Maastricht University

Department of Data Science and Knowledge Engineering

Data Analysis 2018/2019

Final Exam

– Do not turn this page before the official start of the exam! –

First name, Surname: **Glenn Close (and not Glose)**

Student ID: ____

Expected Exam Grade (out of 50): ____ (optional)

Program: Bachelor Data Science and Knowledge Engineering

Course code: KEN3450

Examiner: Dr. Gerasimos (Jerry) Spanakis

Date/time: Monday, 1st April 2019, 13.00-16.00h

Format: Open book exam

Allowed aides: Pens, simple (non-programmable) calculator from the DKE-list of allowed calculators.

Instructions to students:

- The exam consists of 7 questions (plus one bonus question) on 17 (15+2 extra) pages.
- Fill in your name and student ID number on each page, including the cover page.
- Answer every question at the reserved space below the questions (keep your answers short and to the point). If you run out of space, continue on the back side, and if needed, use the extra blank page.
- Ensure that you properly motivate your answers.
- This exam sums to 50 points (plus 1 bonus point) and counts for 50% of your final grade. The rest 50% comes from data clinics (30%), data madness (10%) and Kaggle competition (10%)
- While you shouldn't spend too much time on calligraphy, please make sure that I have at least a chance at deciphering your exam.
- 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. You are not allowed to take any sheets, even blank, home.
- 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 you have not registered for the exam, your answers will not be graded, and thus handled as invalid.
- **Success! Break a pencil!**

The following table will be filled by the examiner:

Question:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**B**	**Total**
Maximum points:	9	8	8	5	5	6	9	1	**51**
Achieved points:									

Question 1: **EDA Warmup (9 points)**

A. The following table is consistent with the results from “Beverage Choices of Young Females: Changes and Impact on Nutrient Intakes” (Shanthy A. Bowman, *Journal of the American Dietetic Association*, *102*(9), pp. 1234-1239):

		Nationwide Food Survey Years			
		1987-1988	1989-1991	1994-1996	**Total**
Drinks Fluid Milk	Yes	354	502	366	**1222**
	No	226	335	366	**927**
	Total	**580**	**837**	**732**	**2149**

1.(1) Compute the following:

a. What percent of the young girls reported that they drink milk?

b. What percent of the young girls were in the 1989-1991 survey?

c. What percent of the young girls who reported that they drink milk were in the 1989-1991 survey?

d. What percent of the young girls in 1989-1991 reported that they drink milk?

a. 1222/2149 = 56.9%

b. 837/2149 = 38.9%

c. 502/1222 = 41.1%

d. 502/837 = 60.0 %

2.(1) What is the marginal distribution of milk consumption?

Yes 1222

No 927

Taking the marginal distribution over time was also a popular answer (or drawing the whole matrix with percentages).

3.(1) Do you think that milk consumption by young girls is independent of the nationwide survey year? Use statistics to justify your reasoning.

It was perfectly fine to compute the chi-square.

Also, just referring to the basic percentages (61%, 60%, 50%) was also fine.

4.(1) Draw (an approximate sketch is enough) one (or more) chart(s) to compare the milk consumption between ~~1989~~1989-1991 and 1994-1996.

Here comes the 1st typo: It was meant to say 1989-1991, which many of you correctly understood (that's what makes us humans), or asked for clarification. Assuming that this is a wrong date and just doing a plot for 1994-1996 was also fine.

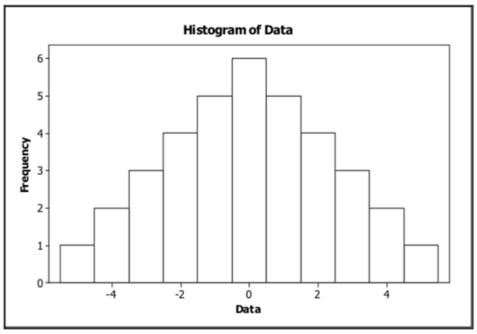
Obviously, any labeled plot here is correct (e.g. piechart -since you are showing ratios that should not be punished), or barplot, as long as you show the difference between the two year ranges. Notice that in one case you had a 50-50 split which means that it should be something that you should highlight.

