

# CS 458/658:

# Introduction to Data Mining

## Data Mining Course Project

Instructor: Lei Yang

Department of Computer Science and  
Engineering, UNR – Fall 2018

Tue, Thur 12:00PM - 1:15PM, WRB 4050

# Project

- ◆ **One key goal** of this course is to take advantage of your intelligence and (limited) experience (so you're audacious and creative) to expand your knowledge in creating something useful and interesting
- ◆ **Group project**
  - 3 students per group
    - ◆ 27 undergraduates —9 groups
    - ◆ 9 graduates —3 groups
  - You can apply whatever techniques you learnt from data mining course and other sources

# Tasks



## Tasks

- Task 1: Document Classification
- Task 2: Exploring Environmental Data at NRDC
- Task 3: Exploring Used Auto Purchase Dataset
- Task 4: Exploring Intent data for B2B marketing and sales (TBD)

# Undergraduates

## ◆ Tasks:

- Task 1
- Pick one of the following:
  - ◆ Task 2
  - ◆ Task 3

## ◆ Bonus:

- ◆ One or two of the remaining tasks
- ◆ Other problems in Task 2

# Graduates

## ◆ Tasks

- Task 1
- Task 2
- Task 3

## ◆ Bonus:

- ◆ Task 4 ?
- ◆ Other problems in Task 2

# Evaluation

- ◆ Final report (due **Dec 12, 2018** in **Webcampus**)  
(**35%**)
  - Each member need to submit your own report and indicate your contribution in %
  
- ◆ Class presentation and/or demo (**5%**)
  - Each group will present their work. Each member needs to present.
  - Your presentation will be evaluated by the other groups using an evaluation form.
  - Each presentation is 20 mins with 5 mins for Q&A.
    - ◆ Nov. 27, 2018
    - ◆ Nov. 29, 2018
    - ◆ Dec. 4, 2018
    - ◆ Dec. 6, 2018

# Task 1: Classification

## ◆ Provided data

- The training set and its label information
- The testing set

## ◆ Hidden data

- The label information of the testing data
- The data will be used for the purpose of evaluation

# Data Format

## ◆ The training set

- training.txt
- The first column is the information ID
- The second column is the feature ID
- The third column is the value of the feature
- The default values of features are zeros

1	16	1
1	23	4
1	27	1
1	29	8
1	30	2
1	33	3
1	42	1
1	54	1
1	72	1
1	81	1



# Data Format

- ◆ The label information of the training set
  - label\_training.txt
  - Each row represents a data point in the training set
  - 1 is true information while -1 is misinformation

1  
-1  
1  
-1  
1  
1  
1  
-1  
-1

# Data Format

- ◆ The testing set
  - testing.txt
  - It has the same format as the training set

```
1 16 1
1 23 1
1 27 1
1 29 2
1 50 1
1 245 1
1 340 1
1 388 1
1 589 1
1 638 1
1 764 1
1 902 1
1 905 1
1 2774 1
1 8066 1
1 10762 2
```

# Model Challenge from Model Selection

- ◆ There are so many classifiers
  - Which one is better?
- ◆ There may be parameters in classifiers
  - How to determine the optimal values?

# Evaluation

- ◆ Classification accuracy will be used to evaluate the quality of the predicted labels
  - Comparing the hidden labels with your predicted labels
- ◆ Your final grades will strongly depend on the rankings of the quality of the predicted labels you provide



## Task 2: Exploring Environmental Data at NRDC

# Data

◆ <http://sensor.nevada.edu/SENSORDataSearch/>

# Sample project

- ◆ Wind speed prediction
  - Predict the quantitative wind speed at different sites using the historical information in the same sites as well as data in other neighbor sites.
- ◆ The data is available from <http://sensor.nevada.edu/SENSORDataSearch/>
  - Snake Range East Sagebrush (EB)
  - Snake Range East Subalpine (EA)
  - Snake Range East Salt Desert Shrub (ED)
  - Snake Range West Subalpine (WA)
  - Snake Range West Montane(WM)
  - Snake Range West Sagebrush (WB)
  - Snake Range West Pinyon-Juniper(WP)

10 minutes wind data

# Download data

Choose data

Choose a site

time

data type

download

Start Formatting

Cancel

Home Map View

Visualize Data... Download Data... Interface... Favorites... Reset

Start Time: 01/01/2013 00:00

End Time: 12/31/2013 16:48

Time Zone Preference: (UTC-08:00) Pacific Time (US & Canada)

