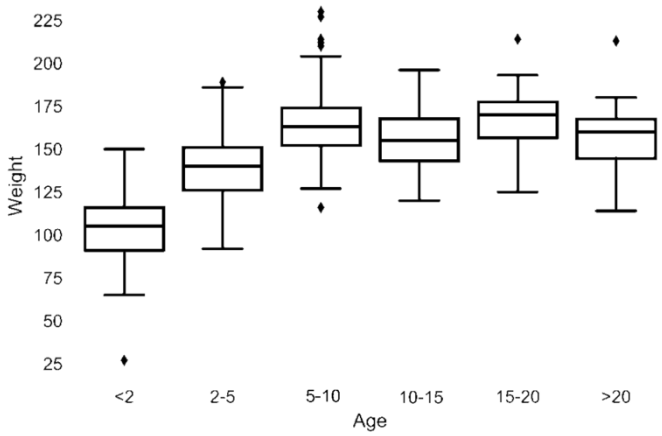
Question 1: **The multiple choice (maximum 10 points, minimum 0 points)**

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).

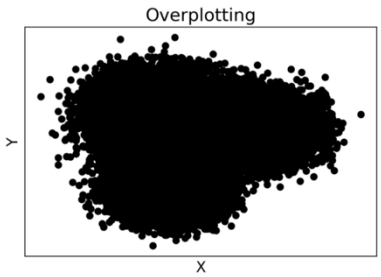
a) When developing a model for a donkey's weight, we considered the following box plots of weight by age category. This plot suggests:

i) Age is not needed in the model
ii) Some of the age categories can be combined
iii) Age could be treated as a numeric variable
iv) All of the above
v) None of the above



b) Which of the following changes will *most* effectively improve the following plot to communicate the relationship between the two variables *x* and *y*?

i) Change the dot size and/or transparency
ii) Remove the outliers
iii) Display a histogram
iv) Change the scale of *x* and *y*



c) A *P*-value indicates:

i) the probability that the null hypothesis is true
ii) the probability that the alternative hypothesis is true.
iii) the probability the null is true given the observed statistic.
iv) the probability of the observed statistic given that the null hypothesis is true.
v) the probability of the observed statistic given that the alternative hypothesis is true.

d) A poll of 120 Texas residents found that 30 had visited the Museum of the Earth, and that 80 had been to Home Depot. If it appeared that going to Home Depot and going to the Museum of the Earth were independent events, how many of those polled had been to both?

i) 10
ii) 15
iii) 20
iv) 24
v) It cannot be determined

		Museum		
		Yes	No	Total
Home Depot	Yes	? ?		80
	No			40
	Total	30	90	120

e) You are given the following optimization problem for solving a learning task.

$$argmin_\theta \left[\sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda \sum_{p=1}^{d} \theta_p^2 \right]$$

Which of the following statements are true?

i) There are *d* data points
ii) There are *n* data points
iii) The data is *d* dimensional
iv) This is a classification problem
v) This is a linear model
vi) This problem uses LASSO ($L1$) regularization
vii) Larger values of λ imply increased regularization
viii) Larger values of λ will likely increase variance
ix) Larger values of λ will likely increase bias
x) None of the above are true

f) Select all statements that are true:

i) If there are two identical features in the data, the Ridge (L2) regularization will force the coefficient of one redundant feature to be 0.
ii) We cannot use linear regression to find the coefficients $\boldsymbol{\theta}$ in $y = \theta_1 x^3 + \theta_2 x^2 + \theta_3 x + \theta_4$ since the relationship between y and x is non-linear.
iii) Introducing more features increases the model complexity and may cause overfitting.
iv) All of the above
v) None of the above

Question 2: Regression (10 points)

Engineers want to know what factors are associated with gas mileage. The regression below predicts the average miles per gallon (MPG) for 82 cars using their engine horsepower (HP) and weight (WT, in 100's of pounds).

Dependent variable is MPG (R-squared = 82.4%, R-squared (adjusted) = 81.9%)				
Variable	**Coefficient**	**SE (Coefficient)**	**t-ratio**	**p-value**
Intercept	66.855	2.079	32.3	<0.0001
HP	-0.0209708	0.015	-1.4	0.1661
WT	-0.990369	0.1047	-9.45	<0.0001

a) Write down the regression equation for this model and compute the predicted gas mileage of a 3500 pound car with a 150 horsepower engine.