B. For each of the following questions circle the correct one (A, B, C, D or E). There is no need to justify your answer.

[A correct answer gives you +1, a wrong answer gives you -1, a non-answer gives you 0]

1.(1) Which is true of the data shown in the histogram?

I. The distribution is approximately symmetric.
II. The mean and median are approximately equal.
III. The median and IQR summarize the data better than the mean and standard deviation.



A) I only

B) III only

C) I and II

D) I and III

E) I, II, and III

2.(1) Jerry tells you your test score was the 3rd quartile for the class. Which is true?

I. You got 75% on the test.
II. You can't really tell what this means without knowing the standard deviation.
III. You can't really tell what this means unless the class distribution is nearly Normal.

A) I only
B) II only
C) III only
D) II and III
E) none of these

3.(1) The five-number summary of credit hours for 24 students in a statistics class is:

Min	Q1	Median	Q3	Max
13.0	15.0	16.5	18.0	22.0

Which statement is true?

A) There are no outliers in the data.
B) There is at least one low outlier in the data.
C) There is at least one high outlier in the data.
D) There are both low and high outliers in the data.
E) None of the above.

C. For the following two questions give short answers:

1.(1) Pearson's correlation ρ between two variables $\boldsymbol{x}$ and $z \in \mathrm{R}^{\mathrm{P}}$ is given by

$$\rho(\boldsymbol{x},\boldsymbol{z}) = \frac{\sum_{j=1}^{p}(x_j - \overline{\boldsymbol{x}})(z_j - \overline{\boldsymbol{z}})}{\sqrt{\sum_{j=1}^{p}(x_j - \overline{\boldsymbol{x}})^2}\sqrt{\sum_{j=1}^{p}(z_j - \overline{\boldsymbol{z}})^2}}$$

where $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{z}}$ are the mean values. If the data are centered, which distance measure you get?

The cosine distance (or similarity measure).

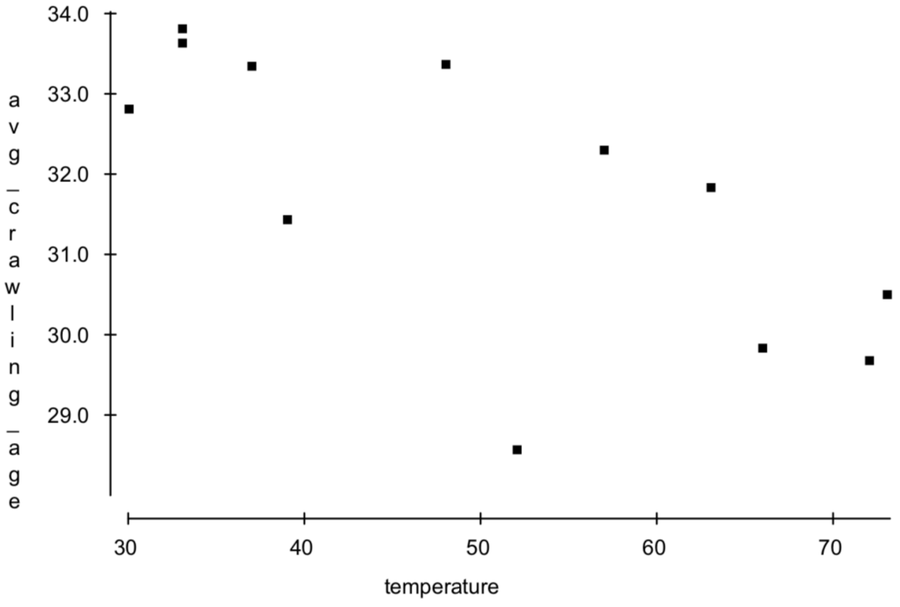
2.(1) Suppose that X, Y, and Z are random variables. X and Y are positively correlated and Y and Z are likewise positively correlated. Does it follow that X and Z must be positively correlated?

