Question 1: **The multiple choice (maximum 10 points, minimum 0 points)**

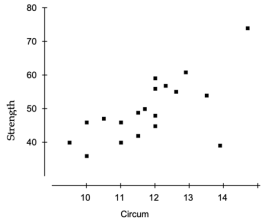
Each question may have any number of correct choices. Clearly mark (tick or circle) all choices you believe to be correct. You get 1 point for each correct choice, -0.5 for each incorrect choice. You can get a maximum of 10 points (note that does not necessarily mean that there are 10 correct choices) and a minimum of 0 points (i.e. there is no negative scoring).

a) Suppose that a Normal model describes fuel economy (miles per gallon) for automobiles and that a Tesla has a standardized score (z-score) of +2.2. This means that Teslas. . .

i) get 2.2 miles per gallon.
ii) get 2.2 times the gas mileage of the average car.
iii) get 2.2 mpg more than the average car
iv) have a standard deviation of 2.2 mpg.
v) achieve fuel economy that is 2.2 standard deviations better than the average car.

NB: The student that pointed out that Tesla does not consume gas, got a full point as well.

b) Researchers investigating the association between the size and strength of muscles measured the forearm circumference (in inches (in)) of 20 teenage boys. Then they measured the strength of the boys' grips (in pounds (lbs)) and made the depicted scatterplot. Which of the following sentences are correct?



i) If the point in the lower right corner (at about 14 in and 38 lbs.) were removed, the correlation become weaker
ii) If the point in the lower right corner (at about 14 in and 38 lbs.) were removed, the correlation become stronger
iii) If the point in the upper right corner (at about 15in and 75 lbs.) were removed, the correlation would become weaker
iv) If the point in the upper right corner (at about 15in and 75 lbs.) were removed, the correlation would become stronger
v) One pound is 0.45 kilograms. If I changed the units of the grips from pounds to kilograms, then correlation would change by a factor of 0.45

c) Which of the following are true for the probability density function (PDF) and the cumulative distribution function (CDF)?

i) PDF cannot be less than 0
ii) PDF cannot be bigger than 1
iii) CDF cannot be less than 0
iv) CDF cannot be bigger than 1

d) We are about to test a hypothesis using data from a well-designed study. Which of the following are true?

i) A small *P*-value would be strong evidence against the null hypothesis
ii) We can set a higher standard of proof by choosing a significance threshold of 10% instead of 5%
iii) If we reduce the significance level, we reduce the power of the test.

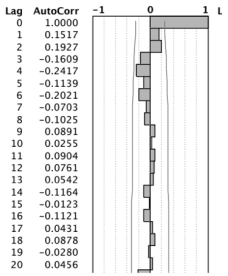
e) Suppose your model is demonstrating high variance across different training sets. Which of the following is NOT a valid way to try and reduce the variance?

i) Increase the amount of training data in each training set
ii) Improve the optimization algorithm being used for minimizing the error
iii) Decrease the model complexity
iv) Reduce the noise in the training data

f) Which of the following sentences are true?

i) Random scatter in the residuals indicates a model with high predictive power.
ii) If two variables are very strongly associated, then the correlation between them will be near +1.0 or -1.0.
iii) The higher the correlation between two variables the more likely the association is based in cause and effect.
iv) Plotting the residuals can help identifying a pattern in the data that was hard to see in the original scatterplot

g) The figure at the right shows the ACF for a timeseries of monthly data. The bars show the autocorrelation values at each lag (1 through 20) and the gray line around the bars shows the significance band. If we are concerned about autocorrelation with lags up to one year, we should…:



i) Identify a proper ARMA model to capture the dependence
ii) Difference the data prior to fitting the regression model to capture any seasonal patterns
iii) Take smoothing differences to phase out any dependence
iv) Recognize that there is not statistically significant autocorrelation

Question 2: Wasting time on Internet (10 points)

A group of DKE students decides to see if there is link between wasting time on the Internet and their GPA. They don't expect to find an extremely strong association, but they're hoping for at least a weak relationship. Here are the findings of their regression model (independent variable is time spent on Internet (measures in hours / week) and the dependent variable is their GPA).

