CIS 490 Machine Learning

Lecture 8

Instructor: (Julia) Hua Fang

1

2

Next two weeks:

- Feb 15: in-class Lecture on Classification.
- Feb 17: No lecture meet; use this common time for group meets to review contents and complete your sectional project; TA in classroom for Q&A and instructor on Zoom for Q&A, a zoom link posted at myCourses.
- Feb 22: follows Monday's class schedule due to the Holiday on Monday. Check Univ. Calendar and Univ. Policy.
- Feb 24: Sectional project 1 presentation on Zoom

2

Last time

3

- Supervised Learning
  - ➢Regression
    - ❖ Linear regression
    - ❖ Regularized linear regression

      You are expected to understand:

      ✓ Overfitting: Random error vs. deterministic error; Bias-Variance tradeoff

      ✓ Regularization for linear regression:
      - Ridge regression
      - Lasso regression

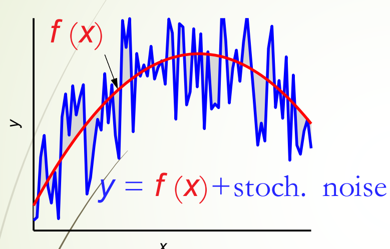
        -- Cross Validation

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

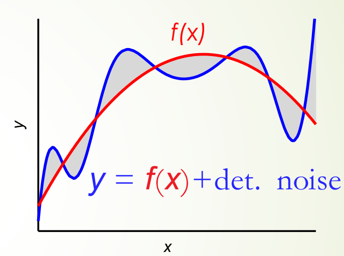
3

Answers to In-class exercise in LS7

Random Error/Stochastic Noise



Deterministic Error/Noise



Where does Stochastic or Deterministic Noise come from?

Answers:

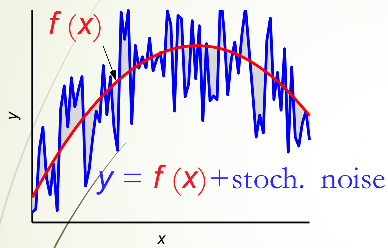
Stochastic Noise Source:
 random measurement errors

Deterministic Noise Source:
 learner f cannot model *y*

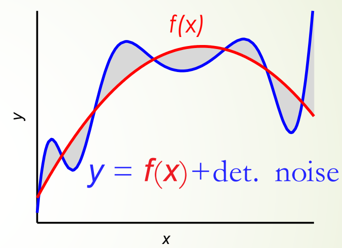
4

Answers to In-class exercise in LS7

Random Error/Stochastic Noise



Deterministic Error/Noise



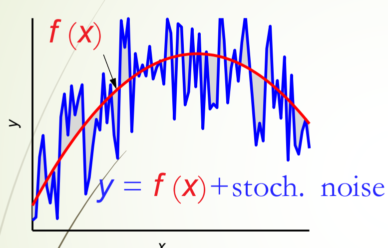
If we remeasure y, would Stochastic Error change? Yes,

If we remeasure y, would Deterministic Noise change? No.

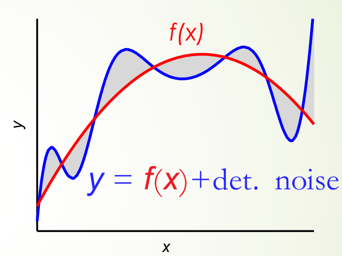
5

Answers to In-class exercise in LS7

Random Error/Stochastic Noise



Deterministic Error/Noise



If we change f(x), would Stochastic Error change? No

If we change f(x), would Deterministic Noise change? Yes

6

7

In class exercise

1. Given the amount of MSE is fixed, if variance increases, will bias increase or decrease? Decrease
2. Given the amount of MSE is fixed, if bias increases, will variance increase or decrease? Decrease
3. Given the amount of bias is fixed, if variance increases, will MSE increase or decrease? Decrease

Recall SSE/N =MSE = Variance + Bias^2

7

8

Outline

- Regularized linear regression (2)
  - ➢Ridge regression
  - ➢Lasso regression
  - ➢Running R for regularized linear regression

