

Machine Learning

100 QUESTIONS

THOMPSON WONG

Introduction

MACHINE LEARNING

THOMPSON WONG

© Thompson Wong

This book is written during the learning of machine learning, it summarized the frequent questions asked by junior or the major points that make people confuse. The original plan is 100 questions, it may be adjusted depends on the practice. Hope it will help you in machine learning field. Enjoy !!!

ABOUT AUTHOR

Thompson Wong

Data Scientists | Machine Learning | Startup

<https://kinshing0930.wixsite.com/portfolio>

Thompson, experienced with data analysis such as equipment optimization, fault detection and diagnosis, performance evaluation; software development includes analytics framework development, machine learning pipelines and application, RESTful API, optimization algorithm design; research and development for next generation technology; dashboard solution using Grafana.

Benefiting from the various working experience in different scale company, a solid experience and good soft skill is developed such as engineering mindset for problem solving, design thinking for the product development, efficiency and effectiveness in software features delivery, leadership to coordinate the works from different stakeholders, communication between client request and team capability.

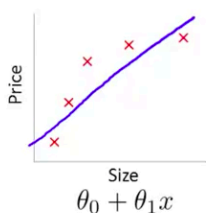
Communication to digest the needs, design thinking to make a innovation and breakthrough, engineering to solve problem step by step, leadership to coordinate the colleagues and make everyone satisfied in their job assignment.

Question 1	6
Bias and variance	
Question 2	7
Regularization	
Question 3	8
Data-centric or model-centric?	
Question 4	9
Model-based vs instance-based learning	
Question 5	10
Differences between rf, xgboost, lightgbm, GBDT	
Question 6	11
Ensemble Learning	
Question 7	12
Random Forest	
Question 8	13
Gradient boosting decision trees	
Question 9	14
XGBoost	
Question 10	15
LightGBM	

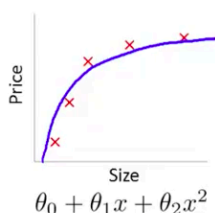
QUESTION 1

BIAS AND VARIANCE

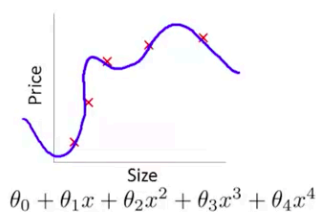
Bias and variance is a basic concept to illustrate the machine learning model performance, like overfitting and underfitting.



High bias
(underfit)
 $d=1$



"Just right"
 $d=2$



High variance
(overfit)
 $d=4$

The bias usually represent the underfitting, as insufficient features makes prediction bias in training set. The variance usually represent the overfitting, too much features make machine learning model only perform well in training set, and get fail in newly unseen dataset or test set or validation set. You can comprehend as not generic model.

QUESTION 2

REGULARIZATION

Regularization is an effective things that control the model fitting performance. It is divided by L1 and L2 regularization. The formula is generally written in:

L1 Regularization

$$\text{Cost} = \sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2 + \lambda \sum_{j=0}^M |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^N (y_i - \sum_{j=0}^M x_{ij} W_j)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=0}^M W_j^2}_{\text{Regularization Term}}$$

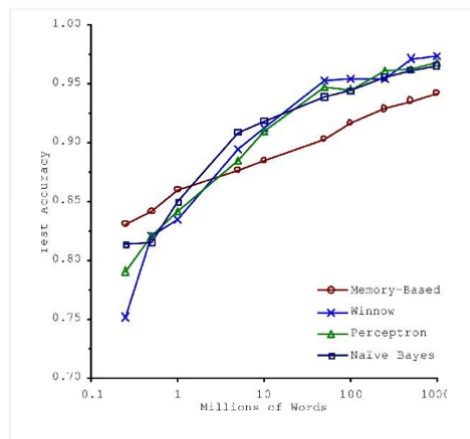
The L1 regularization is usually used in feature selection, L2 regularization is usually used in avoiding overfitting in machine learning. L1 regularization reduce model complexity via feature selection (reducing coefficients parameters/features amount), L2 regularization reduce complexity by controlling coefficients parameters values.

QUESTION 3

DATA-CENTRIC OR MODEL-CENTRIC?

Data quality is always the major factors affecting the machine learning model performance. Garbage in, garbage out. Spending time for controlling data quality is much better than model algorithm. The existing algorithm is good enough, we all standing on the giant's shoulders.

When you have good enough data quality and quantity, the modelling result will be close to each others in different machine learning algorithms. So don't worry on selecting algorithm or developing new algorithm. Focus on improving data quality, Data-centric AI is greater than model-centric AI. The below figure illustrate the relationship between data quantity, quality and its performance



QUESTION 4

MODEL-BASED VS INSTANCE-BASED LEARNING

Model-based usually known as supervised learning such as SVM (support vector machine), instance-based learning usually known as unsupervised learning such as kNN.

In model-based learning, once model is built, the training data can be discarded (e.g. learned weights and bias value in SVM: $y=mx+c$) while instance-based learning use the entire dataset as the model, and looks at the close neighbourhood of the input example in the space of feature vectors and outputs the label that it saw the most often in this close neighbourhood.

Fundamentally, the model-based and instance-based learning has no direct relationship with the terms of data-centric and model-centric. Don't get confused even the terms are similar.

QUESTION 5

DIFFERENCES BETWEEN RF, XGBOOST, LIGHTGBM, GBDT

XGBoost, lightGBM, GBDT, RF are attributed as **ensemble learning***, and boosting as major techniques to improve model performance while random forest adopts bagging method to improve performance.

Generally, the relationship is similar to the iteration of technologies. Therefore, RF -> GBDT -> XGBoost -> LightGBM. I shall create new question to detailedly explain each algorithms.

QUESTION 6

ENSEMBLE LEARNING

The purpose of ensemble learning is that combining multiple base-estimator's prediction result to improve single estimator's robustness and generalization.

The current ensemble learning is divided by two classes. One is strong relationship, serialization is required for result generation. Another one is no strong relationship between estimators, use parallel method to create multiple estimators. The representative of the former is Boosting, and the representative of the latter is Bagging.

Some terms can also be comprehended as ensemble learning as well such as blending, stacking. People may claim this is different. But no worries, it is the same. Just bear in mind is ok, no need to argue for a long time.

QUESTION 7

RANDOM FOREST

QUESTION 8

GRADIENT BOOSTING DECISION TREES

QUESTION 9

XGBOOST

QUESTION 10

LIGHTGBM

