projections onto the first and second principal components in the columns pc1 and pc2, respectively. As in the previous part, fill in each of the missing axis labels using either pc1 or pc2 if the plots were generated using the points projected onto the first or second principal component, or select neither.

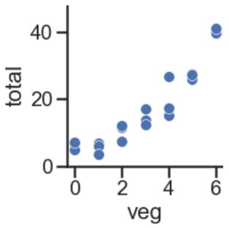
Plot (a): pc1
Plot (b): pc2
Plot (c): neither
Plot (d): neither
Plot (e.i): neither
Plot (e.ii): neither
Plot (f.i): pc2
Plot (f.ii): pc1
Plot (g.i): pc1
Plot (g.ii): pc2
Plot (h.i): pc1 or pc2
Plot (h.ii): pc1 or pc2

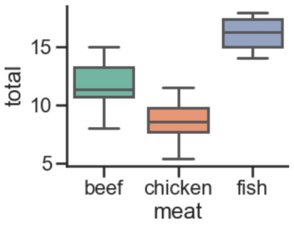
The first PC points from the upper left of the scatter plot to the lower right.
The second PC is perpendicular to the first and points from lower left to upper right.
(a): After projecting on the first PC, there is a gap between the two clusters of data.
(b): The points generally lie along the first PC with large variations occurring less frequently than small variations.
(c-d): Neither of these could have been generated from the projected points.
(e-f): Remember that we subtract the average value from each column before conducting SVD. This means that the points are always centered at 0 after projection, ruling out choice (e). The first PC captures more variance than the second PC, so (f.ii) is pc1 and (f.i) is pc2.
(g): pc1 is the x-axis because the points are divided into two clusters along the first PC.
(h): As in the previous part, either pc1 for both axes or pc2 for both axes are correct

Block Mananas

Every week, Manana goes to her local grocery store and buys a varying amount of vegetables but always buys exactly one pound of meat. We use a linear regression model to predict her total grocery bill. We've collected a dataset containing the pounds of vegetables bought, the type of meat bought (either beef, chicken, or fish), and the total bill. Below we display the first few rows of the dataset and two plots generated using the entire dataset.

veg	meat	total
1	beef	13
3	fish	19
2	beef	16
0	chicken	9





Question 6 – Fill in (multiple) – Question-ID: 109183 (6 points)

Suppose we fit the following linear regression models to predict `total`. Based on the data and visualizations shown in the block introduction, determine whether the fitted model weights are positive (+), negative (-), or exactly 0. The notation `meat=beef` refers to the one-hot encoded meat column with value 1 if the original value in the meat column was beef and 0 otherwise. In the next question shortly justify your answer.

- $f(\mathbf{x}) = \theta_0$
  Weight of θ_0 +

(a)
θ0 is the mean of total, which is positive

- $f(\mathbf{x}) = \theta_0 + \theta_1 \cdot veg^2$
  Weight of θ_0 +
  Weight of θ_1 +

θ0 is the total when veg^2=0 is positive
θ1 is positive as total increases when veg (or its square) increases

- $f(\mathbf{x}) = \theta_0 + \theta_1 \cdot (meat = beef) + \theta_2 \cdot (meat = chicken)$
  Weight of θ_0 +
  Weight of θ_1 -
  Weight of θ_2 -

θ0 is the total when meat=fish which is positive
θ1 is the difference between total when meat=beef and total when meat=fish, which is negative
θ2 is the difference between total when meat=chicken and total when meat=fish, which is negative

- $f(\mathbf{x}) = \theta_0 + \theta_1 \cdot (meat = beef) + \theta_2 \cdot (meat = chicken) + \theta_3 \cdot (meat = fish)$
  Weight of θ_0 not enough info
  Weight of θ_1 not enough info
  Weight of θ_2 not enough info
  Weight of θ_3 not enough info

This model is an extension of the previous model, however θ3 is not needed. There are thus an infinite number of possible combinations of θ0, θ1, θ2, θ3. We cannot say for sure what they will be.

Extra credit from P (in the exam):
Taking θ0 to be the free variable
θ1 = [total when meat=beef] - θ0
θ2 = total when meat=chicken] - θ0
θ3 = total when meat=fish] - θ0.

Question 7 – Open-ended – Question-ID: 109184 (4 points)

Here you can justify your answer to the previous question.

