Top 100 Machine Learning Questions & Answers

机器学习面试常见问题 100 道

Q1. Explain the difference between supervised learning and unsupervised learning machine learning – 有监督学习和无监督学习的区别

1. Supervised learning needs to provide labeled data during training process.
2. Unsupervised learning needs not to provide labeled data.

（有监督学习需提供标签数据，而无监督学习无需提供。）

Q2. What are the parametric models? Give an example. – 介绍参数模型

1. Parametric models are those with a finite number of parameters. i.g Logistic regression, linear regression.
2. Non-parametric models are those with unbounded parameters. i.g SVM, Decision tree, naïve bayes.

（参数模型指带有有限参数的模型）

Q3. Difference between classification and regression? – 分类和回归任务的区别

1. Classification is used to produce discrete results, usually classify the data into specific categories.
2. Regression is used to deal with continuous data, like stock price trend.

（分类任务的结果通常是具体的类别，回归任务的一般用于预测连续数据的趋势）

Q4. Explain overfitting and how to avoid it? – 解释过拟合

1. Overfitting refers that a model learns the training set too well, which will affect the model ‘s ability to generalize and don’t apply to new data.
2. Overfitting: high variance, underfitting: high bias.
3. Methods to avoid overfitting:
   a) Regularization, it adds a cost term to the objective function as a feature
   b) Make the model simpler, reduce the training set or the variables/parameters
   c) Use cross validation like k-folds

（过拟合就是模型过于适应训练集数据，而对新的数据集泛化较差。 正则化/削减模型参数/使用交叉验证可以避免出现过拟合）

Q5. What is meant by Training set and Test set? – 解释训练集和测试集

1. Training set is used to train the model.
2. Test set is used to test the model.

（训练集用于训练模型，测试集用于测试训练过后模型的性能）

Q6. How to handle Missing or Corrupted data in a dataset? – 如何处理缺失/损坏数据?

1. Drop all columns or rows of the missing/corrupted data. Pandas: IsNull() and dropna()
2. Replace those missing/corrupted data with some other value like mean number or median. Pandas: Fillna().

（通过删除整行或是整列的缺失/损坏数据，或是使用其他数据来替换这些缺失/损坏数据）

Q7. Explain Ensemble learning – 解释集成学习

In ensemble learning, many base models like classifiers and regressors are generated and combined together so that they give better results. It is used when building component classifiers that are accurate and independent.

（对于训练数据, 通过训练若干个个体学习器, 通过一定的结合策略来得出更好的预测结果）

Q8. Explain the Bias-Variance Tradeoff. – 解释权衡偏差和方差

1. Bias refers how well the model fits the data, variance refers how well the model can generalize and apply to new data.
2. Simpler model are stable but not fit the data well (low variance, high bias), and more complex model are more prone to overfitting but fit the data better. (high variance, low bias). The best model usually lies in the middle.

（偏差反应模型与训练数据的契合度，方差反应模型与新的数据的泛化性，简单模型有可能出现欠拟合，即低方差高偏差，复杂的模型有可能出现过拟合，即高方差低偏差）

Q9. Difference between stochastic gradient descent (SGD) and gradient descent (GD)? – 随机梯度下降和梯度下降的区别?

1. Gradient descent computes the gradient of all training samples for each set of parameters and then find the path to the minimum one, which is slow and taking big computation resources.
2. Stochastic gradient descent only computes one gradient which is the mean value of gradient from a batch of training samples to find the path. This is much faster in computation and taking small resources.

（梯度下降要求计算出所有样本的梯度再找出到最小值的路径, 随机梯度下降则只需要计算抽取出样本的梯度的平均值再找出最小值的路径。梯度下降耗时长，消耗计算资源多；随机梯度下降耗时短，消耗计算资源少）

Q10. How to choose a classifier based on a training set data size? – 如何根据训练集数据大小来选择分类器?

1. When the training set is small, a model with right bias and low variance is better.
2. When the training set is complex, a model with low bias and high variance works better

（训练集小就选用低方差的模型来防止过拟合，复杂的数据集就选用高方差低偏差的模型）

Q11. What are 3 data preprocessing techniques to handle outliers? – 3 种处理异常值的方法?

1. Winsorize (cap at threshold)
2. Transform to reduce skew (box-cox)
3. Remove outliers

（缩尾处理/数据变换/删除）

Q12. How much data should be allocated for training, validation and test sets? – 将数据集分割成训练集，验证集和测试集的合适比例?