    You are expected to :
    - ❖ Interpret output
    - ❖ Understand Cross-validation (CV)
    - ❖ Understand how to use CV to choose optimal λ

Adapted from Jeff Howbert, Greg Shakhnarovich, Patrick Breheny , M. Magdon-Ismail, Patrick Breheny, Jeff Schneider

8

9

Lasso regression

9

Lasso Regression

Lasso: Least absolute shrinkage and selection operator

- Goal: Lasso "*shrinks some coefficients and sets others to 0, and hence tries to retain the good features of both subset selection and ridge regression*" (Tibshirani)

Author: Robert Tibshirani, (note: Last author of your reference book, ISLR)

Paper title: "*Regression Shrinkage and Selection via the Lasso*" Published in Journal of the Royal Statistical Society 1996).

10

Lasso Regression vs. Ridge Regression

- Lasso: L 1 Regularization

Goal: Find $\hat{\beta}^L_\lambda$ which minimizes:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \boxed{\lambda\sum_{j=1}^{p}|\beta_j|}$$

Or equivalently

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \boxed{\sum_{j=1}^{p}|\beta_j| \le s}$$

- Ridge: L 2 Regularization

Goal: Find $\hat{\beta}R$ which minimizes:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \boxed{\lambda\sum_{j=1}^{p}\beta_j^2}$$

Or equivalently

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \boxed{\sum_{j=1}^{p}\beta_j^2 \le s}$$

Note: other literature uses the symbol t instead of s

$\|\beta\|_1$ denotes the ℓ_1 norm (pronounced "ell 1") of a vector and is defined as

$$\|\beta\|_1 = \sum|\beta_j|$$

$\|\beta\|_2$ denotes the ℓ_2 norm (pronounced "ell 2") of a vector and is defined as

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p}{\beta_j}^2}$$

11

Lasso Regression vs. Ridge Regression

- Lasso: L 1 Regularization
  - Yields sparse models (ie. models that involve only a subset of the variables)
  - The L 1 Penalty $\lambda\sum_{j=1}^{p}|\beta_j|$ forces some coefficients exactly equal to 0 when λ is sufficiently larger!

- Ridge: L 2 Regularization
  - include all p predictors in the final model: produce a coefficient estimate for each predictor.
  - The L 2 penalty $\lambda\sum\beta_j^2$ shrinks all of the coefficients towards zero, but not set any of them exactly to zero (unless λ = ∞), no matter how small the value is.

So, Lasso has a major advantage over Ridge:

- Lasso performs **variable selection** and find models easier to interpret

12

13

Lasso and Ridge: λ

How to find the appropriate tuning parameter λ?

13

14

Cross-validation!

14

15

Cross-validation: Why

Two potential drawbacks with the training/testing set split:

- the test error rate can be highly variable, depending on how you split the dataset
- For only one training set (a subset of an entire set), machine learning methods tend to perform worse when trained on fewer observations.

15

Cross Validation (CV)- 3 types

- Holdout method: 2 subsets (training vs. testing) You knew this and its potential drawbacks already.
- K-fold CV: one fold (subset) as testing while the remaining K-1 as the training, eg., 5- or 10-folds
  - Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times.
- Leave-one-out (LOO): logical extreme of K-fold CV, ie., K = N, (N: the number of observations) e.g. each case is one fold!

We focus on regular K-fold CV: e.g. 5- or 10-folds, as they empirically yield a test error that does not suffer from excessively high bias, nor from very high variance

https://www.cs.cmu.edu/~schneide/tut5/node42.html

16

K-fold CV: Algorithm

i. Partition the data *T* into *K* groups, or folds, of equal size.