Question 8 – Fill in (multiple) – Question-ID: 109187 (3 points)

Suppose we fit the model: $f(\mathbf{x}) = \theta_0 + \theta_1 \cdot veg + \theta_2 \cdot (meat = beef) + \theta_3 \cdot (meat = fish)$. After fitting, we find that $\hat{\theta} = [-3, 5, 8, 12]$. Calculate the following

The prediction of this model on the first point in our dataset. 10 (=-3+5x1+8x1+12x0)

The loss of this model on the second point in our dataset using squared error loss. 25 prediction-real value = 24-19=5

The loss of this model on the third point in our dataset using absolute loss with L1 and $\lambda = 1$
26 |prediction - real value| = 1 and L1 loss = λ * (|θ1|+|θ2|+|θ3|)=1*25=25

Question 9 – Open-ended – Question-ID: 109189 (6 points)

Determine how each change below affects model bias and variance compared to the model described at the previous question, aka: $f(\mathbf{x}) = \theta_0 + \theta_1 \cdot veg + \theta_2 \cdot (meat = beef) + \theta_3 \cdot (meat = fish)$

1) Add degree 3 polynomial features decrease bias, increase variance
2) Add a feature of random numbers between 0 and 1 increase variance
3) Collect 100 more sample points decrease variance
4) Remove the `veg` column increase bias, decrease variance

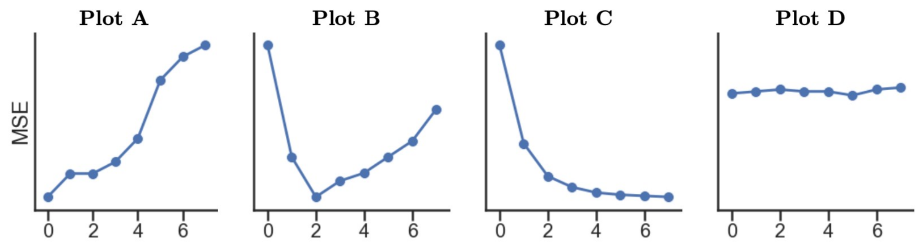
For each option (1-4) justify shortly what will happen to the bias **and** variance of the model.

---

1) Increasing model complexity, so bias decreases and variance increases
2) Assuming the coefficient for the random feature is non-zero, variance will increase and bias will stay the same
3) Assuming that the sample points come from the same stationary distribution as the original set, variance will decrease.
4) Decreasing model complexity, so variance will decrease, and bias will increase.

Question 10 − Fill in (multiple) − Question-ID: 109190 (2 points)

Suppose we predict total from `veg` using 8 models with different degree polynomial features (degrees 0 through 7). Which of the following plots display the training and validation errors of these models? Assume that we plot the degree of polynomial features on the x-axis, mean squared error loss on the y-axis, and the plots share y-axis limits.



Training Error: C

Validation Error: B

Block Boosting, Bagging, Random Forests

The following questions refer to boosting, bagging and random forests

Question 11 − Multiple response − Question-ID: 109191 (2 points)

Consider running a single iteration of AdaBoost on three sample points, starting with uniform weights on the sample points. All the ground truth labels and predictions are either +1 or −1. In the table below, some values have been omitted. Which of the following statements can we say with certainty? In the next question you will be asked to justify your answers shortly.

	True Label	Classifier Prediction	Initial Weight	Updated Weight
X_1	-1	-1	$1/3$	?
X_2	?	$+1$	$1/3$	$\sqrt{2}/3$
X_3	?	?	$1/3$	$\sqrt{2}/6$

A X_3's classifier prediction is -1

[✓ B] X_2 is misclassified

C X_3 is misclassified

[✓ D] X_1's updated weight is $\frac{\sqrt{2}}{6}$

- In the AdaBoost algorithm, all correctly classified points have their weights changed by the same factor. Since we observe two different updated weights, we know one of x2 or x3 is correctly classified, and the other is mis- classified.
- Since x1 is correctly classified, the error rate is err = 1/3. As the error rate is less than 1/2, the weights of correctly classified points will decrease and the weights of misclassified points will increase.
- Hence, X2 is misclassified and X3 is correctly classified.
- As X1 is correctly classified, it has the same updated weight as X3. But we can't tell what X3's classifier prediction is; only that it is correctly classified.
- As an aside, we can confirm the multipliers used for re-weighting of misclassified and correctly classified points (in that order):

sqrt (err/(1-err)) = sqrt ((2/3)/(1/3)) = sqrt(2)

Question 12 − Open-ended − Question-ID: 109192 (5 points)

In the answer box below, please justify shortly your answer to the previous question.

Question 13 − Open-ended − Question-ID: 109193 (2 points)

Why does bagging by itself (without random subset selection) tend **not** to improve the performance of decision trees as much as we might expect?

It is common that the same few features tend to dominate in all of the subsets, so almost all the trees will tend to have very similar early splits, and therefore all the trees will produce very similar estimates. The models are not decorrelated enough.

Question 14 − Multiple response − Question-ID: 109194 (2 points)

Why would we use a random forest instead of a decision tree? Select all that apply.

[✓ **A**] To reduce the variance of the model

B For lower training error

C For a model that is easier for a human to interpret

Block bonus

The last block contains the bonus question

Question 15 − Open-ended − Question-ID: 109643 (1 point)

Describe shortly the most hilarious experience you had in your studies at DKE.

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