Correlation is not transitive, so it does not follow that X and Z must be positively correlated.

Question 2: **Crawling Regression (8 points)**

Researchers at the University of Denver Infant Study Center investigated whether babies take longer to learn to crawl in cold months (when they are often bundled in clothes that restrict their movement) than in warmer months. The study sought an association between babies' first crawling age (in weeks) and the average temperature during the month they first try to crawl (about 6 months after birth). Between 1988 and 1991 parents reported the birth month and age at which their child was first able to creep or crawl a distance of four feet in one minute. Data were collected on 208 boys and 206 girls. The graph below plots average crawling ages (in weeks) against the mean temperatures when the babies were 6 months old. The researchers found a correlation of $r = -0.70$ and their line of best fit was

$$\widehat{AvAge} = 36 - 0.08 AvTemp\ .$$



A.(1) Draw the line of best fit **on** the graph. (Show your method below clearly)

To draw a line you need just 2 points. Use the regression equation to find such 2 points (e.g. 30, 33.6 and 70, 30.4) and draw the line.

B.(1) Describe the association in context.

The association is linear, moderately strong, and negative, with one outlier. Children seem to crawl earlier when the temperature is higher, though there was an unusually early age observed for a temperature just above 50°.

C.(1) Explain (in context) what the slope of the line means.

The model suggests that, on average, babies crawl 0.8 weeks earlier for every 10° higher the temperature is.

D.(2) Explain (in context) what the *y*-intercept of the line means. Comment on its validity/usefulness (in context).

The model predicts that at a temperature of 0° babies would crawl at an average of 36 weeks old.

Since no data was collected for such cold temperatures this does not mean anything and perhaps we should avoid any conclusions based on this. The intercept is there just to "calibrate" the regression line.

E.(2) Compute R^2 for this example and explain (in context) what R^2 means.

In this simple setting, you just take the square of the correlation.

49% of the variability in crawling age can be explained by variations in temperature.

F.(1) In this context, what does a negative residual indicate?

A negative residual would indicate that babies crawled at a younger age than the model predicted.

Question 3: Text basics (8 points)

A. Assume that we have a collection of 5 documents in total. Let the indexing vocabulary consist of 3 terms.

1.(1) Assuming that index terms t_1, t_2 and t_3 appear in 3, 4 and 2 documents respectively, determine their *idf* weights. Comment on the result.

Idf1=log(5/3) = 0.22
Idf2=log(5/4) = 0.09
Idf3=log(5/2) = 0.40

Taking ln was also fine.

2.(1) Let one of the documents be represented by the vector $\boldsymbol{d} = \mathbf{(0, 3, 5)}$, where the elements represent the term frequencies (f_j, where $j \in [1,3]$). Let tf_j be the weights of ***d*** (where $j \in [1,3]$), be defined as:

$$tf_j = \begin{cases} 0, if \;\; f_j = 0 \\ 0.3 + \frac{(1-0.3) \cdot f_j}{argmax_k(f_k)} \end{cases}$$, where $argmax_k(f_k)$ refers to the maximum value of *f*.

Using results of previous question (1) and the above information, determine the representation of *d* given by *tf-idf* weights. You do not need to normalize the *tf-idf* weights.

Remember tf-idf means that you multiply the tf value of each term (as given by the formula) with the idf value.

Tf-idf(d) = (0, 0.3 + (0.7x3)/5 x 0.09, 0.3 + 0.7x5/5 x 0.40) = (0, 0.07, 0.58)

3.(2) Assuming the d vector with *tf-idf* weights and that a query is given by ***q*=(1,0,0)** what is the similarity of the query to the document d? Make appropriate choices and comment on the result.