(1 point)

MPG = 66.855 - 0.021HP - 0.99WT

66.855 -0.021 x 150 - 0.99 x 35 = 29.055 miles per gallon.

Remember that weight is measured in hundreds of pounds, so we plug in 35, not 3500.

b) Write down the hypothesis for the test of the coefficient of horsepower. Conduct the test (assuming a significance level of 5%) and explain your conclusion in the context of this problem.

(2 points)

H_0: $\beta_{HP} = 0$

H_A: $\beta_{HP} \neq 0$

The P-value associated with horsepower is 0.1661, which is larger than significance level of 5%. Therefore, for cars of similar weight, our data shows little evidence that horsepower is associated with gas mileage.

c) Would the hypothesis test conducted in b) have the same conclusion if you ran a simple regression model just using the horsepower variable (thus excluding the weight)? Explain shortly your answer.

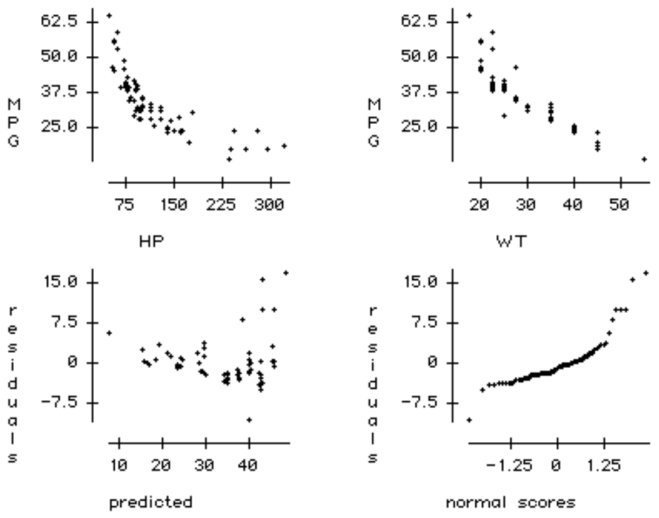
(1 point) Not necessarily. The coefficient is not significant in the presence of WT. For a simple regression model, we would have to make our computations from scratch.

d) Explain the meaning of the coefficient of weight and of the intercept (constant) in the context of this problem.

(2 points) For cars with similar horsepower, on average, each additional 100 pounds of weight is associated with a decrease in gas mileage of about 1 mile per gallon.

For cars with a zero horsepower engine and with a weight of zero pounds, the average gas mileage is 66.85 miles per gallon. This is a meaningless interpretation because it's impossible to have a car that weighs nothing and has a zero horsepower engine

e) The plots given below are: scatterplot MPG vs. HP (top-left), scatterplot MPG vs. WT (top-right), residuals vs. predicted values (bottom-left), and a normal probability plot of residuals (bottom-right).



Given these plots and the regression results, comment on how good the model we fit is.

(4 points) The plots suggest that the model is not a very good fit. Some of the things you could comment on:

* The plot of residuals vs. predicted values is curved, as is the plot of MPG vs. HP. This is a serious problem that undermines all of our regression results. It's possible that a re-expression of MPG and/or HP could correct this.

* The plot of residuals vs. predicted values seems to grow much wider for higher predicted values. This indicates that the variation of the regression errors is not constant. A re-expression of MPG to the log scale might correct this problem.

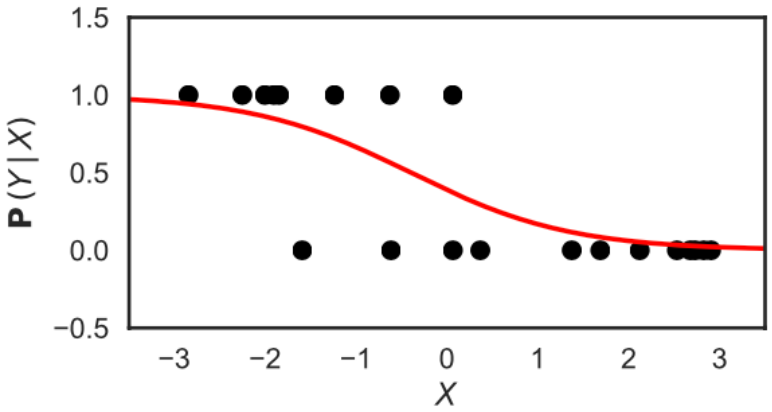
* The normal probability plot of the residuals is curved and shows three large outliers. The regression errors appear to be non-normal. A re- expression of MPG to the log scale might correct this problem.