Generally, a good rule is to use an 80/20 train/test split, and train set can be further split into train/validation set or into partitions for cross validation.

（一般使用 80/20 的比例分割训练集和测试集, 在训练集中可以进一步分割出新的训练集和验证集）

Q13. Explain False Positive and False Negative – 解释 FP/FN

1. False Positive refers cases that wrongly get classified as True but are False.
2. False Negative refers cases that wrongly get classified as False but are True.

（FP 指的是做出 positive 的判定但此判定是错误的，FN 指的是做出 negative 判定但此判定是错误的）

注：

TP(True Positive): 做出 positive 的判定且此判定是正确的

TN(True Negative): 做出 negative 的判定且此判定是正确的

准确率(Accuracy) : $\frac{\text{所有预测正确的样本}}{\text{总样本}} = \frac{TP+TN}{TP+FN+FP+TN}$

精确率(Precision): $\frac{\text{将正类预测为正类}}{预测的正类} = \frac{TP}{TP+FP}$

召回率(Recall): $\frac{\text{将正类预测为正类}}{\text{原本的正类}} = \frac{TP}{TP+FN}$

Q14. Explain difference between L1 and L2 regularization – 解释正则化 l1，l2 的区别

L1 regularization penalizes the sum of absolute values of the weights, whereas L2 regularization penalizes the root of the sum of squares of the weights. The L1 regularization solution is sparse. The L2 regularization solution is non-sparse. L1 regularization corresponds to setting a Lapalacean prior, while L2 corresponds to a Gaussian prior.

（L1 是模型的各个权重的绝对值之和，L2 是模型各个权重的绝对值的平方和的开方值，L1 是稀疏的，可以用于特征选择，L2 是非稀疏的，可以用于防止过拟合，L1 服从拉普拉斯分布，L2 服从高斯分布。在实际使用中，如果特征是高维稀疏的，则使用 L1 正则；如果特征是低维稠密的，则使用 L2 正则。L1 也叫 Lasso regression, L2 也叫 Ridge regression）

Q15. Explain Fourier transform – 解释傅里叶变换

A Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain, it is a very common way to extract features from audio signals or other time series data.

（傅里叶变换可以将音频信号转换成若干正弦波之和，正弦波由频率，振幅，相位构成。傅里叶变换可以将信号从时域转换为频率, 在机器学习中傅里叶变换通常用来提取一下像音视频信号这种时序类型数据的特征）

Q16. Explain deep learning, difference between DL and other ML algorithm? – 解释深度学习，以及深度学习与其他机器学习算法的区别?

Deep learning is a subset of machine learning. And deep learning is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structred data.
（深度学习是机器学习的一个分支，深度学习中使用了神经网络，反向传播算法用来更好地训练预测未标注的或是半结构化大型数据集。）

Q17. Difference between generative and discriminative model? – 生成式模型和判别式模型的区别

1. A generative model focuses on how the data is generated while a discriminative model focuses on the distinction between the categories. A discriminative model performs better on multi-class classification. A generative model less likely occurs overfitting problems.
2. Generative model example: naïve bayes, discriminative model example: LR, DT,SVM.

（生成式模型为源头导向型，关注数据如何生成的。判别式模型为结果导向型，关注类别之间的差别。生成式模型较少遇到过拟合问题，判别式模型在多分类问题上表现良好）

Q18. State the application of supervised machine learning in modern business – 举例监督学习在现代商业中的应用案例

Email Spam Detection, Healthcare Diagnosis, Sentiment Analysis and Fraud Detection.
（垃圾邮件检测，健康诊断，情绪分析，欺诈检测）

Q19. Explain Semi-supervised learning – 解释半监督学习

In semi-supervised learning, the training data contains a small amout of labelled data and a large amount of unlabelled data.
（在半监督学习中，训练集包含了少量的已标注数据和大量的未标注数据）

Q20. State Unsupervised learning techniques – 列举无监督学习的方法

1. Clustering: In Clustering, data is divided into subsets. These subsets are called clusters, contain data that are similar to each other. Different clusters reveal different details about the object.
2. Association: Association algorithm identify patterns of associations between different variables or items. Example: Recommendation system.

