In-class activity: Competition based game for developing classification model

Fa2018 SBC Lab Session
June Young Park

http://nagy.caee.utexas.edu
@Z0ltanNagy

Objective

- Evaluate your progress of learning
- Review various classification problem
- Experience ‘real world’ setting
- Prepare mid-term exam part2

Procedure

1. Dataset description

2. Read dataset & Assign feature and target
3. Develop classification model
4. Evaluate and improve model

5. Save your classification model
6. Send your model via email

7. Fill a short survey question
8. Test models and announce results

Dataset description

Train set: 4,000 buildings

- Feature: various energy consumptions (26)
- Target: location information (3)

Test set: 1,686 buildings (hidden)

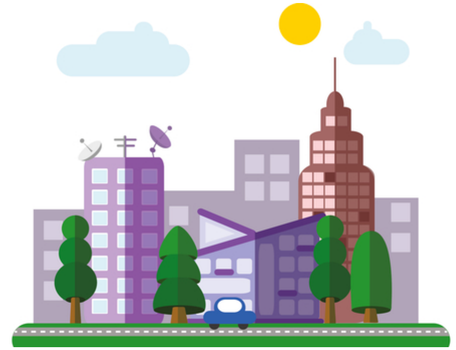
- Feature: various energy consumptions (26)
- Target: location information (3)

DOE RECS DATA
https://www.eia.gov/consumption/residential/

26 features: various electricity usage information

KWHSPH	Electricity usage for space heating, main and secondary, in kilowatthours, 2015
KWHCOL	Electricity usage for air conditioning (central systems and individual units), in kilowatthours, 2015
KWHWTH	Electricity usage for water heating, main and secondary, in kilowatthours, 2015
KWHRFG	Electricity usage for all refrigerators, in kilowatthours, 2015
KWHRFG1	Electricity usage for first refrigerators, in kilowatthours, 2015
KWHRFG2	Electricity usage for second refrigerators, in kilowatthours, 2015
KWHFRZ	Electricity usage for freezers, in kilowatthours, 2015
KWHCOK	Electricity usage for cooking (stoves, cooktops, and ovens), in kilowatthours, 2015
KWHMICRO	Electricity usage for microwaves, in kilowatthours, 2015
KWHCW	Electricity usage for clothes washers, in kilowatthours, 2015
KWHCDR	Electricity usage for clothes dryers, in kilowatthours, 2015
KWHDWH	Electricity usage for dishwashers, in kilowatthours, 2015
KWHLGT	Electricity usage for indoor and outdoor lighting, in kilowatthours, 2015
KWHTVREL	Electricity usage for all televisions and related peripherals, in kilowatthours, 2015
KWHTV1	Electricity usage for first televisions, in kilowatthours, 2015
KWHTV2	Electricity usage for second televisions, in kilowatthours, 2015
KWHAHUHEAT	Electricity usage for air handlers and boiler pumps used for heating, in kilowatthours, 2015
KWHAHUCOL	Electricity usage for air handlers used for cooling, in kilowatthours, 2015
KWHEVAPCOL	Electricity usage for evaporative coolers, in kilowatthours, 2015
KWHCFAN	Electricity usage for ceiling fans, in kilowatthours, 2015
KWHDHUM	Electricity usage for dehumidifiers, in kilowatthours, 2015
KWHHUM	Electricity usage for humidifiers, in kilowatthours, 2015
KWHPLPMP	Electricity usage for swimming pool pumps, in kilowatthours, 2015
KWHHTBPMP	Electricity usage for hot tub pumps, in kilowatthours, 2015
KWHHTBHEAT	Electricity usage for hot tub heaters, in kilowatthours, 2015
KWHNEC	Electricity usage for other devices and purposes not elsewhere classified, in kilowatthours, 2015

Target (3):



1(urban)

2(urban cluster)

3(rural)

Build your model using train set (30 minutes)

