# Homework 2
# STA 221

Bohao Zou 917796070
Bingdao Chen 917781027
University of California, Davis

May 18, 2020

# Question 1

1. Gradient descent is a supervised learning algorithm.
2. Gradient descent is an algorithm to minimize or maximize a function.
3. Logistic regression cannot be performed after linear PCA.
4. Support vector machine is non-linear classification algorithm.
5. A regression algorithm can be modified to be used for classification as well.

## Solution

1. False. Supervised Learning consists of two problems: Regression and Classification. Gradient descent is an optimization algorithm used to find the minimize of functions.
2. True. Gradient descent is an algorithm to minimize a function. For the maximization problem of function $f(x)$, we can transfer it to find the minimize of the $-f(x)$.
3. False. PCA is to reduce the dimension of the data by coming up with linear combinations of the variables that maximize the variance explained. Logistic regression can also be performed on these newly masde variables.
4. False. Support vector machine is linear classification algorithm, because it creates a line or a hyperplane which separates the data into classes. However, SVM with kernel trick is a non-linear classification for mapping the data into the high-dimensional space.
5. True. We can add some function to the output of regression like sigmoid or softmax to transform the regression problem to classification problem.

# Question 2

Because of one page limitation, we have write the report into ***GAN_Report.pdf***.Please see the pdf file ***GAN_Report***.

# Question 3

## Solution of (a)

The estimated $\hat{\vec{\beta}}$ is :

$$\begin{aligned}\hat{\vec{\beta}} = [&-9.79e^{-2}, 4.89e^{-2}, -2.53e^{-2}, 3.45 - 3.55e^{-1}, 5.81, -3.31e^{-3}, -1.02,\\ &2.27e^{-1}, -1.22e^{-2}, -3.88e^{-1}, 1.7e^{-2}, -4.85e^{-1}]^T\end{aligned}$$

The MSE for training data is 24.476.

### Solution of (b)

The MSE for test data is 24.292.

### Solution of (c)

The MSE of polynomial linear model for training data is 5.4316 and the MSE for test data is 34.6864.
Compared with the previous model, the mean squared prediction error of the polynomial linear regression model on train data is smaller, which indicates that it fits better. However, the mean squared prediction error of the polynomial linear regression model on test data gets greater, which indicates that the polynomial model results in overfitting.

## Question 4

### Solution of (a)

For rating value 1, there are 656 reviewers.
For rating value 5, there are 656 reviewers.

The best performance of a constant classifier is that assigning all the reviews to one class which has the highest probability in the data. In this case, because the number of rating = 5 and rating = 1 are same, the misclassfication rate is 0.5.

### Solution of (b)

Because the regularization penalty is comprised of the sum of the absolute value of the coefficients, we need to scale the data such that its coefficients are all based on the same scale. Finally, we used the code $LogisticRegression(penalty =' l1', solver =' liblinear')$ to add L1-regularized into this linear model. The details of codes are in Jupyter file.

### Solution of (c)

There are 443 covariates have non-zero coefficients in Document-Term Matrix. There are 460 covariates have non-zero coefficients in Tf-idf Matrix.

For ***Document-Term Matrix***, the twenty words with the most ***positive*** coefficients are ***'lov', 'teething', 'soph', 'loved', 'voic', 'easy', 'great', 'she', 'highly', 'teeth', 'gift', 'hear', 'chew', 'seen', 'sensit', 'littl', 'channel', 'sony', 'excellent', 'other'***.

For ***Document-Term Matrix***, the twenty words with the most ***negative*** coefficients are twenty words are ***'not', 'returned', 'leak', 'out', 'something', 'off', 'wast', 'stopped', 'after', 'useless', 'back', 'return', 'get', 'even', 'becaus', 'doesn', 'bin', 'leaked', 'disappointed', 'pictur'***.

For ***Tf-idf Matrix***, the twenty words with the most ***positive*** coefficients are twenty words ***'lov', 'teething', 'soph', 'loved', 'she', 'voic', 'highly', 'great', 'gift', 'toy', 'sensit', 'littl', 'easy', 'channel', 'teeth', 'other', 'confined', 'hear', 'her', 'he'***.

For ***Tf-idf Matrix***, the twenty words with the most ***negative*** coefficients are twenty words are ***'not', 'wast', 'off', 'out', 'diap', 'after', 'leak', 'bottl', 'the', 'swing', 'wrap', 'horribl', 'return', 'motorol', 'doesn', 'something', 'beeping', 'returned', 'you', 'pictur'***.

## Solution of (d)

The misclassification rate of Document-Term for linear model is 0.0786.
The misclassification rate of Tf-idf for linear model is 0.0711.

The misclassification rate of Document-Term and Tf-idf for constant classifier is 0.4797.

From the result above, the misclassification rate of both model are very low. However, for the constant classifier, because it ignores the content of reviews and blindly assigns all reviews to one class, the results of document term matrix and Tf-idf matrix are same. It assigns all the review to the "bad" class, and the misclassfication rate is 0.4797, which is much higher than the misclassfication rate of the first model. The logisitic model is better.

## Pledge

Please sign below (print full name) after checking the following. If you can not honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

1. We pledge that we are honest students with academic integrity and we have not cheated on this homework.
2. These answers are our own work.
3. We did not give any other students assistance on this homework.
4. We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
5. We understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

**Team Member 1: Bingdao Chen** **Team Member 2: Bohao Zou**