Without any calculation, the cosine distance (or similarity) will give you 0 (check why).

That makes sense, since with our query q we are looking for t1, but that does not exist in document d (weight is 0), so there is no value in saying that there is any degree of similarity.

B. For the 520-million-word Corpus of Contemporary American English, we have the following counts of unigrams and bigrams:

your	883.614
rights	80.891
doorposts	21
your rights	378
your doorposts	0

1.(1) Estimate the probabilities *P*(rights) and *P*(rights **|** your).

P(rights) = 80.891/520.000.000
P(rights|your) = 378/80.891

2.(2) Estimate the probability *P(doorposts* / *your)*. What seems to be the issue with the result and how would you fix it?

The probability is zero. Which means that we didn't see that sequence (your doorposts) in the training data.
Different fixes can be proposed:

-Do normalization
-Add 1 to all counts so that no probability is zero
-…

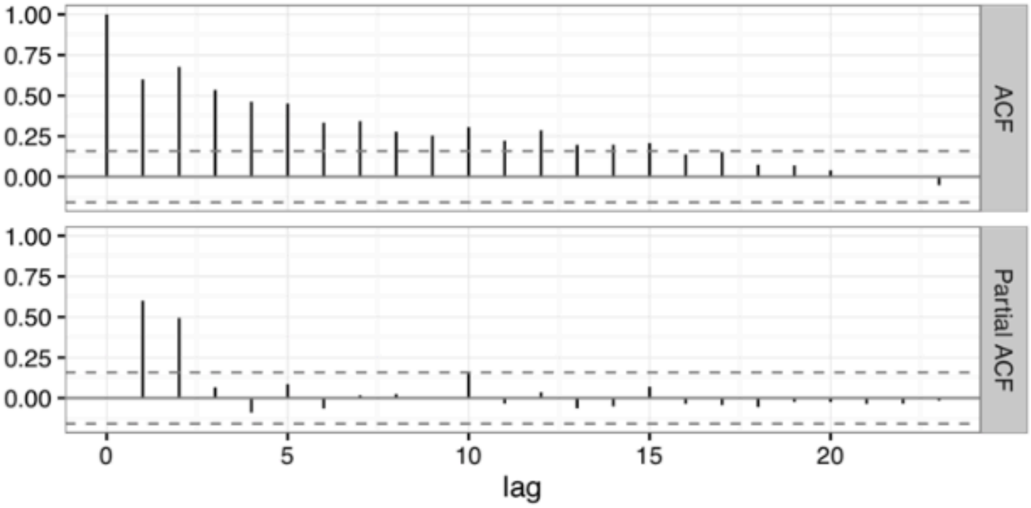
3.(1) Suppose that we train three *n*-gram models on 38 million words of newspaper text: a unigram model, a bigram model, and a trigram model. Suppose further that we evaluate the trained models on 1.5 million words of text with the same vocabulary and obtain the following perplexity scores: 170, 109, and 962. Which perplexity belongs to which model? Provide a short explanation.

Under reasonable assumptions, higher-order models yield lower perplexity scores, because we take into account more context. Thus the proper assignment is: 962, unigram; 170, bigram; 109, trigram.