1. Read dataset & Assign feature and target
2. Develop classification model
3. Cross-validate and improve model

General process

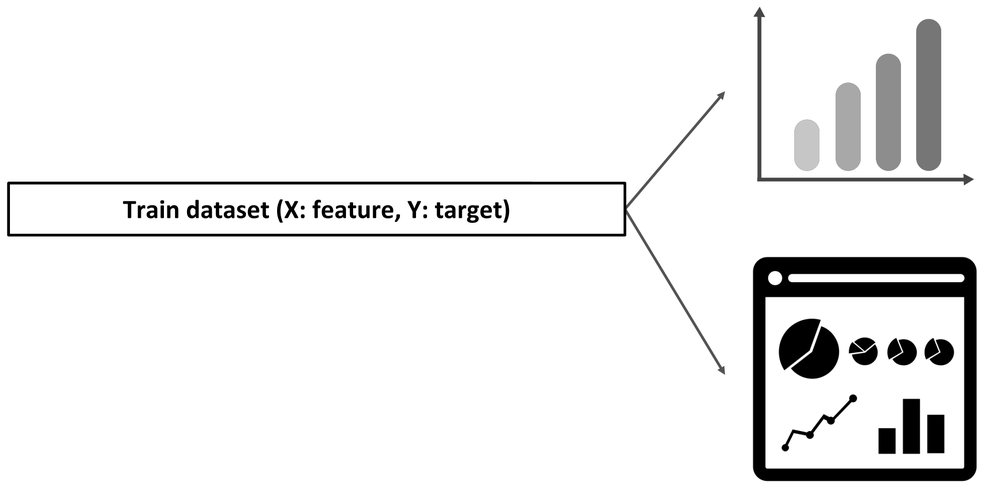
1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

4. Try various classification algorithms
5. Cross validate your model
6. Improve your model
   with feature engineering* & parameter tuning**
7. Cross validate your model again (x n times)

8. Read your test dataset
9. Evaluate your model performance

General process

1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

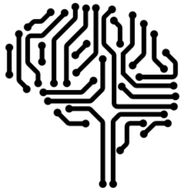


General process

1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

4. Try various classification algorithms
5. Cross validate your model
6. Improve your model
   with feature engineering* & parameter tuning**
7. Cross validate your model again (x n times)

Train dataset (X: feature, Y: target)

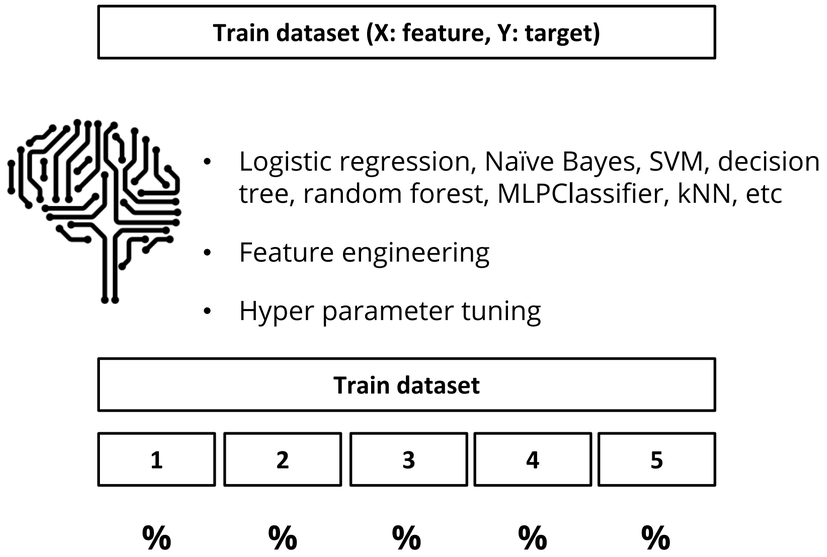


- Logistic regression, Naïve Bayes, SVM, decision tree, random forest, MLPClassifier, kNN, etc
- Feature engineering
- Hyper parameter tuning

General process

1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

4. Try various classification algorithms
5. Cross validate your model
6. Improve your model
   with feature engineering* & parameter tuning**
7. Cross validate your model again (x n times)