Available Properties:

- Radiation: Photosynthetically active
- Radiation: Solar
- Relative humidity
- Temperature
- Wind direction
- Wind speed

Monitoring Hardware

Precipitation

Selected Properties:

Name	Units	System	Re
Wind speed	m/s	Atmosphere	

10m height

- Wind velocity monitor

Map

Download Data

Measurement Intervals:

- ☐ 1-minute
- ☒ 10-minute

Measurement Types:

- ☒ Average
- ☒ Minimum
- ☒ Maximum
- ☒ Standard deviation
- ☒ Resultant average

Aggregate Data Every:

1

Hour

Site Name	Latitude	Longitude	Altitude
Snake Range West Sagebrush	38.92535707774057	-114.4082747078973	1790.3952

Atmosphere - Wind speed

Name	Units	Interval	Type
<input checked="" type="checkbox"/> Wind velocity monitor	m/s	10-Minute	Average
<input type="checkbox"/> Wind velocity monitor	m/s	10-Minute	Minimum
<input type="checkbox"/> Wind velocity monitor	m/s	10-Minute	Maximum
<input type="checkbox"/> Wind velocity monitor	m/s	10-Minute	Standard deviation
<input type="checkbox"/> Wind velocity monitor	m/s	10-Minute	Resultant average

Start Formatting Cancel



# Example data

Site Name:,Snake Range West Sagebrush  
Deployment:,Wind velocity monitor  
Monitored System:,Atmosphere  
Measured Property:,Wind speed  
Vertical Offset from Surface:,10m height  
Units:,m/s  
Measurement Type:,Average  
Measurement interval:,00:10:00  
Time Stamp ((UTC-08:00) Pacific Time (US & Canada))  
1/1/2013 12:00:00 AM,0.513648960000000000  
1/1/2013 12:10:00 AM,0.073761600000000000  
1/1/2013 12:20:00 AM,0.348691200000000000  
1/1/2013 12:30:00 AM,0.291023040000000000  
1/1/2013 12:40:00 AM,0.476544640000000000  
1/1/2013 12:50:00 AM,0.435864000000000000  
1/1/2013 1:00:00 AM,1.001369600000000000  
1/1/2013 1:10:00 AM,0.899444480000000000  
1/1/2013 1:20:00 AM,0.206979520000000000  
1/1/2013 1:30:00 AM,0.604398080000000000  
1/1/2013 1:40:00 AM,0.710793600000000000  
1/1/2013 1:50:00 AM,0.430052480000000000  
1/1/2013 2:00:00 AM,0.198485760000000000  
1/1/2013 2:10:00 AM,0.175239680000000000  
1/1/2013 2:20:00 AM,0.598139520000000000  
1/1/2013 2:30:00 AM,1.322791360000000000  
1/1/2013 2:40:00 AM,0.473415360000000000  
1/1/2013 2:50:00 AM,0.105948480000000000  
1/1/2013 3:00:00 AM,0.760862080000000000

# Challenges

- ◆ Data preprocessing

- Missing data

- ◆ What methods to use?

- ◆ How to tune parameters?

- ◆ ...

# Evaluation

- ◆ For each site, you need to provide prediction accuracy of your proposed approach based on the following measure

- Mean absolute error (MAE)

$$MAE = \frac{1}{\text{number of points}} \sum |forecast - actual|$$

- Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{\text{number of points}} \sum |forecast - actual|^2}$$

- ◆ Compare your approach with the following Benchmark:

- Persistent forecast:  $\text{predicted\_wind}(t) = \text{actual\_wind}(t-1)$



# Task 3: Exploring Used Auto Purchase Dataset

# Exploring Used Auto Purchase Dataset (1)

- ◆ Dataset: the set of all used auto purchases for the past 5 years in US
  - Number of attributes: 280
    - ◆ Vehicle Info (Model, Engine, Drive Type) Home Info (Purchase Price/Date, Value, Year), Address (State, County), Loan Info (Monthly Mortgage), Demographic (Ethnic, Number of Children ), Behavior Info (Investment, Interest in Travel/Reading, Presence of Premium Credit Card) .....
    - ◆ No Personally identifiable information



**Acknowledgement:**  
Thanks Marketing  
Evolution for sharing  
this dataset.

# Dataset(UsedAutoRELEVATEfirst10000-noLatLong.csv)

	Attribute 1	Attribute 2	...
Data entry 1			
Data entry 2			
.....			

