Table of Contents 目录

- Example - Cross Validation for Model Selection 示例 - 模型选择的交叉验证
- Hyper-parameter Tuning 超参数调优
- Data Leakage Problem 数据泄露问题
- Confusion Matrix 混淆矩阵
- Accuracy, Precision and Recall 准确度、精确度和召回率
- Receiver Operating Characteristic (ROC) Curve 曲线ROC
- Area Under the Curve (AUC) 曲线下面积AUC

6. Decision Tree Model 决策树模型

- Introduction of Decision Tree 决策树介绍
- Entropy & Gini Impurity 熵和基尼不纯度
- Decision Tree Algorithm 决策树算法
- Pruning 剪枝

7. Random Forest Model 随机森林模型

- Ensemble Learning 集成学习
- Random Forest Algorithm 随机森林算法
- Feature Importance 特征重要性
- Out-of-bag Error 包外误差

8. Boosting Model: adaboost, gradient boosting, XGBoost Boosting 模型：adaboost、梯度提升、XGBoost

- Introduction to Boosting Method 方法介绍- Boosting
- Additive Modeling 可加模型
- Discrete AdaBoost 离散AdaBoost
- Shrinkage 收缩
- Gradient Boosting Machine - Overview 梯度提升机 - 概述
- Gradient Boosting Machine - Details & Example 梯度提升机 - 详细信息和示例
- Introduction to XGBoost 简介- XGBoost
- XGBoost - Loss Function Details 损失函数细节- XGBoost

9. K-nearest Neighbors Model K近邻模型

- K-nearest Neighbors Algorithm k近邻算法
- KNN Code 模型代码KNN
- Approximate Nearest Neighbors Algorithm 近似k近邻算法

10. K-means Model K均值聚类模型

- K-means Algorithm 算法K-means
- K-means Optimization 优化K-means
- Find the Optimal K 找到最佳K
- K-means ++ algorithsm(k-means clustering) K均值聚类算法

11. Linear Algebra 线性代数

- Matrix & Linear Transformation 矩阵和线性变换
- Eigenvector, Eigenvalue and Singular Value 特征向量、特征值和奇异值
- Eigendecomposition 特征分解
- Covariance Matrix 协方差矩阵
- Geometry of Linear Transformation - Basic Concepts 线性变换几何 - 基本概念
- Geometry of Linear Transformation - Change of Basis 线性变换几何 - 基础变化

12. PCA: Principal Component Analysis 主成分分析PCA

- PCA Formula 公式讲解PCA
- PCA Algorithm 算法步骤PCA
- PCA Code 模型代码PCA

IV. Deep Learning 深度学习

---

1. backpropagation 反向传播演算法

2. basics& optimization 基础知识和优化

3. convolutional neural network 卷积神经网络

4. recurrent neural networks 循环神经网络

5. Recommendation system design 推荐系统设计

```
-- x-mind?
```

Jupyter keyboard shortcuts: Command + Shift + P

I. Statistics 统计

Q1: What are the advantages and disadvantages of using sample mean as an estimator?

1. advantages:

i. Unbiasedness: The sample mean is an unbiased estimator of the population mean, which means that, on average, it will give an accurate estimate of the true population mean.

—> sample mean是population mean的无偏估计

ii. Consistency: As the sample size increases, the sample mean converges to the population mean, and variance decreases. This property ensures that with larger sample sizes, the sample mean becomes a more accurate estimator.

随着数据量增加，sample mean的variance以线性速度不断减小，精确度不断提高。

2. Disadvantages:

i. Sensitivity to outliers: The sample mean is sensitive to extreme values or outliers in the data. Outliers can disproportionately influence the estimate of the mean, leading to biased results.

ii. Sample dependence: The sample mean relies solely on the available sample data and does not take into account the underlying population distribution. It may not accurately estimate the population mean if the sample is not representative of the population.

iii. Sample size requirements: The sample mean performs well when the sample size is large enough, but it may not provide accurate estimates with small sample sizes. The precision of the estimate depends on the sample size, and small samples may lead to

Q2: What is a 95% confidence interval?