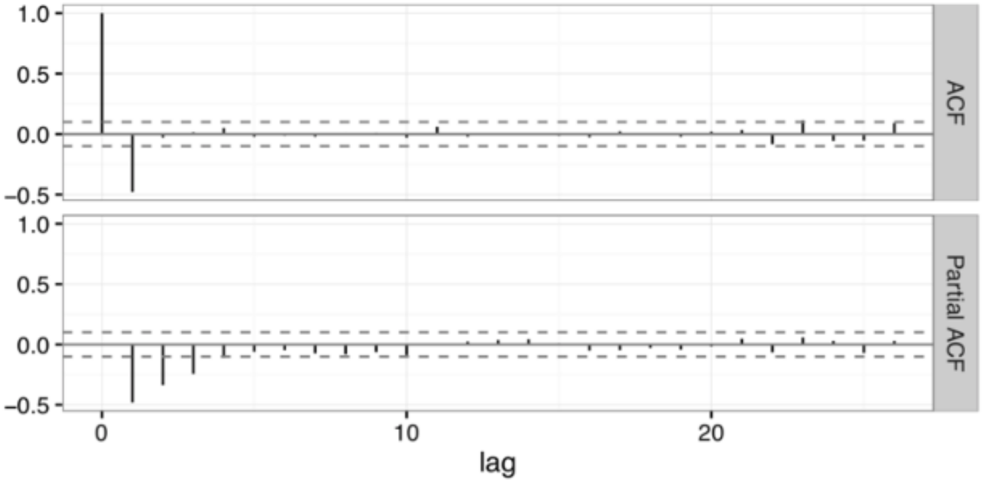
Question 4: TimeSeries (5 points)

A.Below are the ACF and PACF for ~~three~~two time series. For each series, state whether it is autoregressive **or** moving average, and the order (p or q). Justify your answer in 1-2 lines.

i.(1) Circle one: **AR** or MA order (p or q) = **2**



ii.(1) Circle one: AR or **MA** order (p or q) = **1**



B.(3) ARIMA models include a parameter, d, that controls the number of times a time series is differenced before being modeled by an ARMA process. Why is differencing a time series sometimes necessary? How could you choose the amount of differencing required for a particular time series?

Mainly to achieve stationarity.
(Differencing removes non- stationarity beyond just trend, e.g. can remove the non-stationary variance as well)

Observe the series after each difference, and difference until it appears stationary.

Question 5: Experimental Design

A.(3) An article in a local newspaper reported that dogs kept as pets tend to be overweight. Veterinarians say that diet and exercise will help these chubby dogs get in shape. The veterinarians propose two different diets (Diet A and Diet B) and two different exercise programs (Plan 1 and Plan 2). Diet A: owners control the portions of dog food and dog treats; Diet B: a mixture of fresh vegetables with the dog food and substitute regular dog treats with baby carrots. Plan 1: three 30-minute walks a week; Plan 2: 20-minute walks daily. Sixty dog owners volunteer to take part in an experiment to help their chubby dogs lose weight.

Design an experiment to determine whether the diet and exercise programs are effective in helping dogs to lose weight. What more information would you need in order to make this a randomized block experiment and mention an example of when that might be useful.

Different setups possible, as long as you mentioned how you split the groups to random ones (of equal samples) and what is the measure variable you compute in the end and how you are checking that for significance.

For randomized blocked experiment we need more info about the dogs (e.g. specific breed, age, etc.)

B.(2) Suppose that we have a multiple-choice exam with 20 questions and T/F answers. Explain (briefly) how the examiner (without knowing the real answers) could use crowdsourcing to know the real answers. Is there a way that the examiner can take into account the "reliability" of each student (without having any background info on them)?

Majority voting (no need to perform another experiment, since we already have the answers. Any exam could be seen as a form of crowdsourcing but obviously and in the spirit of assessment nobody is grading like this.

For the "reliability" check the slides for the alternating calculations between the correct answer percentage and then the reliability of the student, etc.

Question 6: Dimensionality Reduction (6 points)

A.(2) A data scientist runs a principal component analysis (PCA) on their data and tells you that the percentage of variance explained by the first 3 components is 80 %. How is this percentage of variance explained computed? Explain the intuition (and do not blindly copy what you see in the slides).

Refer to how we need to find new axes (components) that explain as much variance possible from the original data as possible. Referring to the variance matrix and the eigenvectors also is fine. You should also justify that the result suggests that by using the three new components (suggested by PCA) explains 80% ot the variance in the original data.

B.(2) The same data scientist suggests using PCA as a pre-processing step before regression. Which of the following sentences are correct about this action (circle the ones that are correct)?

