UTS

Machine Learning Foundations

Workshop 1: Introduction

Mingshan
June 17, 2020

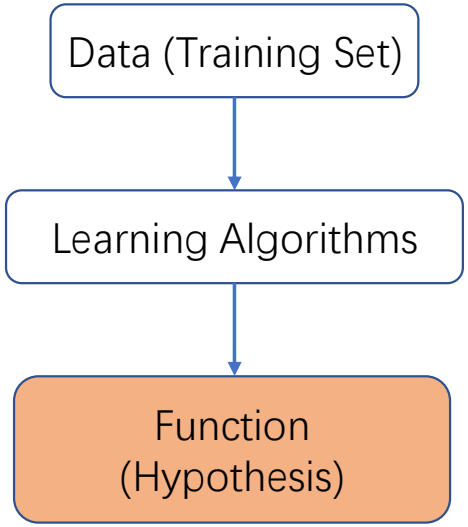
Table of Contents

What is Machine Learning?

- The field of study that gives computers the ability to learn without being explicitly programmed.

- **Learn a relationship (hypothesis / function) between features (input variable) and target (output variable) from experience (Training data).**

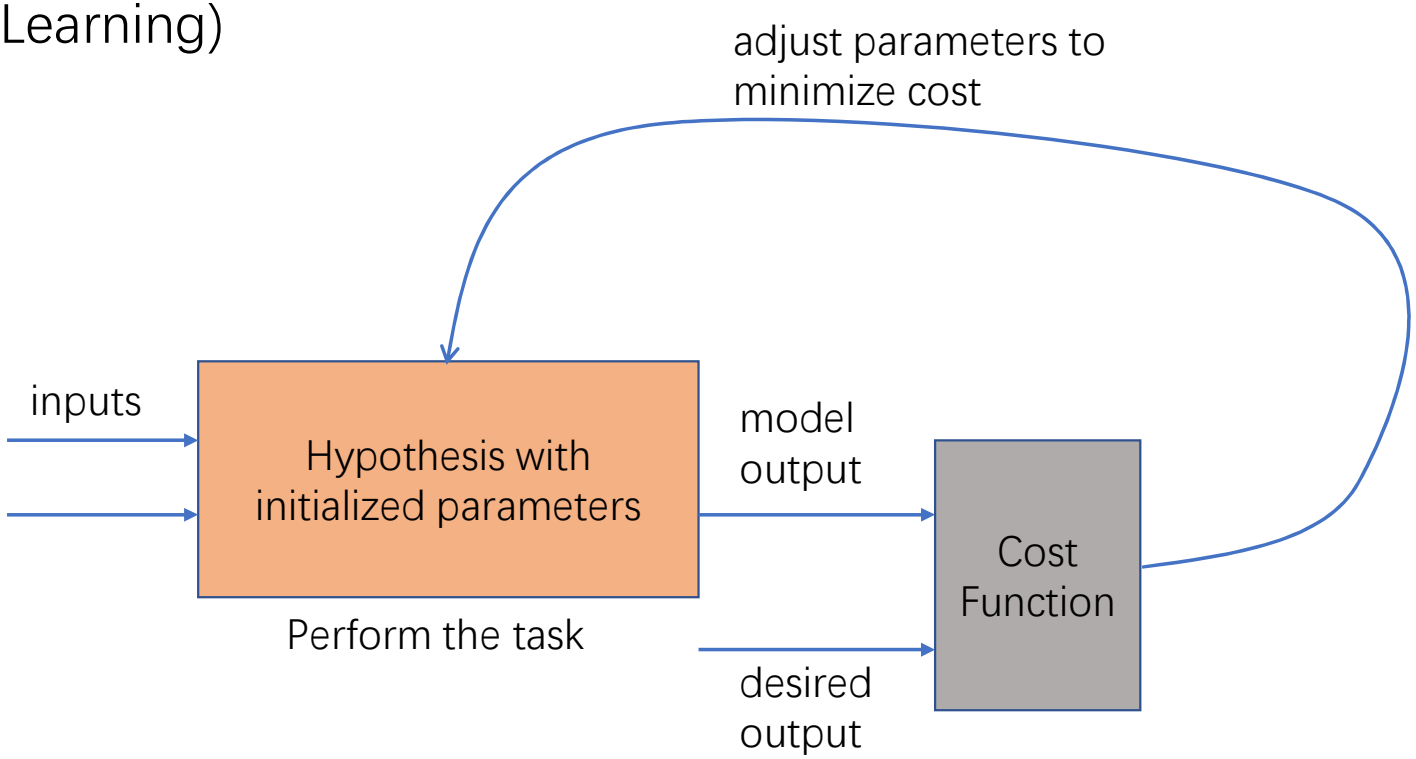
- **Data + Learning Algorithm (Hypothesis + Performance Measure)**



Non-Machine-Learning approach: Using math or expert knowledge to build the function explicitly.

Machine Learning approach: Using data and learning algorithm to learn the function.