In statistics, a 95% confidence interval means that if we were to repeat the sampling procedure many times and calculate confidence intervals, approximately 95% of those intervals would contain the true population parameter.

Q3: Explain P-value definition.

P value is the probability of obtaining as or more extreme results than the current observation, under the null hypothesis.

① it is a probability

② it is calculated under the null hypothesis

③ as or more extreme results

p-value 是计算值，本身也是随机变量，随机性来源于sample data 随机性

significance level α不是计算出来的值，是预先设定的threshold

α 确定 HT中 p-value小到多少才能满足要求，一般定为0.05

Q4: How can we verify if the training data and validation data follow the same distribution?

Q5: How can we verify the randomness of missing values in our training dataset?

Q6: Describe the procedure of hypothesis testing.

Hypothesis testing is a way to assess the plausibility of an assumption regarding a population parameter by using sample data; help figure out the odds that results happened by chance

假设检验是一种通过使用样本数据来评估关于总体参数的假设的合理性的方法； 帮助找出结果偶然发生的几率。

i. 使用HT时因为可以控制样本随机性造成的错误。

ii. 找到结果发生的几率

① form the hypothesis 提出假设

② construct the estimator and test statistics 预测量 μ se z

③ obtain the distribution of test statistics under the null hypothesis 分布

④calculate p-value P值

⑤draw conclusion (reject or not reject null hypothesis) 结论

※设计hypothesis的原则是，试图reject null hypothesis, confirm alternative hypothesis。这是无罪推定，先假定无罪，再通过足够的理由推翻假定，证明有罪。这样做可以控制false positive。

H0：零假设是两者没有差异，观察到的差异是由于抽样或实验误差

Ha：备择假设是试图证明的有意思结果，是两者有差异。

Q7: What's the difference between p-value, type I error rate and significance level, Z-score?

i. Type I error = false positive = significance level = α

= P(reject H0| H0 is true) = P(p value < significance level | H0 is true)

the probability of rejecting the null hypothesis given that the null hypothesis is true.

the probability of having type I error is type I error rate α

零假设正确却拒绝了零假设，hypothesis testing犯错的概率

误判为阳性

ii. Type II error = false negative = β ＝1 - power

the probability of having type II error is type II error rate β

power is the probability that an experiment will flag a real change as statistically significant

零假设错误却接受了零假设，hypothesis testing犯错的概率

误判为阴性

iii. Type I error rate 和Type II error rate是此消彼长的

iv. Z-score:

对于sample mean estimator, 为后续计算方便，将其normalization，得到服从N(0, 1)分布的Z-score：

σ sample mean = σ population / n

Z = (X - μ) / σ sample mean

Q8: How to deal with the tradeoff between Type I error rate and Type II error rate?

(ie. How to set the significance level?)

大部分情况下，Type I error 比Type II error严重，比如法院定罪，互联网公司判定新产品效果

type I error: 本来无罪却判为有罪，本来新旧产品无差异却判定新产品有效，所以更严重

方法：

①choose a relatively lower significance level; 设定低的α

②increase sample size(collect more data) to decrease type II error rate. 提高样本数

Q9: Given a fixed-size sample data and 0.05 significance level, shall we reject the null hypothesis if the p-value is 0.049?

it depends

讨论use case中type I error 和type II error的严重程度

Q: 查准率&查全率

accuracy 精度 = 预测正确的样本 / 总样本 = TP + TN / TP + TN + FP + FN

precision 查准率 = 挑出的西瓜有多少是好西瓜 = TP / TP + FP

recall 查全率 = 所有好西瓜有多少被挑出了 = TP / TP + FN

Q. z-test & t-test

Q: Independence

Two events are independent, if the occurrence of one does not affect the probability of occurrence of the other.

