SGCODE CAMPUS

Machine learning on AWS

AWS Educate

COURSE OBJECTIVES

- Understand the concepts of machine learning
- Lean how to solve regression and classification problems via linear model examples
- Be able to train, select, validate and evaluate the linear models
- Be able to make use of AWS sagemaker to train and deploy your learning model and make predictions

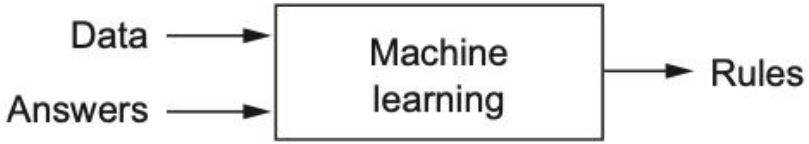
SGCODE CAMPUS

Prerequisite

- Statistics: expected value, variance, normal distribution
- Coding: python, Jupyterlab, pandas, numpy
- AWS: AWS console, S3

Machine Learning Introduction

- Machine learning algorithms learn from data to find hidden relations, to make predictions, to interact with the world, …



- Data is the key to generate rules and it decides how predictive your model is

SGCODE CAMPUS

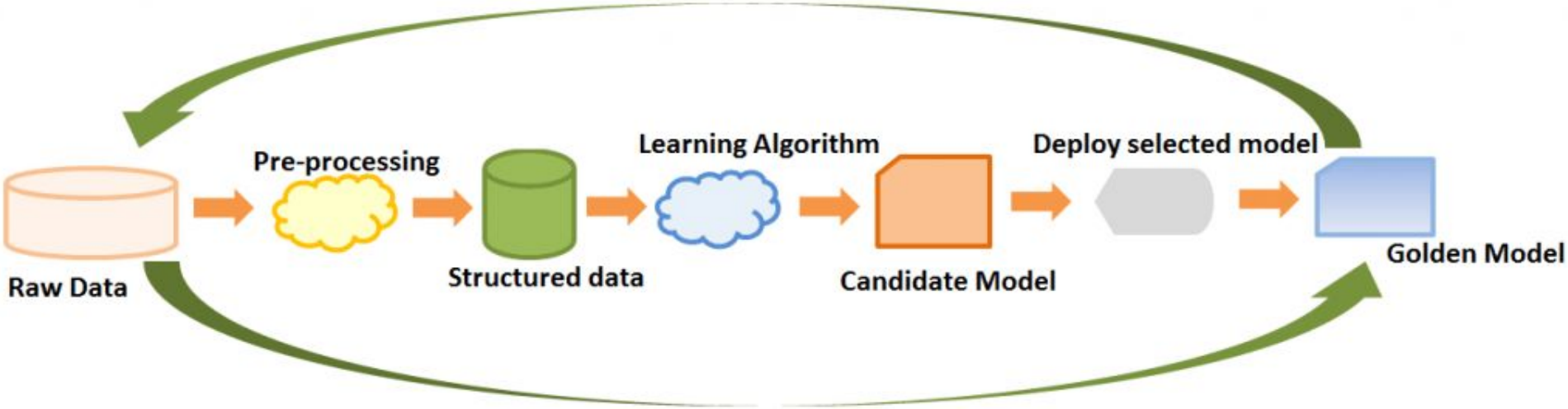
Machine Learning Categories

- Supervised Learning
  - Learn through examples of which we know the desired output (what we want to predict).
  - Classification problem, regression problem
- Unsupervised Learning
  - There is no desired output. Learn something about the data. Latent relationships.
  - Clustering problem
- Reinforcement Learning
  - An agent interacts with an environment and watches the result of the interaction.
  - AlphaGo

Machine Learning Algorithms

- Supervised
  - • Linear classifier • Naive Bayes • Support Vector Machines (SVM) • Decision Tree • Random Forests • k-Nearest Neighbors • Neural Networks (Deep learning)
- Unsupervised
  - • PCA • t-SNE • k-means • DBSCAN
- Reinforcement
  - • SARSA–λ • Q-Learning

Machine Learning Process



Machine Learning Process

- Data preprocess: check and clean the dataset
  - Missing data
  - Outliers
  - Bias (Classification problem)
- Feature engineering:
  - Label and convert data into model input

Scrubbing Data - Missing Data

- Sometimes, we have missing data
  - E.g. Customers leaving survey questions blank
- Missing values usually fall into these categories
  - Not applicable (e.g. Do you have a pet?)
  - Unknown (there is an answer, but you didn't collect it) (e.g. What is your age?)
  - Missing (someone forgot to type it in, or accidentally deleted it!)
- Understanding which of these would help us fill in the data!