[A correct answer gives you +1, a wrong answer gives you -1, an non-answer gives you 0]

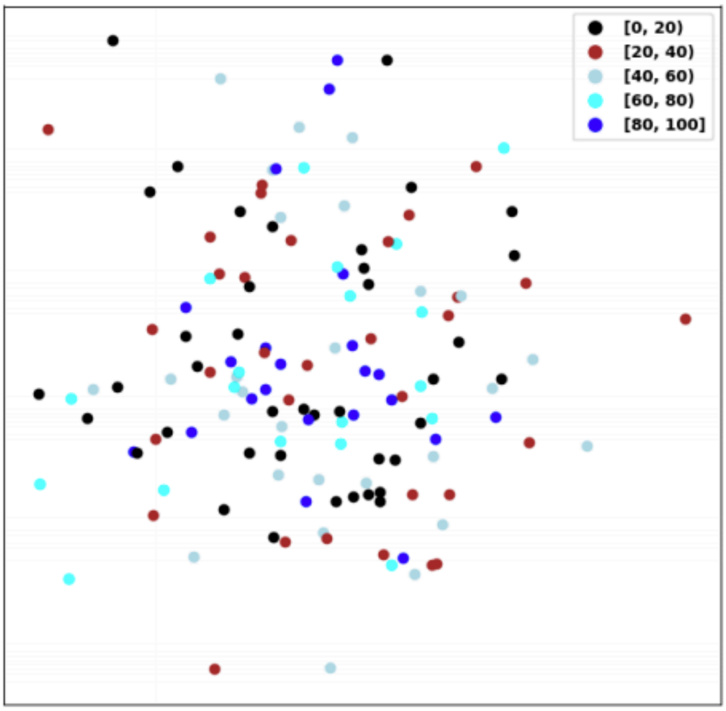
(A) It will reduce overfitting by removing poorly predictive dimensions

(B) It will expose information missing from the input data

(C) It will make computation faster by reducing the dimensionality of the data

(D) For inference and scientific discovery, we prefer features that are not axis-aligned

C.(2) In a recent computer science conference, Jerry attended a paper presentation where the author showed the following figure of PCA (showing 2 principal components) of a large dataset of 5 groups (shown in different color). Based on this visualization and your intuitive knowledge of PCA, explain why Jerry rolled his eyes (🙄) and then make a rough estimate of the variance explained by these 2 principal components.



THIS QUESTION IS EXCLUDED FROM THE EXAM SINCE IT WAS SUPPOSED TO BE PRINTED IN COLOR (BUT IT WAS NOT).

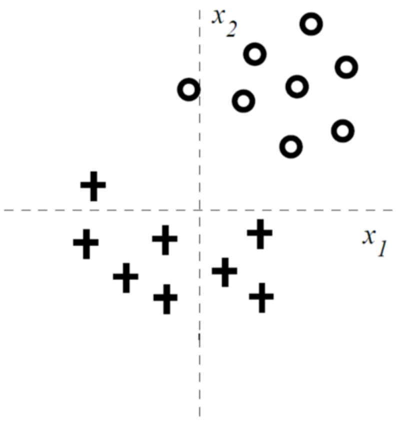
What you could have answered (some of you did):

1)Bad plot since there are no labels.

2)The result suggests that the different data points (belonging to different classes) are almost everywhere. There is no separation at all. That should suggest that the variance explained is very low (which indeed it was, around 5%).

Question 7: **Regularization (9 points)**

Consider the 2-dimensional labeled training set of the following Figure where **'+'** corresponds to class $y=1$ and **'O'** corresponds to class $y = 0$.



We attempt to solve the binary classification task depicted with the simple linear logistic regression model.

$$P(y = 1|\boldsymbol{x}, \boldsymbol{w}) = g(w_0 + w_1x_1 + w_2x_2) = \frac{1}{1 + \exp(-w_0 - w_1x_1 - w_2x_2)}$$

Consider training regularized linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} \log(P(y_i|x_i, w_0, w_1, w_2)) - \lambda \cdot w_j^2$$

for very large λ.