Q: Conditional Independence

条件独立的基本定义: P(F1, F2 | C) = P(F1 | C) * P(F2 | C) 也就是说已知C条件的条件下，F1和F2是独立的。

Q: Bayes' Rule P(B|A)= P(B)* P(A|B) / P(A) 新信息出现后B的概率 = B原先的概率*信息带来的调整

P(B|A)被称作后验概率 (Posterior Probability): the likelihood of event B occurring given that A is true P(B)被称作先验概率 (Prior Probability)，也是单独B的概率 P(A|B)被称作似然函数(Likelihood)，也是条件概率: the likelihood of event A occurring given that B is

true P(A|B)P(A)被称为调整因子，使得预估概率更接近真是概率

Q: Bernoulli and Binomial Distributions

i. 伯努利分布 (Bernoulli distribution), 也叫做两点分布(two-point distribution) 伯努利试验（Bernoulli trial）是只有两种可能结果的单次随机试验. 一次试验只有两个可能结果，即“成功”和“失败” . “成功”是指我们感兴趣的某种特征 . P(Y=y)=py(1-p)1-y , 0<p<1, y=0,1

一个离散型随机变量X只取0和1两个可能的值

ii. 二项分布 Binomial distribution 二项分布即重复n次独立的伯努利试验. 重复进行 n 次试验，出现“成功”的次数对应的离散型随机变量X的概率分布称为二项分布.
P(X=x)=Cnxpxqn-x , x=0,1,2,...,n

Q: 中心极限定理 （Central Limit Theorem）

大量相互独立的随机变量，其均值（或者和）的分布以正态分布为极限（意思就是当满足某些条件的时候，比如Sample Size比较大，采样次数区域无穷大的时候，就越接近正态分布）。

即样本量极大时， 样本均值的抽样分布趋近于正态分布。这和样本所属的总体的分布类型无关。

中心极限定理： 大量相互独立的随机变量，其均值（或者和）的分布以正态分布为极限（意思就是当满足某些条件的时候，比如Sample Size比较大，采样次数区域无穷大的时候，就越接近正态分布）。

而这个定理最重要的地方在于，无论是什么分布的随机变量，都满足这个定理。

在自然界与生产中，一些现象受到许多相互独立的随机因素的影响，如果每个因素所产生的影响都很微小时，总的影响可以看作是服从正态分布的。中心极限定理就是从数学上证明了这一现象 。最早的中心极限定理是讨论n重伯努利试验中，事件A出现的次数渐近于正态分布的问题。

Q: 大数定理 Law of large numbers

大数定律讲的是，样本量极大时，样本的均值必然趋近于总体的期望， Xn ≈ u，因此我们可以用样本均值来估计总体的期望。

简单的可以描述为，如果有一个随机变量X，你不断的观察并且采样这个随机变量，得到了n个采样值，X1,X2,..., Xn，然后求得这n个采样值得平均值Xn，当n趋向于正无穷的时候，这个平均值就收敛于这个随机变量X的期望。

如果有一个随机变量X，你不断的观察并且采样这个随机变量，得到了n个采样值，X1,X2,..., Xn，然后求得这n个采样值得平均值Xn，当n趋向于正无穷的时候，这个平均值就收敛于这个随机变量X的期望。

在linear regression, logistic regression 都会用到

Q: 最大似然估计Maximum Likelihood Estimation

极大似然估计，利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值。

换句话说，极大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“模型已定，参数未知”。

参数未知，通过对结果的推测得到最有可能出现的参数

比其他估计方法更加简单； 收敛性：无偏或者渐近无偏，当样本数目增加时，收敛性质会更好； 如果假设的类条件概率模型正确，则通常能获得较好的结果。但如果假设模型出现偏差，将导致非常差的估计结果。

最大似然估计的一般求解过程：写出似然函数；对似然函数取对数，并整理；求导数；

概率是在特定环境下某件事情发生的可能性，也就是结果没有产生之前依据环境所对应的参数来预测某件事情发生的可能性，概率越大说明这件事情越可能会发生。 似然刚好相反，是在确定的结果下去推测产生这个结果的可能环境（参数），该事件在不同条件下发生的可能性，似然函数的值越大说明该事件在对应的条件下发生的可能性越大。