– Suppose T = (T_1, T_2, . . . ,T_K)

– Commonly chosen K's are K = 5 and K = 10

ii. For each fold *K*= 1,2,…,*k*, use the *k*th-fold (called held-out fold) for testing, and fit a model *f (x) to the rest of the data.*

1. K=1 used for testing, fit f(x) on the remaining, compute MSE_1
2. K=2 used for testing, fit f(x) on the remaining, compute MSE_2

…

K. K=k used for testing, fit f(x) on the remaining, compute MSE_K

iii. Compute k-fold CV error

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

17

Lasso and Ridge: K-fold CV for choosing λ

Step 1: Choose a grid/range of λ values,
eg. λ =0.1, 10, 1000

Step 2: apply K -fold CV (see previous slide), where f(x)= $\hat{f}^{\lambda^*}(x)$ (Lasso/Ridge).

For λ = 0.1, compute CV_error $_{\lambda=0.1}$=? (e.g. CV_error $_{\lambda=0.1}$ =20)

For λ = 10, compute CV_error $_{\lambda=10}$=? (e.g., CV_error $_{\lambda=10}$ =2)

For λ = 1000, compute CV_error $_{\lambda=1000}$=? (e.g., CV_error $_{\lambda=1000}$ =100)

Step 3: choose λ* as the minimizer of CV error. (e.g, λ* =10 in this example)

Step 4: Then refit the model with λ* (e.g., λ* =10) on the entire training set

18

19

R: Running Ridge, Lasso and K-fold CV for choosing λ

19

20

R: Running Ridge and K-fold CV for choosing λ using Credit data

- Real Data Set: "Credit.csv", posted at MyCourses. You can also download it from

  `https://www.kaggle.com/ishaanv/ISLR-Auto?select=Credit.csv`

Goal: Perform regularized regression using Ridge, Lasso and select the best lambda, λ^* , using k-fold Cross Validation.

Predict Y: 'Balance', using X.

20

21

R: Credit data

```
credit<read.csv("/Users/hfang/Downloads/CIS490_2020Spring
/Credit.csv")
credit <- credit[, 2:12] # we don't need the first column
```

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903

```
#change text variables to numeric, e.g. Gender Male/Female
changed to 1/0.

credit.mat <- model.matrix(Balance ~ .-1, data=credit)

# Delete unnecessary info at the bottom
credit.mat <- credit.mat[,-8]
set.seed(1) # Set seed for reproducibility

# Separate the features (independent) from the target
(dependent) variables
x <- credit.mat
y <- credit[, 'Balance']
```

21

22

R: Running Ridge and K-fold CV for choosing λ* using Credit data

22

23

R: Steps for Running Ridge and K-fold CV for choosing λ* and finding the final model

- Create a list of possible lambda values at which to evaluate the ridge model
- Run the ridge model at each lambda values
- Plot the coefficient results at different values of lambda
- Output numerical cross-validated results
- Choose λ* as the minimizer of CV error: Plot MSE performance at each lambda and calculate at which lambda (ie. finding the optimal lambda, λ*) the MSE has the minimum value, or the error is within 1 standard error of the minimum MSE.
- Use the optimal λ* to refit the model on the entire data and evaluate the model using MSE/RMSE.

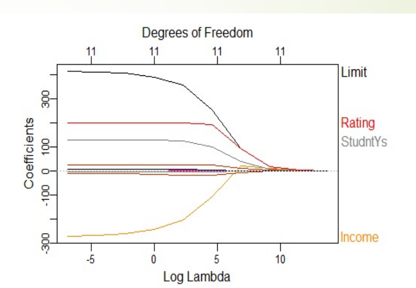
23

24

R: Running Ridge and K-fold CV for choosing λ* using Credit data

Run Ridge regression at each lambda

```
#glmnet run ridge and lasso
install.packages('glmnet')
#plotting
install.packages('plotmo')
library(glmnet)
library(plotmo)
#specify a range of possible value
lambda for testing
grid <- 10^seq(6, -3, length=10)
grid
#run ridge at each lambda; #Alpha=
Ridge model;Alpha =1: Lasso

ridge.mod <- glmnet(scale(x), y,
alpha=0, lambda=grid, thresh=1e-2,
standardize = TRUE)

# Plot each coefficient over different
values of lambda; label specifies the
number of coefficient labels to display

plot_glmnet(ridge.mod, xvar = "lambda",
label = 4)
```