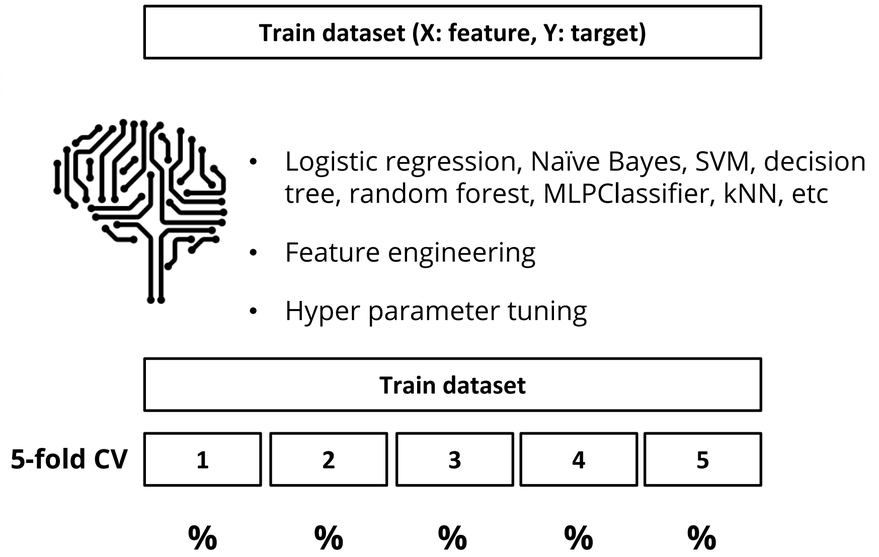
Cross validation (5 folds)

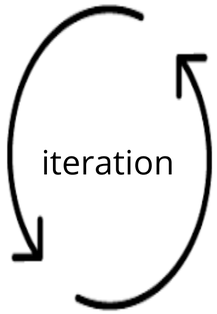
- Statistical method of evaluating generalization performance
- More stable and thorough than using a split into a training and a test set.
- Data is instead split repeatedly and multiple models are trained.
- Most commonly version: *k*-fold cross-validation, where k is a user-specified number, usually 5 or 10.

General process

1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

4. Try various classification algorithms
5. Cross validate your model
6. Improve your model
   with feature engineering* & parameter tuning**
7. Cross validate your model again (x n times)





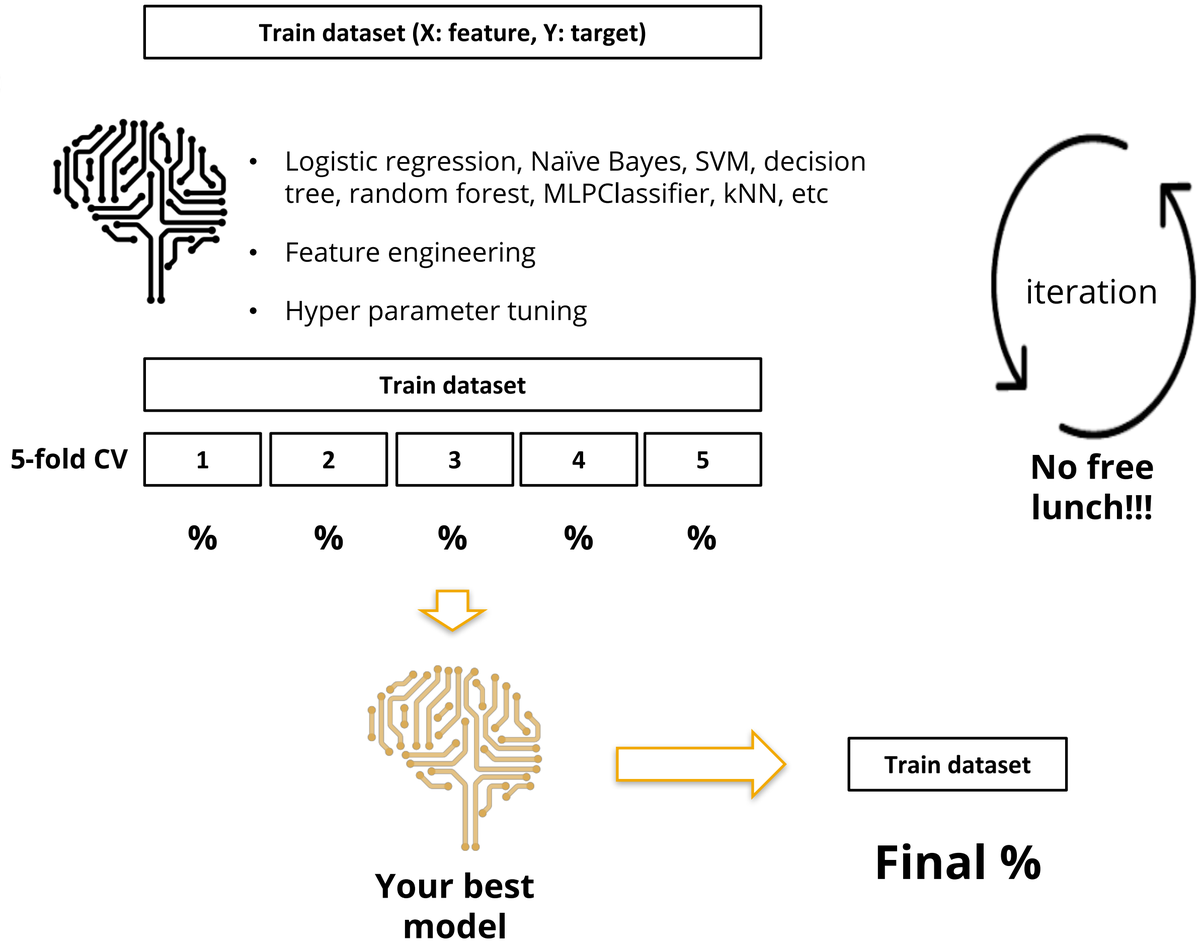
No free lunch!!!