概率：已知参数，推测结果的可能性 概率描述的是在一定条件下某个事件发生的可能性，概率越大说明这件事情越可能会发生 似然：参数未知，通过对结果的推测求参数 似然描述的是结果已知的情况下，似然函数的值越大说明该事件在对应的条件下发生的可能性越大。

Q:

Discrete random variables Continuous random variable Probability density function (PDF) Cumulative Distribution Functions (CDFs) PMF

1 随机变量：“事件结果和数的对应关系”就形成了随机变量 随机变量的取值：直接实验结果，或者结果映射（函数） 随机变量的类型：离散型随机变量，或者是，连续型随机变量 随机变量是说，某个变量的值，不是一个确定的值；但是各取值的概率分布，即各取值之可能程度的大小关系，是确定的。概括总结，第一，随机变量所有的取值是清楚的；第二，随机变量的具体取值是不确定的。

随机变量一般用大写字母表示，其具体的取值一般用小写字母来表示。随机变量 X 取值为 x 的概率，本质上也是一个事件的概率，这个事件就是{X=x} ，我们将他记作：PX(x)=P({X=x}) 。

2 离散型随机变量 离散型随机变量（Discrete random variables） Definition: A discrete random variable is a random variable that has only a finite or countably infinite (think integers or whole numbers) number of possible values.

如果一个随机变量的全部可能取值，只有有限多个或可数的无穷多个，则称它是离散型随机变量，比如上面的计算两个骰子点数之和。

研究随机变量的方法就是穷举，由于离散型随机变量的取值可以一个个列出，故可以用分布律研究。连续情形不可一一列出，那就穷尽所有区间上的概率，分布函数恰好能够胜任。

Expected Value and Variance of Discrete Random Variables: 离散型随机变量的期望 (Expected Value) or Mean 描述离散型随机变量取值的集中程度

离散型随机变量的方差 (Variance=The square of Standard Deviation) 描述离散型随机变量取值的分散程度

离散型随机变量对应的常见分布有： 两点分布 二项分布 几何分布 超几何分布 均匀分布 泊松分布

3 连续型随机变量 如果随机变量的取值为连续的（如全部实数，一段区间），则称它为连续型随机变量 A continuous random variable is a random variable with infinitely many possible values (think an interval of real numbers, e.g., [0,1]). 连续型随机变量对应的常见分布有： 均匀分布 指数分布 正态分布

4 当我们使用概率函数描述连续概率分布时，我们称其为概率密度函数（probability density function），通常缩写为pdf。 概率密度函数（probability density function），连续型随机变量的概率密度函数是一个描述某个确定的取值点附近的可能性的函数。

5 累积分布函数Cumulative Distribution Functions (CDFs) for Discrete Random Variables CDF对于离散型随机变量的定义也适用于连续型随机变量 CDF是PDF的（从负无穷-oo到当前值的）积分，PDF是CDF的导数。 CDF相当于其左侧的面积，也相当于小于该值的概率，负无穷的CDF值为 0，正无穷的CDF值总为 1 。

累积分布函数CDF：概率函数取值的累加结果，所以它又叫累积概率函数。对于连续随机变量来说，概率是PDF的积分，累计概率函数也就是PDF的积分，只是区间不同。

6. 概率质量函数 Probability Mass Function (PMF)

概率质量函数就是将随机变量的每个值映射到其概率上。也就是说，我们可以计算理算随机变量等于一个特定值的概率。

Definition· The probability mass function (pmf) (or frequency function) of a discrete random variable X assigns probabilities to the possible values of the random variable. More specifically, if x1,x2,… denote the possible values of a random variable X, then the probability mass function is denoted as p and we write f(xi)=P(X=xi) 由于概率质量函数返回概率，所以它必须遵循概率法则（公理）。也就是说，概率质量函数输出0到1之间的值（含），而所有结果的概率质量函数输出之和等于1。

7 Normal Distributions

均值设为零（μ=0），标准差设为1（σ=1） 正态分布大概是所有概率和统计学问题中最常见的分布了。它如此常见的原因之一是中心极限定理。 使用错误的参数值会得到离你的期望相差很远的结果。