* (Additional thought given): We aren't told how these 82 car models were selected for the study. We need to know that they were selected randomly, or at least in a way that seems to guarantee that the cars are not related in some way

Question 3: Logistic Regression (10 points)

Note: This question contains 3 sub-questions: c) can be answered independently of a) and b)

Suppose you are given the following dataset consisting of X and Y pairs. Suppose we are using logistic regression to solve the problem of predicting the outcome Y ($Y \in \{0, 1\}$) based on X (predictor).



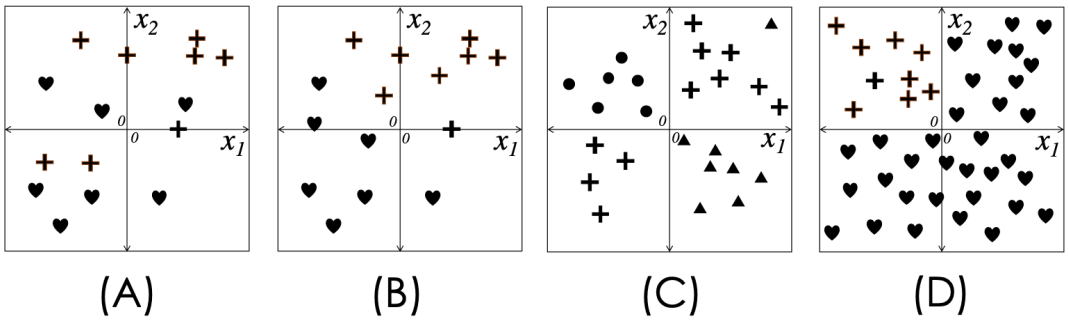
a) Roughly sketch the predictions (on figure) made by the logistic regression model for $P(Y=1|X)$. No justification needed. (2 points)

b) Given this data, what is your estimation for the value $P(Y=1 \mid X=3)$? Explain shortly your answer.

(2 points)

From the plot it's clear that for X=3, the probability is going to be close to 0.

Now, consider the following figures of different shapes plotted in a two-dimensional feature space (x_1, x_2). Suppose we are interested in classifying the type of shape based on the feature space.



c) For each of the cases (A-D) explain which algorithm between k-NN and logistic regression would perform best in terms of minimizing the classification error? Explain shortly your answer.

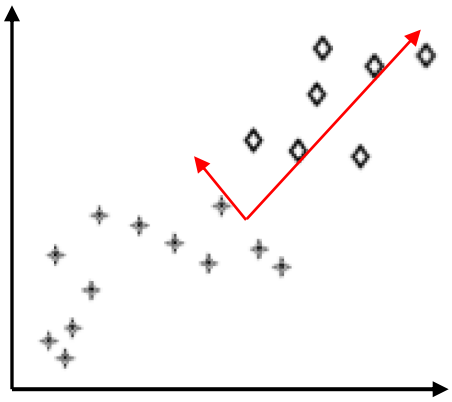
(6 points) In short: (B) is the only case that is linearly separable, so LR would be a good fit there.

In other cases, k-NN would probably perform best. Other answers, commenting on the value of k for k-NN were also considered correct, e.g. in (D) LR could classify everything correctly by making 2-3 mistakes. K-NN could possibly make more if we do not find a proper value of k.

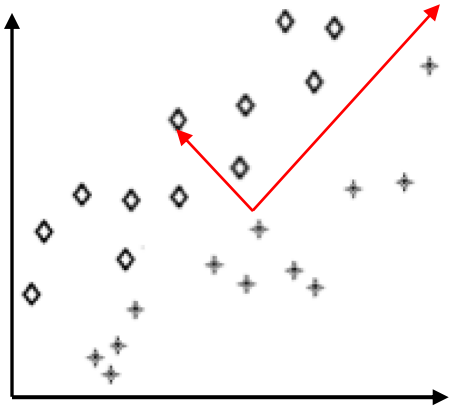
Question 4: Let's PCA (10 points)

Note: This question contains 3 sub-questions: c) can be answered independently of a) and b)

In the following plots, two datasets of points belonging to two classes (rhombuses and crosses) are given. Both datasets are two-dimensional.



(A)



(B)

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.

(3 points) (very) rough sketch above.

In both cases we expect a similar image. Remember that PCA looks for the direction of variance in the data (disregarding info about classes). Perhaps in the second case the 2nd PC is slightly larger.

b) Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain shortly why it is not possible.