The regularization penalties used in penalized conditional log- likelihood estimation are $-\lambda w_j^{\ 2}$, where j = {0,1,2}. In other words, only one of the parameters w_j is regularized in each case.

A. Given the training data of the figure, how does the training error change with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j for very large λ. Provide a brief justification for each of your answers.

1.(1) By regularizing w_2

Increases. When we regularize w2, the resulting boundary can rely less and less on the value of x2 and therefore becomes more vertical. For very large λ, the training error increases as there is no good linear vertical separator of the training data.

2.(1) By regularizing w_1

Remains the same. When we regularize w1, the resulting boundary can rely less and less on the value of x1 and therefore becomes more horizontal and the training data can be separated with zero training error with a horizontal linear separator.

3.(1) By regularizing w_0

Increases. When we regularize w0, then the boundary will eventually go through the origin (bias term set to zero). Based on the figure, we can not find a linear boundary through the origin with zero error. The best we can get is one error.

B.(2) If we change the form of regularization to L1-norm (Lasso) and regularize w_1 and w_2 only (but not w_0), we get the following penalized log-likelihood to maximize.

$$\sum_{i=1}^{n} \log(P(y_i|x_i, w_0, w_1, w_2)) - \lambda \cdot (|w_1 + w_2|)$$

Consider again the problem in the previous figure and the same linear logistic regression model. As we increase the regularization parameter λ which of the following scenarios do you expect to observe? Circle only one and briefly explain your choice.

(A) First w_1 will become 0, then w_2.
(B) First w_2 will become 0, then w_1.
(C) w_1 and w_2 will become zero simultaneously.
(D) None of the weights will become exactly zero, only smaller as λ increases.

The data can be classified with zero training error and therefore also with high log-probability by looking at the value of x2 alone, i.e. making w1 = 0. Initially, we might prefer to have a non-zero value for w1 but it will go to zero rather quickly as we increase regularization. Note that we pay a regularization penalty for a non-zero value of w1 and if it does not help classification why would we pay the penalty?

As λ increases further, even w2 will eventually become zero. We pay higher and higher cost for setting w2 to a non-zero value. Eventually this cost overwhelms the gain from the log-probability of labels that we can achieve with a non-zero w2.

Note that the absolute value around the sum of the parameters should not affect the result as λ increases.

C.(2) For very large λ, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (Note that the number of points from each class is the same. You can give a range of values for w_0 if you deem necessary).

For very large λ, we argued that both w1 and w2 will go to zero. Note that when w1 = w2 = 0, the log-probability of labels becomes a finite value, which is equal to nlog(0.5) and that suggests that w0 = 0.

In other words, P(y = 1|x,w)=P(y = 0|x,w)=0.5. We expect so because the number of elements in each class is the same and so we would like to predict each one with the same probability! How cool is logistic regression?!

D.(2) Assume that we obtain more data points from the '+' class that corresponds to $y=1$ so that the class labels become unbalanced. Again, for very large λ, with the same L1-norm regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (You can give a range of values for w_0 if you deem necessary).

For very large λ, we argued that both w1 and w2 will go to zero.

With unbalanced classes where the number of '+' labels are greater than that of 'o' labels, we want to have P(y = 1|x,w) > P(y = 0|x,w). For that to happen the value of w0 should be greater than zero which makes P (y = 1|x,w) > 0.5. Some of you argued that this is going to be relative to the number of data points, or some of you went even further and did actual calculations of the probability. That was great work which I did not expect, so kudos to you!

Bonus Question (1 point)

Glenn ~~Glose~~Close holds the record for the actress with the most Oscar nominations without winning.

How many nominations does she have? Google is your friend.

Extra answer sheet.

Extra answer sheet.