8 总结： 概率分布是结果及相应概率的列表。 我们可以用表格罗列小分布的结果和概率，但大分布用函数概括更方便。 离散概率分布的表示函数称为概率质量函数 PMF 连续概率分布的表示函数称为概率密度函数 PDF 表示概率分布的函数同样遵循概率法则。 概率质量函数PMF的输出是概率，概率密度函数PDF曲线下面积表示概率。 概率函数的参数在定义随机变量结果概率上起关键作用

Linear Regression

1.Data 类型，来源，size，clean dataset？pre-processing，分布，分析 数据集大，要不要抽样分析？ 2.parameters? Coefficient? Assumption Loss function / cost function MLE， LS 3. Evaluation MSE Overfitting LASSO，Ridge Bias vs. Variance

Loss function

(1) How to measure the performance of a linear regression model? use MSE to see if the model give the least error to test data,

(2) What is MSE? To calculate the difference of the predict value of the dataset, to find the average error of the prediction value

(3) Is MSE the smaller the better? Yes, the smaller the better on our testing dataset because MSE here is using for performance on testing dataset. But the model could be overfitting

(4) when the overfitting happens? Model too complex, fit perfect on training dataset, not applicable to testing dataset. Overfitting is coming when model too complex, less data

(5) how to solve the overfitting problem? 常见的： 最优：regularization Features reduction Increase data 主要考察regularization L1, L2 regularization, balance bias and variance , to make more stable

(6) what's the difference between L1 and L2? 首先： 从定义 其次：效果 L1 LASSO, not

Logistic Regression 1.Data 类型，来源，size，clean dataset? pre-processing，分布，分析 数据集大，要不要抽样分析？ 2.parameters? Coefficient? Assumption Loss function / cost function MLE， LS 3.Evaluation MSE Overfitting ROC, AUC Confusion matrix

（1） Cost Function是什么? argmini=1n[-yilog(h(xi))-(1-yi)log(1-h(xi))] （2） 如何interpret coefficient： Logistic regression parameter represents? 考察对logistic regression的定义的理解 （3） 如何construct ROC curve True positive rate vs. False positive rate, under different threshold （4） 如何interpret AUC? Does AUC the bigger the better? What does AUC=0.9 mean? （5） 如何estimate logistic regression的coeff Maximum Likelihood Estimation (6) How does MLE works? (7) Can you write MLE for logistic regression? (8) Can you write code to do MLE?

II. A/B testing A/B测试

Q10: Why do we need secondary metrics in A/B testing?

Acquisition → Activation → Retention → Revenue

i. track long-term user experience change

page bounce rate（点了又推出）; page load latency（加载延迟）; customer service calls/ messages

app-install; logins; searches; high-intent page views

ii. explain the revenue change

sustainable revenue growth = short-term revenue growth + user experience improvement

Q10.1 Direct metrics vs compound metrics: 直接指标和复合指标

i. Direct metrics: clicks, conversions, views, page bounce rate

ii. Compound metrics: LTV(User lifetime value), Page performance score(Google lighthouse)

Q11: Why do we need to do randomization in A/B testing?

i. 确保randomness: 一个user进入两个group的概率是一样的，完全随机。

ii. 确保单一变量控制：除experiment effect之外，两个group所有feature完全一样。

确保上述两个条件的构造实验过程，叫做create counterfactual.

—> 只有通过设计randomized experiment, 确保randomness和单一变量控制，才能证明因果关系causal effect。

Q12: How to do randomization?/ Given a random number generation between 0 and 1, how to implement the randomization process in A/B testing?

利用unique identifier生成一个随机数，然后将这个随机数映射到[0,1]之间，再进行上述处理。

Q13: Why do we usually set the same sample sizes in the treatment & control group?

因为在total sample size (m+n) 一定的前提下，当m = n时，two sample test得到的test statistics绝对值最大，假设检验的结果更容易显著(significant)。

Q14: How to test if the treatment control group assignment is randomized?

run A/A test first.