SGCODE CAMPUS

Scrubbing Data - Missing Data

- Fill with 0
  - If the data Not applicable
- Fill with mean, median, or random
  - Data is unknown
  - But we know that the feature has some impact on the model
- Fill forward then fill backwards
  - For time series data to prevent lookahead bias

Scrubbing Data - Dropping Data

- Irrelevant Features
  - E.g. Phone number versus customer segmentation
- Missing too much data in a feature
  - Above 20 - 30% of data missing

Scrubbing Data - Dealing with Categorical Data

- Categorical / Qualitative
  - Male / Female
  - Membership tier (Free, Pro, Pro Plus, Enterprise)
- Binning Quantitative Data
  - Age vs Age groups
    - 0 - 20, 20 - 30, 30 - 40, 40 - 50, 50 - 60, >60
  - Salary
    - 0 - 20k, 20 - 30k, ...

SGCODE CAMPUS

Scrubbing Data - One-Hot Encoding for Categorical Data

- Categorical with no clear ordinal (ordered) relationship
  - Example: Type of car
    - Can we say that Honda < Toyota < Hyundai < Mercedes?
      - It could be that it reflects the purchasing power, but does it reflect spending habits?
    - Can we say that Sunday > Mon > … > Fri > Sat?
      - It could be that Fri, Sat, and Sun are better for sales, but what about Mon - Thu?

Scrubbing Data - One-Hot Encoding for Categorical Data

- Car Example

Y	...	has_car_brand
		Toyota
		Hyundai
		Hyundai
		Honda
		Mercedes

→

Y	...	has_toyota	has_hyundai	has_honda	has_mercedes
		1	0	0	0
		0	1	0	0
		0	1	0	0
		0	0	1	0
		0	0	0	1

SGCODE CAMPUS

Machine Learning Process

- Model training and hyperparameter tuning
  - Choose the right model
  - This is an iterative process to gain incremental improvement
  - Aim for the best prediction results
  - Validation dataset are used for hyperparameter tuning
- Model deployment and prediction
  - Save model as model file
  - May deploy to cloud or save to local

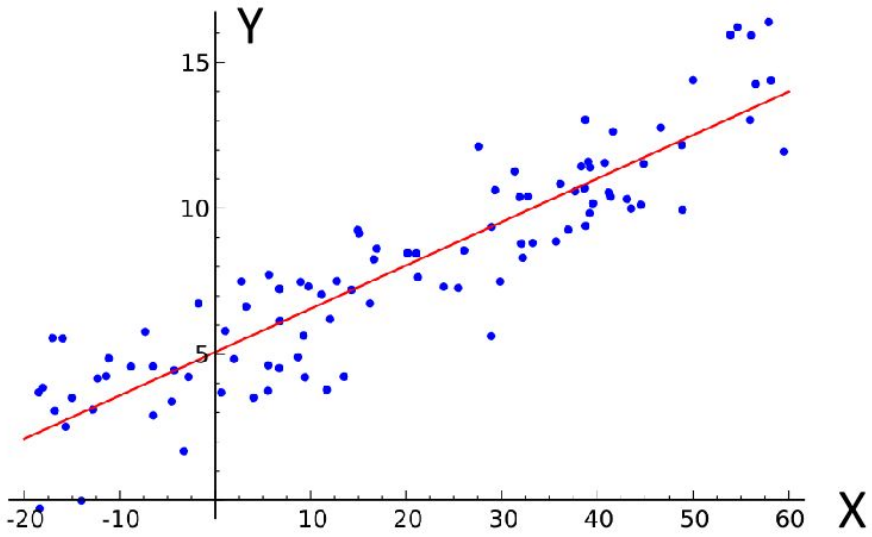
Machine Learning Process

- Examples of training models
  - Regression model
    - Linear regression
    - Model evaluation: feature selection, p value
    - Error analysis: MSE, R-square
  - Classification model
    - Logistic regression
    - Error analysis: AUC score, confusion matrix

Linear Regression

- Basic supervised learning model
- Assuming the relation of output and features is linear
- One of the most widely used techniques
- Fundamental to many larger models
- Easy to interpret
- Efficient to solve