Learning Process (Supervised Learning)



Data + Learning Model (Hypothesis + Performance Measure)

Machine Learning Applications

- Video recommendation, Product Recommendation
- Playing board game (Chess, Go), Playing video games.
- Traffic control
- Epidemic control
- Self-driving car
- Medical diagnosis

...

Machine Learning Applications in Telecom

- Discussion

When to use Machine Learning?

- Task is difficult to solve analytically
- Data is available
- Data is relevant to the task

Predictive Maintenance

Suppose you would like to predict the amount of maintenance of a particular type of equipment for the next quarter, in order to arrange your maintenance team.

- [ ] Data is the price of the equipment
- [ ] Data is the maintenance record of the equipment
- [ ] This problem is easy to solve analytically
- [ ] This is a Machine Learning problem

Machine Learning Algorithms

- **Supervised Learning** (labeled data)
  - **Linear regression**
  - **Logistic regression**
  - Neural Network
  - Support Vector Machines (SVM)
  - **Decision Tree**
  - Random Forest (Bagging strategy)
  - Adaboost (Boosting strategy)
- **Unsupervised Learning** (unlabeled data)
  - K-Means (Clustering)
  - Principle Component Analysis (PCA) (Dimensionality reduction)
- Federated Learning (decentralized data)
- Reinforcement Learning (sequential data)

Tools

- With installation
  - Anaconda + Jupyter Notebook

    Installer: https://www.anaconda.com/products/individual

    Installation instruction: https://docs.anaconda.com/anaconda/install/

- Without installation (Jupyter notebooks in the browser)
  - Google Colab: https://colab.research.google.com/

    Introduction: https://colab.research.google.com/notebooks/intro.ipynb#

    Features: https://colab.research.google.com/notebooks/basic_features_overview.ipynb

  - Kaggle Notebook: https://www.kaggle.com/notebooks

Python & Libraries

New to Python:

DataCamp Introduction to Python:
https://www.datacamp.com/courses/intro-to-python-for-data-science

- Numpy
- Scipy
- Scikit-learn (sklearn)
- Pandas
- Matplotlib
- Pytorch

...

Data

Dataset : House in Ames, Iowa

A table of 1460 rows and 81 columns

What is the target?

How many features (attributes) ?

How many training examples?

```
Index(['Id', 'MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street',
       'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig',
       'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType',
       'HouseStyle', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd',
       'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType',
       'MasVnrArea', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual',
       'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1',
       'BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', 'Heating',
       'HeatingQC', 'CentralAir', 'Electrical', '1stFlrSF', '2ndFlrSF',
       'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
       'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual',
       'TotRmsAbvGrd', 'Functional', 'Fireplaces', 'FireplaceQu', 'GarageType',
       'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual',
       'GarageCond', 'PavedDrive', 'WoodDeckSF', 'OpenPorchSF',
       'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'PoolQC',
       'Fence', 'MiscFeature', 'MiscVal', 'MoSold', 'YrSold', 'SaleType',
       'SaleCondition', 'SalePrice'],
      dtype='object')
(1460, 81)
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	...	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	...	0	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	...	0	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	0	NaN	MnPrv	Shed	700	10	2009	WD	Normal	143000
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	8	2007	WD	Normal	307000
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	0	NaN	NaN	Shed	350	11	2009	WD	Normal	200000
8	9	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	4	2008	WD	Abnorml	129900
9	10	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	Corner	...	0	0	NaN	NaN	NaN	0	1	2008	WD	Normal	118000

10 rows × 81 columns

$$\left(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_{79}^{(1)}, y^{(1)}\right)$$

$$\left(x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \dots, x_{79}^{(2)}, y^{(2)}\right)$$

$$\vdots$$

$$\left(x_1^{(1460)}, x_2^{(1460)}, x_3^{(1460)}, \dots, x_{79}^{(1460)}, y^{(1460)}\right)$$

$x_j^{(i)}$ the value of feature j in i^{th} training example

m = number of training examples

n = number of features

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	...	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	...	0	0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	...	0	0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	Inside	...	0	0	NaN	MnPrv	Shed	700	10	2009	WD	Normal	143000
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	8	2007	WD	Normal	307000
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	Corner	...	0	0	NaN	NaN	Shed	350	11	2009	WD	Normal	200000
8	9	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	Inside	...	0	0	NaN	NaN	NaN	0	4	2008	WD	Abnorml	129900
9	10	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	Corner	...	0	0	NaN	NaN	NaN	0	1	2008	WD	Normal	118000

10 rows × 81 columns

In the dataset above, what is $x_{77}^{(8)}$?

- [ ] YrSold 2007
- [ ] MoSold 8
- [ ] YrSold 2009
- [ ] MoSold 11

Q & A (Week 1)

Training, validation and test dataset

- **Training set**: a majority of data examples that used to train the model. (Experiences, what we learn from)

- **Test set**: a part of data examples used to evaluate the model.
  Whether the model perform well with unseen data ?

- **Validation set**: a part of data examples used to tune hyperparameters

- A toy example (univariate linear regression)

  $h(x) = \theta_0 + \theta_1 x$