A/A test: 对treatment group 不施加treatment effect. 实验结果应是两个groups没有显著区别。

用A/A test来检测allocation操作是否有bias。

• A/A experiment. Compare people seeing the same thing to each other. See if the metric picks up the difference between the two. Any differences that you measure are due to the underlying variability, maybe of your system, of the user population, what users are doing, etc

Q15: What can we do if we realize the treatment control group assignment was not randomized after the data collection, i.e. impression process, has been finished?

segmentation：在metric calculation的时候，把数据按照有问题的feature分成subgroup，然后做weighted average。

avg_clicks = female_weights * female_group_avg_clicks + male_weights * male_group_avg_clicks

Q16: What's the relationship between A/B testing sample size and the sample data variance?

if sample variance is larger, need larger sample size.

if desired effect size is smaller, need larger sample size.

if desired power is larger(or desired type II error rate is smaller), need larger sample size.

if desired type I error rate is smaller, need larger sample size.

Q17: Metrics in A/B testing are significantly positive (negative) at the beginning, but become neutral later. What's the possible reason?

positive: novelty effect（novelty effect: users welcome changes）

negative: primacy effect. （primacy effect: users are reluctant to change.）

Q18: How to reduce the novelty effect in A/B testing?

1）把实验运行久一点

2）避免在单个实验引入过多或过于剧烈的改动

3）在enrollment阶段，只把new visitors加入到实验中

4）在analysis阶段，把new visitors和old visitors的treatment effect分开讨论

Q19: How to reduce the primary effect in A/B

Q20: How to resolve interference between treatment and control group?

spillover between control and treatment group will impact A/B testing result.

2 types of interference:

network effect: underestimate the treatment effect.

limited resource in two-sided markets: overestimate the treatment effect.

resolve interference using different allocation strategies:

network effect: network-based clustering

two-sided market: geo-based clustering

Q21: Analyze A/B testing result:

i. In the control group, we collected the app-usage-time of 1010 users {x1, x2,…, x10}, its sample mean value is 1 hour/day

ii. In the treatment group, we collected the app-usage-time of 990 users {y1, y2, … y990}, its sample mean value is 1.2 hour/day

Suppose their sample variances are the same σ^2 = 1. Can you tell which group has a longer app-usage time?

Q22: Power Analysis:

在enrollment中，需要估算sample size，从而计算实验需要运行的时长。

power analysis is to estimate the minimum sample size (n) required an a/b testing experiment, given a desired significance level α, effect size, and statistical power(1 - β)。

Q: Effect size:

the mean difference between two groups in the unit of population standard deviation

ES = |μ1 - μ0| / σ

III. Machine Learning 机器学习

Q1: When and why do we need to do feature normalization?

值间差异大，避免值大变量weight大。

feature scaling is helpful when one feature is much larger (or smaller) than another feature.

Its main purpose is to standardize the values of different features.

feature normalization is useful when the dataset has different feature scales, avoiding weight bias, require optimization algorithms(gradient descent, are sensitive to the input feature scales).

eg. the house size could be 2000, #of bedrooms is [1,5] when predict house price.

Q2: Why do we usually use squared loss as the loss function of ordinary linear regression?

1）Convexity: The squared loss function is convex, which means it has a single global minimum. This property makes it easier to find the optimal solution using optimization techniques like gradient descent. The convexity ensures that the optimization process converges to a unique solution.

2）Sensitivity to outliers: The squared loss function penalizes large errors more than linear or absolute loss functions. Consequently, it is more sensitive to outliers in the data. This sensitivity can be beneficial when outliers are considered as significant deviations from the underlying linear relationship and should have a larger impact on the model.

3）Maximum likelihood estimation: When the errors in the linear regression model follow a Gaussian (normal) distribution, minimizing the squared loss is equivalent to maximizing the likelihood of the observed data. This connection to maximum likelihood estimation provides a statistical interpretation for using squared loss in linear regression.