General process

1. Read your train dataset
2. Explore your dataset
   (e.g., exploratory data analysis, descriptive statistics)
3. Assign features (X: input) and target (y: output)

4. Try various classification algorithms
5. Cross validate your model
6. Improve your model
   with feature engineering* & parameter tuning**
7. Cross validate your model again (x n times)

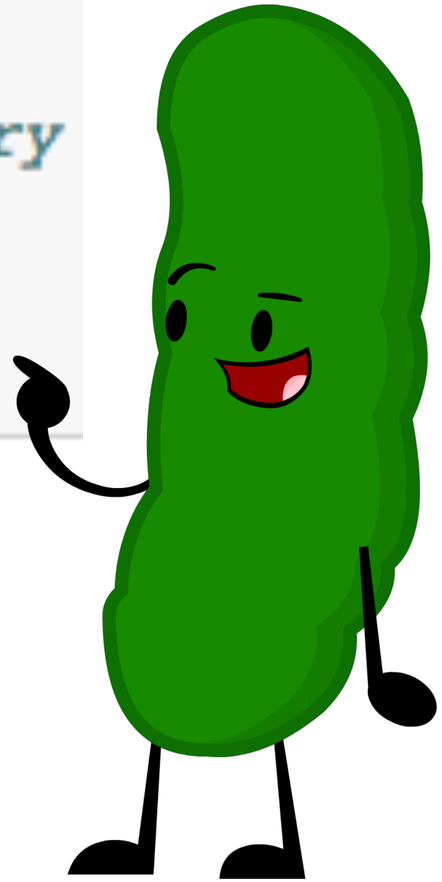
8. Read your test dataset
9. Evaluate your model performance



Save and send your model and jupyter notebook

```python
import pickle

# Save to file in the current working directory
pkl_filename = "model_team_1.pkl"
with open(pkl_filename, 'wb') as file:
    pickle.dump(model1, file)
```



Naming rule: “model_T**#_lastname1_lastname2**”

Please send .pkl and .ipynb to
juneyoungpark@utexas.edu

Test your model with test set ……

Please fill out survey questionnaires ☺

Game result

Accuracy (%)	Group(s)	Detail?
[0,33.3]		
[33.3,66.6]		
[66.6,100]		
June: [67.8%] Average: [XX %]		