Linear Regression



Response covariate noise

$$Y = \beta_1 X + \beta_0 + \epsilon$$

slope intercept

SGCODE CAMPUS

Linear Regression

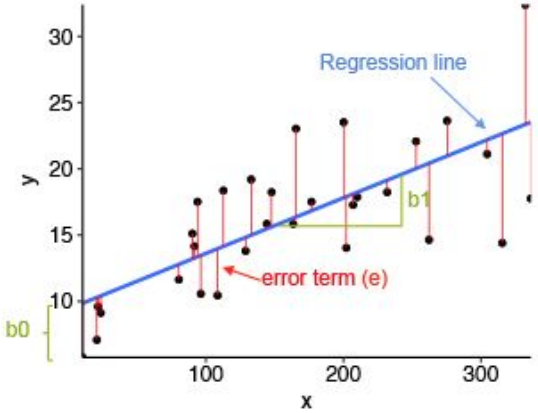
- Consider the following representation:

$$y = \sum_{i=1}^{k} x_i \beta_i + \beta_0 + \epsilon$$

- In other words, we say target variable y is only dependent on some linear combination of the the input variables, x_i.
- This is also called a Linear Regression Model.
- We expect that there will be some error to modelling the problem.
- This error is denoted by ϵ, which allows us to gauge how well the model fits data.

Linear Predictors contd.

- If we set $x_0 = 1$, then: $y = \sum_{i=0}^{k} x_i \beta_i + \epsilon$
- For the linear model prediction: $y^{pred} = \sum_{i=0}^{k} x_i \beta_i$
- We are in fact using the regression model to predict the $E(y)$
- This model is geometrically represented by a straight line (of best fit):



SGCODE CAMPUS

Objective Functions

- Need to find the best estimates of β from training dataset of size n
- The function should minimize the error of the predicted value

$$\min_{\beta} \sum_{j=0}^{n} \left(y_i - y_i^{pred}\right)^2$$

- It is solving for a linear system:

$$\sum_{l=0}^{k} \beta_l \sum_{j=0}^{n} x_{ji} = \sum_{j=0}^{n} x_{ji} y_j \; \forall i = 1, \ldots, k$$

or in matrix form $X^T X\beta = X^T Y$

Solution in the form $\beta = (X^T X)^{-1} X^T Y$

- The solution is also called Maximum Likelihood Estimator (MLE)

Source: Paarth Bir

Choose the best model

- Feature selection and model reduction
  - Systematic way of deleting irrelevant features
  - In the model summary, looking at the p-value of each feature
  - If p-value large then the feature is most likely irrelevant
- Cross validation of the model
  - To avoid overfitting problem of the model
  - Split the dataset into training and testing
  - K-fold cross validation

Model evaluation

- Mean squared error (MSE): $\frac{\sum_{j=0}^{m}(y_i - y_i^{pred})^2}{m}$
- Root mean squared error (RMSE): square root of MSE
- R-squared: closer to 1 means better fit:

$$R^2 = 1 - \frac{\sum_{j=0}^{m}(y_i - \bar{y})^2}{\sum_{j=0}^{m}(y_i^{pred} - \bar{y})^2}$$

- Adjusted R-squared: $1 - (1 - R^2)\frac{m-1}{m-1-k}$

SGCODE CAMPUS

Regression vs Classification

- What if the target variable is a discrete value?
- Binary valued? E.g: y is a binary value {0, 1}
- Multiple Classes?
- The Linear Model is still valid, but we need a transformation
- Transformations are usually performed via an *activation* function.
- This is called a **Linear Classifier Model**.

Logistic Regression

- A basic model for linear classification problem:
  - Still assuming linear relation $y = \sum_{i=0}^{k} \beta_i x_i$
- Use a sigmoid function to convert value of y from [-inf, inf] to [0,1] representing the probability of y being 1