Q3: What's the difference and correlation between ordinary linear regression and logistics regression?

i. 同属generilized linear regression例子

ii. Linear Regression

① Y|X 服从Normal Distribution，

② 直线，predicting continuous values

③ g(E(Y|X)) = Xβ ; g() 是identity function, 即它自己

ii. Logistic Regression

① Y|X 服从Bernoulli Distribution,

② 曲线，predicting probabilities and making binary/ categorical classifications

③ g(E(Y|X)) = Xβ；g() 是logitfunction，反函数是g^-1()

Q4: What's the probabilistic assumption of ordinary linear regression?

normal distribution

Q5: What's the loss function of logistic regression?

Q6: What's the probabilistic assumption of logistic regression?

bernoulli distribution

Q7: How to deal with a multi-class classification problem?

cross entropy loss在binary classification和multi-class classification中都可以应用。

Q8: What's the trend of testing error and training eroor when we increase the model complexity?

Q9: What is overfitting?

i. overfit: when a model fits the training data well but does not work well with new examples that are not in the training set.

model too closely to training dataset, fail to fit test dataset or predict future observation reliably.

if have too many features, the model may fit the training set well, but almost too well and have high variance. 每次新的test dataset 都和真实值有较大偏差。

ii. underfit: too few features, it underfits and has bias.

iii. just right: use the model to predict outcomes correctly for new examples. generalize well.

Q9.1：What's the general strategies to reduce model overfitting?

i. more training date 增加训练数据量

ii. improve model formula 改进模型结构：

a. use fewer features 减少Feature个数 eg. feature selection, PCA

b. 改进loss function形式: 正则化 regularization

Q9.2 Bias& Variance: What is bias and variance? What is bias and variance trade off?

Bias: model space中多个模型的平均输出结果与真实值相比的差距。即整个model space的平均准确性。

Variance: model space中某个model输出结果与model space平均水平的差距的期望。

Model Error = Bias^2 + Variance + σ^2

Squared Error: Bias^2 + Variance

underfit: 模型与training set不准，bias大

overfit:模型与test set不准，variance大

Q10: What's L1 regularization and L2 regularization?

Q11: What's the difference between L1 regularization and L2 regularization?

Q12: How to find the optimal value of hyper-parameter in the regularization term?

Q13: How to prioritize precision and recall metrics in different use cases?

Q14: How to explain AUC from a probability perspective?

Q29: How to deal with categorical features at the

Q30: What are the common dimensionality reduction methods?

Q15: How to reduce overfitting of decision tree models?

Q16: Why do u consider a decision tree as a non-linear model?

Q17: Can tree-based models work well with sparse features? Why?

Q18: Why do random forest models usually provide better result than decision tree models?

Q19: How to tune hyper-parameters in random forest models?

Q20: What does the feature importance mean in random forest models and boosting models?

Q21: What's the difference between bagging methods and boosting methods?

Q22: Compare the advantages and disadvantages of decision tree, random forest and GBDT.

Q23: Compare the difference between XGBoost and GBDT.

Q33: What's the difference between batch gradient descent and stochastic gradient descent?

Q24: What's the time complexity of K-nearest neighbor? Can u further improve its time complexity with approximation methods?

Q25: What are the factors that influence performance of the K-means algorithms?

Q26: How to find the optimal K for K-means algorithms?

Q27: What to optimize centroids selection for K-means algorithm?

Q31: Describe the process of PCA.

Q32: Why do we need feature normalization during Principal Component Analysis?

Q28: What are the common techniques to fill missing values?

IV. Deep Learning 深度学习

Q34: What's the difference between backpropagation and gradient descent?

Q35: Why is backpropagation more efficient than forward-pass-only gradient calculation?

Q36: How to train your neural network model with multiple machines?

Q37: What's the difference between batch and epoch?

Q38: What is the dying ReLu problem?

Q39: How to reduce the overfitting problem in Neural Network Models?

Q40: What are the common techniques to reduce the gradient vanishing/ exploding problem?

Q41: What's the difference between kernel and filter in Convolutional Neural Network?

Q42: What's the difference process of doing backpropagation in a Recurrent Neural Network?