Log base is 10

When $\lambda = 10^5$, Log (λ)=5

24

R: Running Ridge and K-fold CV for choosing λ* using Credit data

25

- Run K-fold CV (10-fold here) for choosing λ*

```
cv.out <- cv.glmnet(scale(x), y, alpha=0, nfolds = 10)
cv.out
```

```
Measure: Mean-Squared Error

    Lambda Measure    SE Nonzero
min  39.66   14054 516.6      11
1se  39.66   14054 516.6      11
```

Lamda.min: value of lambda that gives minimum "cvm" (the mean cv error)
Lamda.1se: largest value of lamda such that error is within 1 standard error of the minimum mse

The optimal λ* = 39.66

Tips: type "help ( )": e.g. help(cv.glmnet) to get the help documents. type glmnet, you will see the source code for this function. Do the same thing for any function you want to check

25

R: Running Ridge and K-fold CV for choosing λ* using Credit data

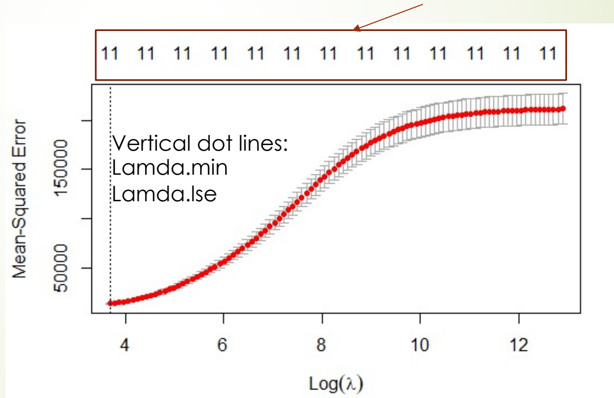
26

- Plot the MSE for each lambda value

```
plot(cv.out)

best.lambda <-
cv.out$lambda.min

best.lambda

[1] 39.65627
```

The optimal λ* = 39.66

How many predictors stay at each λ



26

27

R: Running Ridge and K-fold CV for choosing λ* using Credit data

- Use best λ* to run ridge on the entire data and evaluate using MSE/RMSE

```
ridge.final <- glmnet(scale(x), y, alpha=0,
lambda=best.lambda, thresh=1e-2, standardize = TRUE)
predict(ridge.final, type="coefficients",
s=best.lambda)
```

```
12 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)        520.015000
Income            -179.068431
Limit              386.880531
Rating             142.012772
Cards               26.339482
Age                -17.675691
Education           -1.690490
Gender Male          2.423412
StudentYes         115.165538
MarriedYes          -5.664827
EthnicityAsian       5.948937
EthnicityCaucasian   5.011724
```

```
ridge.pred <- predict(ridge.final, s=best.lambda, newx=scale(x))
print(paste('MSE:', mean((ridge.pred - y)^2)))
[1] "MSE: 13026.371868871"
print(paste('RMSE:', sqrt(mean((ridge.pred - y)^2))))
[1] "RMSE: 114.133132213529"
```

27

28

Final Ridge Model using the optimal λ*

```
(Intercept)          520.015000
Income              -179.068431
Limit                386.880531
Rating               142.012772
Cards                 26.339482
Age                  -17.675691
Education             -1.690490
Gender Male            2.423412
StudentYes           115.165538
MarriedYes            -5.664827
EthnicityAsian         5.948937
EthnicityCaucasian     5.011724
```

*Balance = 502.02 – 179.07*Income + 386.88*Limit + 142.01*Rating + 26.34*Cards - 17.68*Age - 1.69*Education + 2.42*Gender – 115.17*Student - 5.66*Married + 5.95*Asian + 5.01*Caucasian*