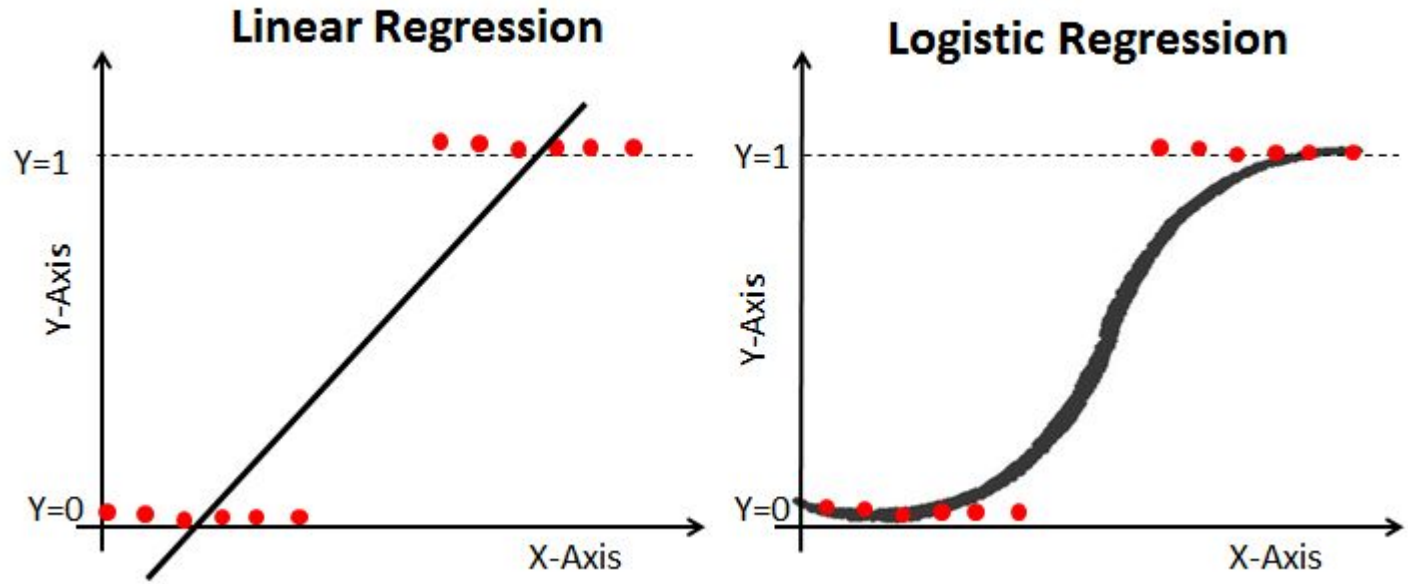
$$y^{pred} = sigmoid(y) = \frac{1}{1 + e^{-y}}$$

- We can then set a cut of probability $\tilde{p}$ so that

$$y_{class} = \begin{cases} 1 \; if \; y^{pred} \geq \tilde{p} \\ 0 \; if \; y^{pred} < \tilde{p} \end{cases}$$

SGCODE CAMPUS

Logistic Regression Vs Linear Regression



SGCODE CAMPUS

Data bias problem

- Sometimes the data is heavily imbalanced
  - This problem only occurs in the classification problems
  - y =1 is much more than y=0
  - This will cause the model been weak in predicting y=0
  - SMOTE algorithm(Synthetic Minority Oversampling Technique) is used to oversample the data in the lower class

Model evaluation

- Feature selection and Cross validation also applies
- Evaluation metrics:
  - Accuracy: number of correct classification over the total predictions
  - AUC: the area under curve (will be illustrated with examples)
  - Confusion matrix:

	Predict Positive	Predict Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Confusion matrix

AUC score

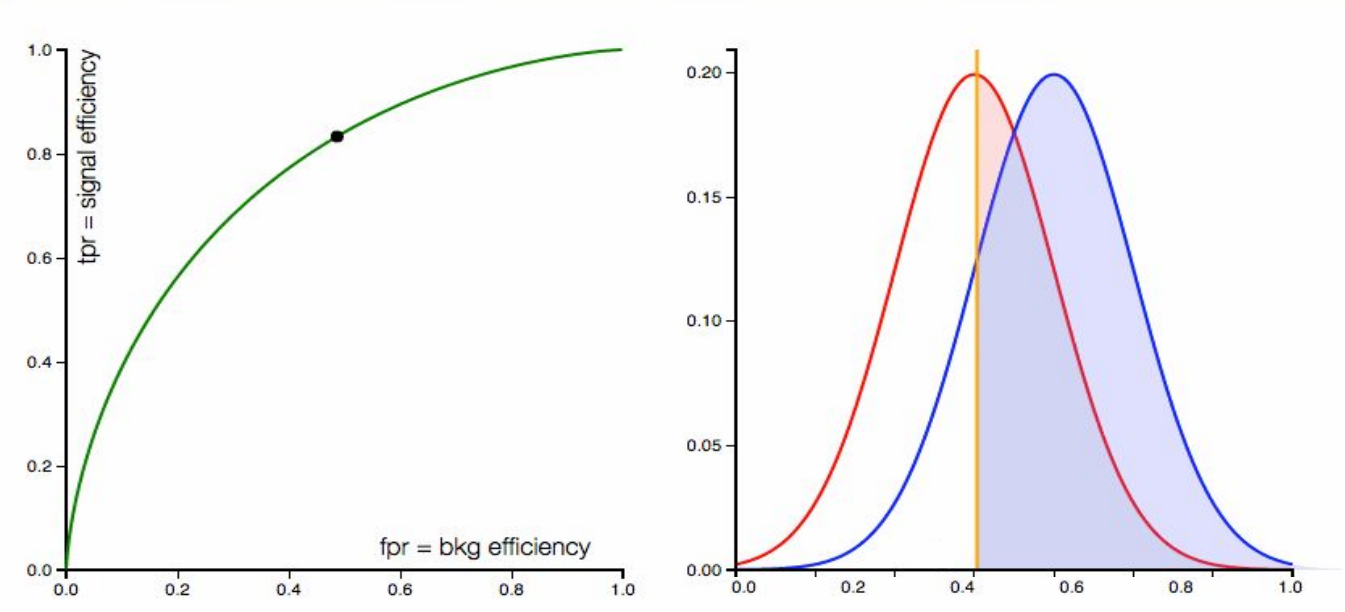
$$FalsePositiveRate = \frac{FalsePositive}{TrueNegative + FalsePositive}$$