（聚类算法：将数据分成若干个聚类簇，这些聚类簇中的对象之间具有较高的相似度，而不同聚类簇中的对象差别较大。 关联规则：探索寻找不同变量或是物体之间的关系）

Q21. Explain ‘naive’ in the Naïve Bayes Classifier – 解释朴素贝叶斯算法中的朴素的意思

‘naïve’ means that the classifier makes assumptions that may not may not be correct. The algorithm assumes that the presence of one feature of a class is not related to the presence of any other feature. Each feature affect the assumptions independently.
For example, a fruit may be considered to be a watermelon if it is green in color or roud in shape, regardless of other features. This assumption may or may not be right.
（朴素的意思为在此分类器中每个特征独立地对分类结果产生影响。例如当一个水果具有绿色或是圆形时，这个水果会被认为是西瓜，且这个分类结果有可能是正确的也有可能是错误的。）

Q22. Explain Latent Dirichlet Allocation(LDA) – 解释隐性狄利克雷分布

LDA is a common method of topic modeling, or classifying documents by subject matter. LDA is a generative model that represents documents as a mixture of topcis that each have their own probability distribution of possible words. In LDA, documents are distributions of topics that are distributions of words.

Q23. Explain Principle Component Analysis(PCA) – 解释主成分分析

PCA is a method for transforming features in a dataset by combing them into uncorrelated linear combinations. These new features (principal component) sequentially maximize the variance represented. For example, the first principal component has the most variance, the second principal component has the second most variance, and so on. Until the number of principal component meets the requirements.

As a result, PCA is useful for dimensionality reduction because the arbitrary variance cutoff can be set.

（PCA 可以将数据集中的特征转换一系列不相关的线性组合，新的特征通过转换后的最大方差值来进行线性筛选，例如第一个特征，即第一个主成分就是拥有最大方差值的那个特征，以此类推，直到主成分的数量满足需求。PCA 常常用来对数据进行降维）

Q24. Explain F1 score, and How would you use it? – 解释 F1 分数

F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being best, and those tending to 0 being worst. F1 score can be used in classification tasks where True Negatives don't matter much.

（F1 分数是用来评价模型性能的一个手段。F1 分数为准确率和召回率的倒数求取平均数后再取倒数。F1 分数趋近于 1 则模型性能好，F1 分数趋近于 0 则模型性能差。F1 分数被用于分类任务）

Q25. When should you use classification over regression? – 什么时候该使用分类任务而不是回归任务？

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. If you wanted the results to reflect the belongingness of data points in your dataset to certain explicit categories.

（分类任务的结果是离散的数值，即若干个类别，并且将数据集当中的数据点归类到这些类别中。回归任务的结果则是连续的，可以表明各个数据点之间的区别。当任务要求的结果为反应数据点的归属，则需要用到分类任务）

Q26. How to ensure the model is not overfitting? – 如何保证模型不会过拟合？

The possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

1. Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.
2. Use cross-validation techniques such as k-folds cross-validation

3. Use regularization techniques like LASSO that penalize certain model parameters if they are likely to cause overfitting.

（减少模型的参数以移除训练数据的噪音/使用交叉验证/使用正则化）

Q27. How to choose machine learning algorithm for the classification problem? – 如何为分类问题选择机器学习算法?

1. If accuracy is concerned, test different algorithms and cross-validate them.
2. If the training dataset is small, use low variance and high bias models.
3. If the training dataset is large, use high variance and low bias models.

（若模型准确重要则多测试几个算法模型并且使用交叉验证。若训练集很小，则选用低方差高偏差的模型。若训练集很大则选用高方差低偏差的模型）

Q28. How to design an email spam filter? – 如何设计一个垃圾邮件过滤器?

1. The email spam filter will be fed with thousands of emails.
2. Each email will have label: spam or not spam.
3. The supervised machine learning algorithm will then determine which type of emails are being marked as spam based on spam words like the lottery, free offer, no money, refunded, etc.
4. The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like DT, SVM to determine how likely the email is spam.
5. If the likelihood is high, it will label it as spam, and the email will not hit your inbox
6. Based on the accuracy of each model, the algorithm with the highest accuracy will be used after testing all models.