28

29

R: Running Lasso and K-fold CV for choosing λ* using Credit data

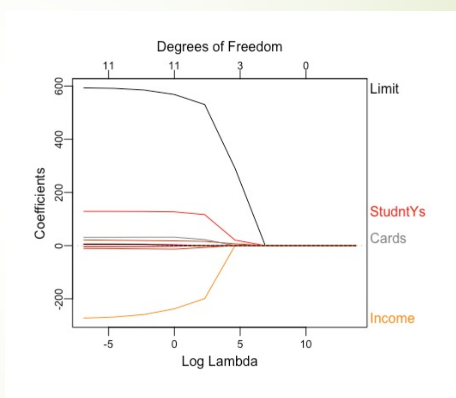
29

R: Running Lasso and K-fold CV for choosing λ* using Credit data

30

- Run Lasso regression using each lambda

```
# Create a grid of possible
values of lambda to test
grid <- 10^seq(6, -3,
        length=10)
# Create linear model
lasso.mod <- glmnet(scale(x),
             y, alpha=1,
             lambda=grid,
             thresh=1e-2,
             standardize =
             TRUE)
#alpha=1 is default: lasso
# Plot coefficient values at
different lambda
plot_glmnet(lasso.mod,
xvar="lambda", label = 4)
```



30

R: Running Lasso and K-fold CV for choosing λ* using Credit data

31

- Run K-fold CV (10-fold here) for choosing λ*

```
lasso.cv.out <- cv.glmnet(scale(x), y, alpha=1, nfolds = 10)
lasso.cv.out
```

```
Measure: Mean-Squared Error

    Lambda Measure    SE Nonzero
min  0.589   10045 535.4      11
1se  6.027   10506 457.2       6
```

The optimal λ* = 6.027

Recall:
Lamda.min: value of lambda that gives minimum “cvm” (the mean cv error)
Lamda.1se: largest value of lamda such that error is within 1 standard error of the minimum of MSE

31

R: Running Lasso and K-fold CV for choosing λ* using Credit data

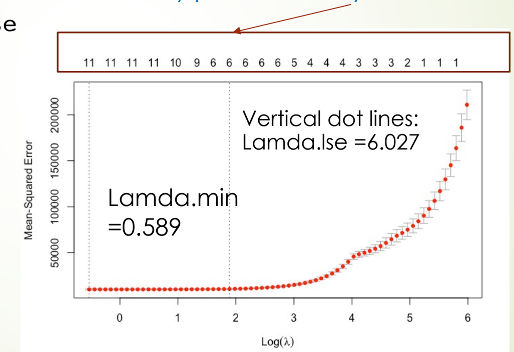
32

- Plot the MSE for each lambda value

```
plot(lasso.cv.out)

lasso.best.lambda <-
lasso.cv.out$lambda.1se
lasso.best.lambda

lasso.best.lambda
[1] 6.0274
```

The optimal λ* = 6.027

How many predictors stay at each λ



32

33

R: Running Lasso and K-fold CV for choosing λ* using Credit data

- Use best λ* to run ridge on the entire data

```
lasso.final <- glmnet(x, y, alpha=1, lambda=grid)
predict(lasso.final, type="coefficients", s=lasso.best.lambda )
```

```
12 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)         520.015000
Income             -201.186637
Limit               547.774004
Rating                4.306531
Cards                27.936428
Age                 -11.171937
Education             .
Gender Male           .
StudentYes          120.701548
MarriedYes            .
EthnicityAsian        .
EthnicityCaucasian    .
```

```
# Calculate MSE and RMSE for Lasso model with optimal lambda
lasso.pred <- predict(lasso.final, s= lasso.best.lambda,
newx=scale(x))
print(paste('MSE:', mean((lasso.pred - y)^2)))
[1] "MSE: 11701.0165977173"
print(paste('RMSE:', sqrt(mean((lasso.pred - y)^2))))
[1] "RMSE: 108.171237386458"
```

33

34

Final Lasso Model using the optimal λ*

```
(Intercept)         520.015000
Income             -201.186637
Limit               547.774004
Rating                4.306531
Cards                27.936428
Age                 -11.171937
Education             .
Gender Male           .
StudentYes          120.701548
MarriedYes            .
EthnicityAsian        .
EthnicityCaucasian    .
```