TPR = 1 and
FPR = 1

	Predict Positive	Predict Negative
Actual Positive	True Positive	0
Actual Negative	False Positive	0

SGCODE CAMPUS

AUC score

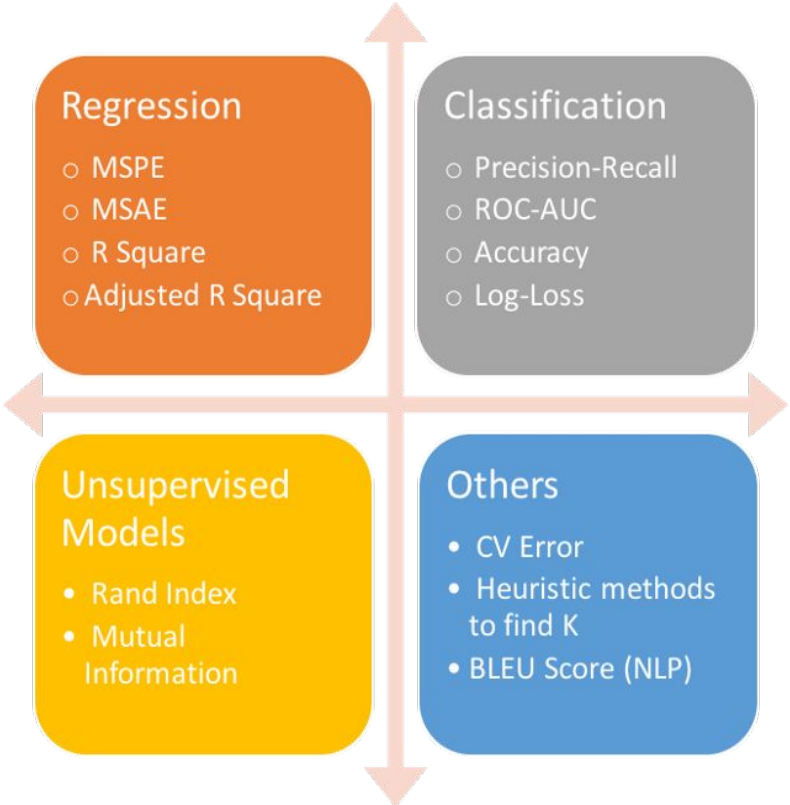


SGCODE CAMPUS

Multi Class Classification

- Reduce the problem to binary classification via to ways: [0, 1, 2]
  - One Vs rest (ovr) method: Model for 0 Vs no 0s
    - Model for 1 Vs no 1s
    - Model for 2 Vs no 2s
  - One vs one (ovo) method
    - Model for 1 Vs 2
    - Model for 0 Vs 1
    - Model for 0 Vs 2
  - the final prediction depends on the class with the highest score

Model metrics



SGCODE CAMPUS

Machine Learning on AWS

- For ML examples
  - ML examples
- For references of sagemaker:
  - Introduction to sagemaker
  - Examples sagemaker

SGCODE CAMPUS

Machine Learning on AWS

- Step 1: Data preparation
  - prepare the input data (training data, validation data, testing data)
  - Different ml models have different data input format
  - list of input format
  - reference for checking different input format
- Step 2:
- model training
  - Specify the container or use sagemaker build-in package
  - Hyperparameter setting
  - Model fitting
- Step 3: model deployment and validation
  - Can use deploy or create_end_point

SGCODE CAMPUS

Summary

- Machine learning essentials:
  - Supervised, unsupervised, reinforcement
  - Machine learning process
  - Data scrubbing
- Learned linear regression and logistic regression model for dealing with regression and classification problems
  - Examples of boston house price and direct marketing campaigns
- model training and deployment with AWS sagemaker
  - Converting the data input
  - Specifying the input and output uri
  - Creating endpoint
  - Invoke endpoint to predict

SGCODE CAMPUS