Customer	Home Purc	Latitude	Investmen	Living Unit	County	CarHome	Lanc	Acct#	Longitude	Home Lanc	Carrier Ro	Vehicle 1	THome	Impri	Area Code	Investmen	Home Tax	Home Year	Transactio	Home Base	Home Lan
XY63114H2	99		1.83E+09	339		0	4.59E+08			0 R028	SE		133				2596	1997	20160512	15	7
XY6311rueXSbZPo899lvW5wuBoN3T-WYHPKKzgT19qCoGAVw							4.46E+08			C022	LT								20161101		
UNMATCHED							4.11E+08			C003	BASE								20161205		
XY6311DIN	1050		1.04E+09	3		0	4.52E+08			0 C002	LX		845				9999	1988	20170130	0	503
XY6311mKnDv9rKvjd9pAFXNIWzGmq2ccDrAr0uAdwRHFAC							3.7E+08			C005	SLT								20160301		
XY6311eByCtpT12dTWfOruOgl9YaYR3WY42TJun6Yzk1hDZl							1.85E+08			C027	SLT								20151229		
XY6311ky5	85		2E+09	25		0	4.36E+08			0 R014	BASE		85	443			1372	1996	20161101	10	35
XY63115E0	0		1.93E+09	153		117	4.04E+08			65 C044	BASE		95	515			2994	1971	20161205	0	21

# Data Dictionaries (EXP REL Custom.xls)

ID	Field Name	Description	...
Data entry 1			
Data entry 2			
.....			