(3 points) For dataset (A) we can project on PC1 (smaller values will be crosses and larger values will be rhombues) and for dataset (B) we can project on PC2. Note that in the second case PC2 will not hold that much information as PC1.

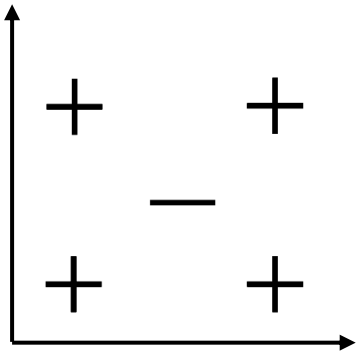
c) What would be one reason to use PCA as a pre-processing step for a regression task? Explain shortly your answer.

(4 points) Different answers possible (e.g. dimensionality reduction, remove correlated dimensions, etc.)

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

Note: This question contains 4 sub-questions: d) can be answered independently of a), b) and c)

Consider training a boosting classifier using decision trees of depth 1 (i.e. stumps) on the following two-dimensional dataset:

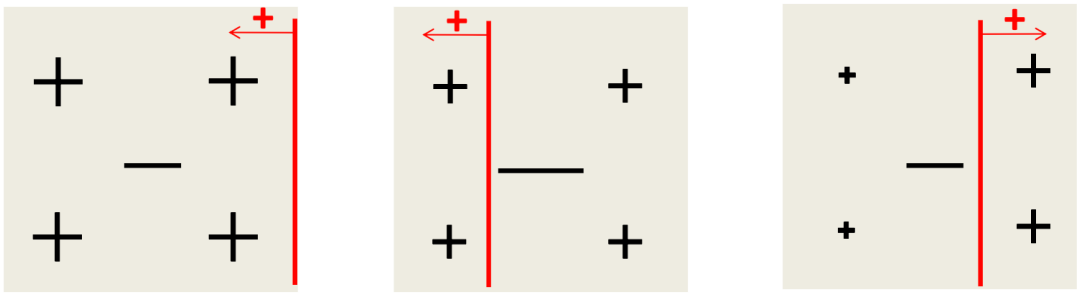


a) Which examples will have their weights increased at the end of the first iteration? Circle them (or write them down by describing their position) and explain shortly your answer.

(2.5 points) The negative example since the decision stump with least error in first iteration is constant over the whole domain. Notice this decision stump only predicts incorrectly on the negative example, whereas any other decision stump predicts incor- rectly on at least two training examples.

b) How many iterations will it take to achieve zero training error? Explain shortly your answer.

(2.5 points) At least three iterations. The first iteration misclassifies the negative ex- ample, the second iteration misclassifies two of the positive examples as the negative one has large weight. The third iteration is needed since a weighted sum of the first two decision stumps can't yield zero training error, and misclassifies the other two positive examples. (see rough figures below)



c) Can you add one more example to the training set so that boosting will achieve zero training error in two steps? If not, explain why.

(3 points) No. Notice that the simplest case is adding one more negative example in center or one more positive example between any two positive examples, as it still yields three decision regions with axis-aligned boundaries. If only two steps were enough, then a linear combination of only two decision stumps $\text{sign}(\alpha_1 h_1(x)+\alpha_2 h_2(x))$ should be able to yield three decision regions. Also notice that at least one of h1 or h2 misclassifies two positive examples. If only h2 misclassifies two positive examples, the possible decisions are (1) $\text{sign}(\alpha_1 - \alpha_2)$ on those two positive examples, (2) $\text{sign}(\alpha_1 + \alpha_2)$ on the remaining positive examples and (3) $\text{sign}(\alpha_1 - \alpha_2)$ on the negative examples - which don't yield zero training error since signs on (1) and (3) agree. If both h1 and h2 misclassify two positive examples, we have (1) $\text{sign}(\alpha_1 - \alpha_2)$ on two positive examples, (2) $\text{sign}(-\alpha_1 + \alpha_2)$ on the remaining positive examples and (3) $\text{sign}(-\alpha_1 - \alpha_2)$ on the negative - which again don't yield zero training error since signs on (1) and (2) don't agree.

d) Why do we want to use "weak" learners when boosting?

(2 points) To prevent overfitting, since the complexity of the overall learner increases at each step. Starting with weak learners implies the final classifier will be less likely to overfit.

Bonus Question (contributes to your bonus part)

How many followers (5% margin for the error) does @jerryinlaw have on Instagram? 290 (any answer in the range 275-305 got the bonus)