（垃圾邮件过滤器将被输入许多分别带有“是垃圾邮件”和“不是垃圾邮件”标签的邮件，有监督学习算法将会设置若干个垃圾邮件关键词来识别垃圾邮件，当有邮件想要送入你的邮箱时，机器学习算法会自动判断该邮件是垃圾邮件的可能性，若可能性高则该邮件不会进入你的邮箱。最后基于每个模型的准确率来选择最优的模型。）

Q29. What evaluation approaches would you work to gauge the effectiveness of a machine learning model? – 用什么评估方法来测量一个机器学习模型的效率?

1. Spliting dataset into training and test sets, or perhaps use cross-validation techniques to further segment dataset into composite sets of training and test sets.
2. Implementing a choice selection of performance metrics: F1 score/the accuracy/confusion matrix.

（首先，先将数据集分成训练集与测试集，或是使用交叉验证进一步将数据集分割成若干组训练集与测试集。其次，选择一种性能评估标准例如：F1 分数，准确率，混淆矩阵）

Q30. How would you implement a recommendation system for our company’s users?

A lot of machine learning interview questions of this type will involve the implementation of machine learning models to a company’s problems. You’ll have research the company and its industry in-depth, especially the revenue drivers the company has, and the types of users the company takes on in the context of the industry it’s in.

Q31. Explain Bagging – 解释装袋法

Bagging, or Bootstrap Aggregating, is an ensemble method in which the dataset is first divided into multiple subsets through resampling.
Then, each subset is used to train a model, and the final predictions are made through voting or averaging the component models. Bagging is performed in parallel.
（装袋法，又名自助聚集法，是集成方法。首先将数据集通过重采样分成若干个子集，接着每个子数据集都会被用来训练一个模型。最终的预测结果将通过所有子模型的训练结果决定。装袋法是并行执行的。）

Q32. Explain ROC curve and AUC – 解释 ROC 曲线和 AUC 值
ROC(Receiver Operating Characteristic) is the performance plot for binary classifiers of True Positive Rate(TPR, y-axis) vs. False Positive Rate(FPR, x-axis).
AUC is the area under the ROC curve. And it is a common performance metric for evaluating binary classification models. It is equivalent to the expected probability that a uniformly drawn random positive is ranked before a uniformly drawn random negative.
（ROC 曲线用于评估一个二分类器的性能，x 轴为假阳率 FP/FP+TN，y 轴为真阳率 TP/TP+FN。根据 ROC 曲线的位置将图分为两个区域，曲线下方的部分被称为 AUC，用来表示预测准确性，AUC 值越高，即曲线下方面积越大，则准确率越高。曲线越接近左上角，准确率越高）

Q33. Why is AUC better than raw accuracy as an out-of-sample evaluation metric? – 为什么 AUC 对于样本外评估的表现要优于原始精度?
AUROC is robust to class imbalance, unlike raw accuracy
（AUROC 对类不平衡具有鲁棒性）

Q34. What are the advantages and disadvantages of neural networks? – 神经网络的优点和缺点？
Advantage: Neural network(specifically dnn) have led to performance breakthroughs for unstructured datasets such as images, audio, and video. Their incredible flexibility allows them to learn patterns that no other ML algorithms can learn.
Disadvantages: Neural network require a large amount of training data to converge. It's also difficult to pick the right architecture, and the internal hidden layers are incomprehensible.
（优点：神经网络在非结构化数据集例如图片，音视频等的训练和学习上性能优越。缺点：神经网络需要大量的训练数据来使模型收敛，并且神经网络中的隐藏层是不可解释的）

Q35. Define Precision and Recall – 定义精确率和召回率
Precision is the ratio of several events that are correctly predicted to the total number of events that are predicted as positive. Precision = TP/TP+FP
Recall is the ratio of a number of events that are correctly predicted to the total number of events are predicted correctly. Recall = TP/TP+FN
（精确率为将正类预测为正类和所有被预测为正类的比率，即精确率=真阳/真阳+假阳。召回率为将正类预测为正类和原本的正类的比率，即召回率=真阳/真阳+假阴）

Q36. Explain Decision Tree classification – 解释决策树算法
A decision tree builds classification models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with

branches and nodes. Decision trees can handle both categorial and numerical data. The policy is a top-down divide-and-conquer.
（开始时，构建一个根节点，所有数据都存放在此根节点中，选择若干个最优特征，将数据依照特征分割成若干个子集，是为子节点。接着在所有子节点内部按照特征继续分割。分割停止条件：当前节点内包含的数据都为同一类/当前节点内的数据无法明确分类/当前节点内的数据为空。决策树的策略为自上而下的分而治之。）

Q37. Explain Pruning in Decision Tree, and how is it done? – 解释决策树中的剪枝，并且如何执行？

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final model, and hence improve the predictive accuracy by reducing of overfitting.
Pruning has two styles:

1. Top-down pruning will traverse nodes and trim subtrees starting at the root.
2. Bottom-up style will begin at the leaf nodes.