Parameter	Estimate	Std. Err.
Intercept	9.07191	0.74405
Time	-0.0297	0.02616
R (correlation) = -0.372		

a) Describe the association in context. What is the value of R^2 here and what is the meaning of it?

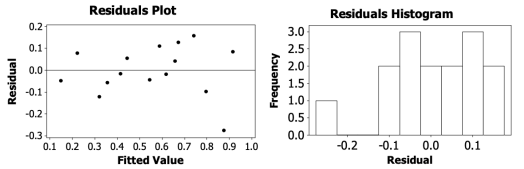
(2p) There is only a weak negative association. Correlation is -0.372, so R2 = 13.8%. That is the variance explained in the GPA due to time spent on internet is 13.8%. We expect that this model is not very predictive.

b) Interpret the slope and the intercept in context.

(2p) Intercept is 9.07. That is the value we get if hours_spent_on_internet=0, i.e. that is the grade that a student would get if no hours spent on internet

Slope is -0.0297, i.e. for every hour spent on internet, there is a grade reduction of -0.297

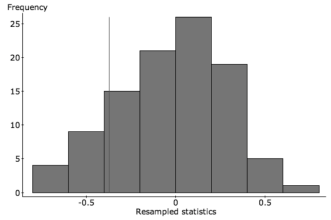
c) Below we see the plot of the residuals vs. fitted values (note that the fitted value is shown on a 0-1 scale here and not 0-10) and the histogram of the residuals. Do these plots suggest a good fit or not?



(3p) Residuals seem to be randomly distributed and there is no pattern learned according to the fitted value.
The residual histogram shows that there is one outlier, but that is not unexpected. Many of you argued that you expected a normal distribution, however you failed to notice that since we do not have many points that might not be the case (which means that uniform should not be that bad).

NB: For the online exam that the plots were not visible the points were shared with the other questions a), b), d)

d) One student is concerned that the relationship is so weak, there may not actually be any relationship at all. To test this concern, the student decides to collect more data, therefore asks 10 more students about their GPA and how much time they waste on Internet every week. Then, the student does some random re-matching of the GPAs and the hours spent on Internet. After each random assignment, the correlation is (re)calculated. This process is repeated 100 times. Below we see a histogram of the 100 correlations. The correlation coefficient of –0.372 is indicated with a vertical line.



Do the results of this simulation confirm the suspicion that there may not be any relationship? Refer specifically to the graph in your explanation and explain the process that the student followed.

(3p) The experiment is the heart of hypothesis testing (randomization test). By applying all permutations possible, students wanted to show how usual/unusual the observed value is.

It appears there may be no relationship at all. The value -0.372 does not appear to be unusual. 15 out of 100 times the correlation was even closer to negative one. So the association we are observing could be due to random variation.

Question 3: Logistic Regression (10 points)

In this question we are going to look at different methods for selecting features for a simple logistic regression model, namely: $P(y = 1|\boldsymbol{x}) = \text{logistic}(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ as we saw in class.

The available training examples are very simple, involving only binary valued inputs, namely:

number of copies	x_1	x_2	y
5	1	1	0
10	0	1	1
10	1	0	1
10	0	0	0

So, for example, there are 5 copies of the data point with (x_1=1, x_2=1), all labeled y=0. If you check the examples carefully, you will see that, practically, the output variable y is 1 when x_1=x_2 and 0 otherwise.

a) Would logistic regression be a good algorithm for the classification problem above? Explain shortly your answer.

(2.5p.) Simple answer: Logistic Regression is a linear model so it won't work well in this specific problem. If you don't see why the problem is non-linear, you can always draw the points (make a picture)

Suppose we do a "greedy" feature selection as follows: we start with no features (we train only with β_0) and then we add one feature at a time, provided that each addition *strictly increases* the training accuracy. Here we define training accuracy as the percentage of training data that are going to be classified correctly. We use no other spotting criterion.

b) Which feature would greedy selection add first, x_1 or x_2 or neither? Explain shortly your answer.

(2.5p) A constant model (meaning that there is only β_0) would classify correctly 20 out of 35 data points. Note that onstant model means that it always decides one class or the other.