Relevate Gold Consumerview

Field Id	Field Name	Long Description	Start Position	End Position	Field Length	Field Type	Mask	Field Values
3415	Address ID	Address ID - Unique identifier assigned to each address in the Consumerview repository. The Address ID remains with an address even in the event that the occupants relocate. Values: 10 byte numeric.	1	10	10	AN	9509990099	
0337	State Code	State Code	11	12	2	AN	95	01=ALABAMA,02=ALASKA,04=ARIZONA,05=ARKANSAS,06=CALIFORNIA,08=COLORADO,09=CONNECTICUT,10=COLUMBIA,12=FLORIDA,13=GEORGIA,15=HAWAII,16=IDAHO,17=ILLINOIS,18=INDIANA,19=IOWA,21=KANSAS,22=LOUISIANA,23=MAINE,24=MARYLAND,25=MASSACHUSETTS,26=MICHIGAN,27=MINNESOTA,28=MISSISSIPPI,29=MISSOURI,30=NEBRASKA,31=NEVADA,32=NEW HAMPSHIRE,33=NEW JERSEY,34=NEW MEXICO,35=NEW YORK,36=NEW HAMPSHIRE,37=NORTH CAROLINA,38=OHIO,39=OKLAHOMA,40=OREGON,41=PENNSYLVANIA,42=RHODE ISLAND,43=SOUTH CAROLINA,44=TEXAS,45=UTAH,46=VERMONT,47=VIRGINIA,48=WASHINGTON,49=WISCONSIN,50=WYOMING
10114	State Abbreviation	State abbreviation	11	14	3	char		AK=ALASKA,AL=ALABAMA,AR=ARKANSAS,AZ=ARIZONA,CA=CALIFORNIA,CO=COLORADO,CT=CONNECTICUT,DE=DELAWARE,FL=FLORIDA,GA=GEORGIA,HI=HAWAII,IL=ILLINOIS,IN=INDIANA,IO=IOWA,KS=KANSAS,KY=KENTUCKY,LA=LOUISIANA,MA=MAINE,MD=MARYLAND,ME=MAINE,MI=MICHIGAN,MN=MINNESOTA,MO=MISSOURI,MS=MISSISSIPPI,MT=MONTANA,NC=NORTH CAROLINA,ND=NORTH DAKOTA,NE=NEBRASKA,NH=NEW HAMPSHIRE,NJ=NEW JERSEY,NM=NEW MEXICO,NV=NEVADA,NY=NEW YORK,OH=OHIO,OK=OKLAHOMA,OR=OREGON,PA=PENNSYLVANIA,RI=RHODE ISLAND,SC=SOUTH CAROLINA,SD=SOUTH DAKOTA,TN=TENNESSEE,TX=TEXAS,UT=UTAH,VA=VIRGINIA,VT=VERMONT,WA=WASHINGTON,WI=WISCONSIN,WY=WYOMING
10581	Zip Code	Zip Code	15	19	5	char		
10579	Zip+4	Zip+4	20	23	4	char		
13272	Delivery Point bar code	DirectDPV - Delivery Point barcode / Check digit	24	26	3	char		
13357	Carrier Route	carrier route code	27	30	4	char		
10217	WORKFLOW FIELD Short City Name to be Inverted V2	special 15byte field - tied to FCARD for 20 byte field	31	43	13	char		
10370	City Name	City name	44	71	28	char		
11247	House Number	Primary (house) number	72	81	10	char		
11249	Pre Direction	Street pre-directional	82	83	2	char		E=East,N=North,NE=Northeast,NW=Northwest,S=South,SE=Southeast,SW=Southwest,W=West
11023	Street Name	Street name	84	111	28	char		
10633	Street Suffix	Street suffix	112	115	4	char		ALY=ALLEY,ANX=ANEX,ARC=ARCADE,AVE=AVENUE,BCH=BEACH,BG=BURG,BLF=BLUFF,BUF=BLUFF,ANCH=ANCH,BRG=BRIDGE,BRK=BRICK,BRS=BRICKS,BTM=BOTTOM,BYP=BYPASS,BYU=BYWAY,CR=CLIFF,CMN=COMMON,CMNS=COMMONS,CON=CONCRETE,COR=CORNER,CORS=CORNERS,CP=CAMP,CPE=CAP,CRST=CRST,CSWY=CAUSEWAY,CT=COURT,CTR=CENTER,CTRS=CENTERS,CTS=COURTS,CURV=CURVE,DALE=DALE,DAV=DAVE,DR=DRIVE,DRS=DRIVES,DV=DIVIDE,EST=ESTATE,ESTS=ESTATES,EXPY=EXPRESS,ELD=ELD,FELDS=FIELDS,FLS=FLATS,FLT=FLAT,FLTS=FLATS,FRD=FORDE,FRG=FORGE,FRK=FORK,FRKS=FORKS,FREWAY=FREEWAY,GDN=GARDEN,GDMS=GARDENS,GLN=GLEN,GLNS=GLEN,GRN=GREEN,GRNS=GREENS,GP=GP,HBR=HARBOR,HBS=HARBORS,HL=HILL,HLS=HILLS,HOLI=HOLLOW,HTS=HEIGHTS,HVN=HAVEN,LE=LE,LS=LANDS,KT=KITCHEN,KNL=KNOLL,KNS=KNOLLS,KY=KEY,KYS=KEYS,LAN=LAND,LCK=LOCK,LGT=LIGHT,LGTS=LIGHTS,LK=LAKE,LKS=LAKES,LAN=LANE,LDNG=LANDING,LOOP=LOOP,MALL=MAIL,ML=NEWS,ML=MAIL,MLS=MILLS,MNR=MANOR,MNRS=MANORS,MNT=MOUNT,MSN=MISSION,MT=



# Exploring Used Auto Purchase Dataset (2)

## ◆ Project description:

- Selection: Due to the size/heterogeneity of the original data, we need to select a target data.
- Preprocessing: Data exist in many types (continuous, nominal) and forms, and may have missing values.
- Transformation: To better extract useful patterns from dataset.
- Data mining: Explore different data mining algorithms
- Interpretation/Evaluation

## ◆ Goal: extraction of useful patterns from dataset

- What car type will be purchased, given customer's info?
- What customer's type, given a car purchased?
- How to divide a market into distinct subsets of customers?



# Report Format

- ◆ Cover Page

- Team members and their contribution in %

- ◆ Introduction

- ◆ Literature review for each task

- ◆ Task 1

- Your approach (e.g., Preprocessing, Model selection, Parameter selection, Your solution)
  - Your conclusion

- ◆ Task 2

- ◆ Task 3

- ◆ Task 4

- ◆ List of documents/codes you submitted

# Report requirements

- ◆ The report should be as concise as possible while providing all necessary information required to replicate your plots.
- ◆ The plots should contain multiple curves and can be formatted so that many plots can fit on one page (so that your report is not longer than it should be).
  - **Don't use screenshot!**
- ◆ In literature review, you need to show your understanding of the literature by reading and comparing the existing work.
  - Cite your references properly. You can use google scholar to download citation.
- ◆ Your submitted code should have proper comments.