*Balance = 520.02 – 201.19 *Income + 547.77 *Limit +4.31*rating + 27.94*Cards-11.17*Age+120.70*Student*

34

35

Summary Table of Ridge and Lasso with CV choosing optimal λ*

Lasso reduces variance and performs variable selection

	Ridge Regression (α = 0)	Lasso Regression (α = 1)
min λ	39.656	0.589
1se λ (λ*)	39.656	6.027
MSE (at λ*)	13026.37	11701.02
RMSE (at λ*)	114.13	108.17

Note: for your projects, just report optimal lambda, or one of the two measures, MSE or RMSE.

35

36

Sectional Project 1: (35 points) Written Report, and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

36

37

Sectional Project 1: (35 points) Written Report, and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

Instruction:

Apply **multiple linear**, **Ridge** and **Lasso** **regression** for Boston Housing data

download at https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

Refer to Lecture slides, Reading assignments, and R Instruction Files for linear regression and regularized linear regression, complete the following:

1. Explore and describe Boston Housing data (ie. Attributes/predictors) using graphics, tables, and descriptive statistics, as appropriate
2. Identify Y and X for this dataset (Read carefully about the document: what Y is? What X are?). Please **name** Y and X (don't call them x1, x2…etc.)

37

38

Sectional Project 1: (35 points) Written Report, Code and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

Instruction continued:

3. Copy and paste your code and output in your Word document
4. Write the final estimated model in the format of, e.g., Y = Beta*X, for Boston Housing dataset, from each of the three models you applied:

   (a)multiple linear regression, (b)Ridge regression and (C) Lasso regression.

   Note clearly the actual names of these attributes for your Boston housing data in your model.

3. Report a summary table of your accuracy checking and cross-validation results, as appropriate. E.g., Create a summary table for MSE/RMSE, etc.. Refer to the summary table listed LS6 for multiple linear regression and LS8 for Ridge and Lasso.
4. Describe your Cross Validation algorithm for choosing your optimal regularization parameter, λ^* , in your **Lasso model**, for Boston Housing dataset
5. References: quote the citations you used for your project

   Note: all discussions and notes are in the context of Boston Housing data example.

38

39

Sectional Project 1: (35 points) Written Report, Code and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

Submission Instruction

- Submit **two** files:
  - Written report in Word or PDF format: No page limit; no template for sectional project report;

    requirement: make it nice and neat (note: Final project does have a template)

    -- Include your code and output.
  - PowerPoint slides for presentation (~8 slides)

39

40

Sectional Project 1: (35 points) Written Report, Code and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

Grading Rubric: total 35 points.

- Project written report (25 points: 15 points for the written report, 3points *5 items; and 10 points for coding): including coding and output; no page limit.
- Slide presentation (10 points): Each group has ~10 min to present and demo your project, so using ~ 8 slides.

  Suggestions: using graphs/images/tables/flowcharts for data description, models/algorithms illustration and comparison. Summarize and compare results from three methods.

  -- Highlight the specific techniques you applied and learnt from these lectures

  -- Highlight the part you are most proud of in this project

Zoom presentation: All group members are required to turn on videos when you are presenting, while other groups, please turn off your videos. Please turn on your videos if you have questions raised for the presenting group.

40

41

Sectional Project 1: (35 points) Written Report, Code and Presentation Slides Submission due Feb23; in-class presentation, Feb 24

On your last slide: Please show

Summary of your group meet time and duration

In person or Zoom:

Group meet time and duration (e.g., 5pm-7pm, Feb 1st):

Average time in communication and discussion regarding assigned group work (via email or other social media, e.g. What's app.):

Participants (Print and sign your names):

Contribution report:

If your team members contribute equally to this project, please make this statement "Each member contributes equally" on your last page, so that each of you will receive the same score.

If your team members do not contribute equally to this project, please note your team members' names, and mark the percentage of effort each member makes (e.g., Sukumar: 80% then if your group receives a project score of 30, then this member with 80% effort will only get 24).

Participants: Print and sign your names

41