There is no improvement for either x1 or x2.

c) What is the maximum accuracy on the training examples that we could achieve by including both x_1 and x_2 in the logistic regression model? Explain your answer.

(2.5p) If we include both x1 and x2, the maximum accuracy will be 30/35 = 85.7%

In that case, we misclassify the 5 cases of x1=0 and x2=0.

d) Suppose we define another possible feature to include, a transformation based on x_1 and/or x_2. Which of the following features, if any, would permit us to correctly classify all the training examples when used in combination with x_1 and x_2 in the logistic regression model: x_1-x_2, x_1x_2, x_2^2 (specify all that apply)?

(2.5p) The question asks to include another feature in combination with x1 and x2. So we keep both x1 and x2 in the model (I was a bit lenient with this).
The key point here is that we need some non-linearity to solve the problem.

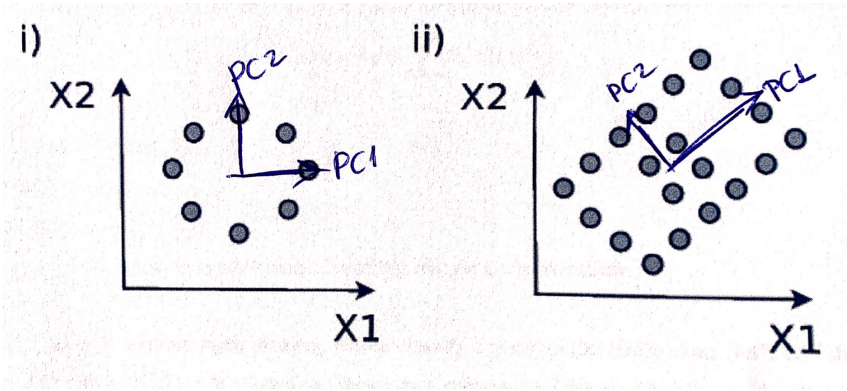
X2^2 won't work since it's a just a copy of x2!

x1-x2 still gives 0 on interaction for both x1 and x2.

x1x2 makes sure that samples misclassified by only using x1 and x2 are now correctly classified, since they have a different value. Think about adding an interaction term.

Question 4: Let's PCA (10 points)

Consider the following data sets (i) and (ii) where we have two features X1 and X2.



a) What would be your expectation from applying PCA on these two datasets? What would be the approximates for the first two principal components? You can sketch your answer on the figure and explain shortly below.

(4p.)

See figure

b) In which of the two cases (i) or (ii) will the first principal component capture a larger proportion of the variance in the data? Explain your answer shortly.

(3p.)

in case ii: variance in 1st component is more.

in case i: variance is the same in both directions

c) In which of the cases (i) or (ii) will the first two principal components (combined) capture a larger proportion of the variance? Explain your answer shortly.

(3p.)

same in both cases: data is 2-dimensional so 2 components always capture 100%

Question 5: Let's boost things up (10 points)

In class we discussed boosting and how we use several weak learners h_t in order to solve a classification problem (with t=1, we use one learner, with t=2, we use two etc.). Here we formalize this by taking their weighted sum as follows:

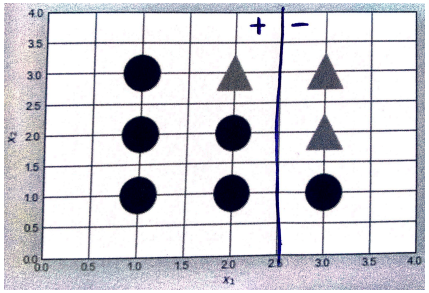
$$H(\boldsymbol{x}) = sgn\left(\sum_t a_t \cdot h_t(\boldsymbol{x})\right)$$

where:

- H is the final weighted sum
- $sgn(*)$ is the sign of *
- h_t is the weak learner
- a_t is the weight we give on each learner, assuming that on every iteration

We will use simple decision trees as weak learners, which classify a point in the circle class (let's call that class positive) or in the triangle class (let's call that class negative) based on a sequence of threshold splits on the data features (here we only have 2 features, x_1 and x_2). In all questions below, make sure that you clearly mark down which regions are circles vs. triangles (positive vs. negative class) and assume ties are broken arbitrarily.

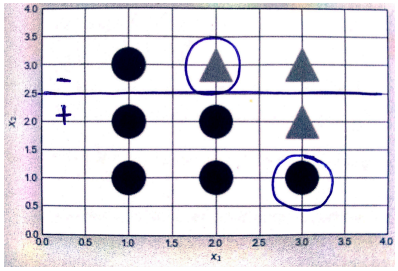
a) Assume that our weak learners are decision trees of depth 1 (i.e. stumps), which minimize the weighted training error. Using the dataset below, draw the decision boundary learned by the h_1 (i.e. the first stump). Explain shortly your answer.



(2p.) Decision trees of depth 1 means that essentially, we can only make one split, i.e. one line. Equivalent to a linear model.

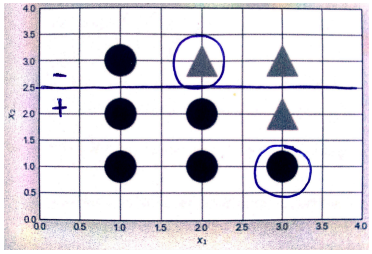
To minimize the error, you can either split at horizontal line X2=2.5 or vertical line X1=2.5. Here we pick the X1=2.5. The misclassified points are the triangle at the top middle and the circle at the bottom right.

b) We are now doing a second iteration (remember we still use stumps). Clearly mark the data point(s) with the highest weights on the second iteration and draw the decision boundaries by h_2. Explain shortly your answer.



(2p.) The misclassified points (from the first round) that get higher weight (in the second round) are circled. We are still using decision trees of depth 1, meaning that we can only add one more line, taking into account that the circled points have higher weight if misclassified now. Obviously, the 2 circle points need to be classified correctly now.

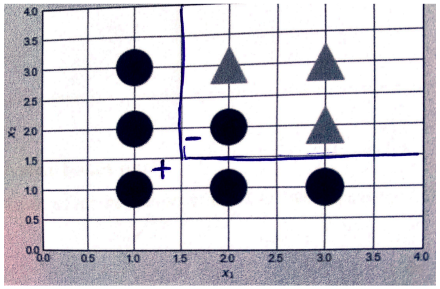
c) Now draw the decision boundary of $H = sgn(a_1h_1+a_2h_2)$. You do not need to compute the a_1 and a_2. Explain shortly your answer.



(2p.) A right answer: Although h_1 and h_2 misclassify the same number of points, the points which h_2 gets wrong have been downweighted. Thus, we have $err_1 > err_2$, so $\alpha_2 > \alpha_1$, and h_2 will dominate over h_1 whenever there is a conflict. In other words, H has the same boundary as h_2.

Other answers: Including a comment about a (no need to compute them, does not mean we ignore them), e.g. if you mention we assume equal a or any other assumption. Not commenting on this value and just interpolating gave partial credit.

d) Now assume that our weak learners are decision trees of maximum depth 2, which minimize the weighted training error. Draw the decision boundary learned by h_1. Explain shortly your answer.



(2p.) Different possible answers, as long as you get that miniziming the error, means only one point is misclassified. Also, having a decision tree of depth 2, means that we can have the space further divided (see also related notebook "extra stuff on trees").

e) How would *k-NN (*with k=3) perform on the above problem/dataset? Explain shortly your answer and write down any further assumptions you make (e.g. breaking ties, distances used).

(2p.) Assuming Euclidean distances, every point has 3 closest neighbors, except the circle in the middle that we have to decide how to break the ties (e.g. taking the majority class which would make it a circle). With that into account, we will misclassify the triangle on the top-middle and the triangle on the middle-right. So in total, we have 2 (max. 3) miclassifications, making k-NN (with k=3) perform similarly to the stump.

Bonus Question (contributes to your bonus part)

Make a prediction for the number of new COVID cases in the Netherlands for today 6718

(Any answer in the range 6300-7100 gets the bonus)