A popular pruning algorithm: reduced error pruning:

1. Starting at leaf nodes, each node is replaced with its most popular class.
2. If the prediction accuracy is not affected, the change is kept.
3. Advantage: simpler and faster.

（剪枝是一种减少决策树尺寸的方法，剪枝能降低最终模型的复杂度，降低模型过拟合从而提高模型的准确率。剪枝有两种形式：预剪枝和后剪枝。预剪枝自上而下提前终止某些分支的生长，后剪枝自下而下先形成完整的模型然后再回头剪枝。一种流行的剪枝算法：降低错误率剪枝算法：这是一种后剪枝算法，即当树生成后，将子树替换成其叶节点，类别按叶节点中数据最多的类，然后判断剪枝与不剪枝之间的性能差异，若性能不被影响，则保留剪枝。）

Q38. What is a recommendation system? – 什么是推荐系统？

It is an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.
这是一种信息过滤系统，这种系统在基于之前用户的习惯或是选择模式之上预测用户想要听到或是看到的内容。

Q39. What is Kernel SVM? – 什么是支持向量机？

Kernel SVM is the abbreviated version of the kernel support vector machine. Kernel SVM uses the function of kernel to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions.
（支持向量机使用核函数来将原本线性不可分的数据转化为高维数据以此来进行线性划分，不同的支持向量机算法会使用不同的核函数。）

Q40. What are some methods of reducing dimensionality? – 有哪些数据降维的方法？

You can reduce dimensionality by combing features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction. Example: PCA
（降维方法：特征工程，移除共线特征，或者算法降维。例子：PCA）

Q41. What are the three stages of building a model in machine learning? – 构建机器学

习模型的三个阶段是?

1. Model Building: choose a suitable algorithm for the model and train it according to the requirement.
2. Model Testing: check the accuracy of the model through the test data.
3. Applying the model: make the required changes after testing and use the final model for real-time projects. The model needs to be checked to make sure it's working correctly. It should be modified to make sure that it's up-to-date.

（首先选择一个合适的算法并且训练模型直至符合需求，接着进行模型测试，使用测试数据对模型的准确率进行测试，最后按照测试结果对模型进行进一步修改之后部署模型。）

Q42. How is KNN different from k-means clustering? – KNN 与 k-means 之间的区别?

1. K-Nearest Neighbors is a supervised classification algorithm, the mechanism of KNN is based on calculating the distance between the unlabeled sample and the labeled samples, the unlabeled sample will be classified into the class which has the shortest distance(nearest neighbor).
2. K-means is a unsupervised clustering algorithm. It requires a set of unlabeled data and a threshold. The K-means will take unlabeled data and iteratively learn how to cluster them into groups by computing the means of the distance between different data points in each cluster.

（KNN 是一种监督学习分类算法, KNN 的机制是通过计算未知样本和所有样本之间的距离，以与该未知样本拥有最短距离的类别作为分类类别（最近邻者）。K-means 算法则是一种无监督学习聚类算法，K-means 的机制是给定一个 K 值和一组未标注的数据，K 值分配了 K 个初始簇类中心点，开始时把数据点分到离其最近的簇类中心点，所有数据分配完毕之后，再根据一个簇类内的所有点重新计算该簇内的中心点（取平均值），然后迭代进行分配数据和更新簇类中心的步骤，直到中心点变化很小或者达到指定迭代次数。）

Q43. Mention the difference between data mining and machine learning – 数据挖掘和机器学习的区别

Machine learning related to the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this processing, machine learning algorithms are used.

（机器学习是告诉计算机如何去学习和预测数据, 数据挖掘是用来从大量的数据中提取特定的模式，知识或是规则。）

Q44. What are the different algorithm techniques in machine learning? – 列举机器学习中的不同算法类别

Supervised learning, Unsupervised learning, Semi-supervised learning and Reinforcement learning.

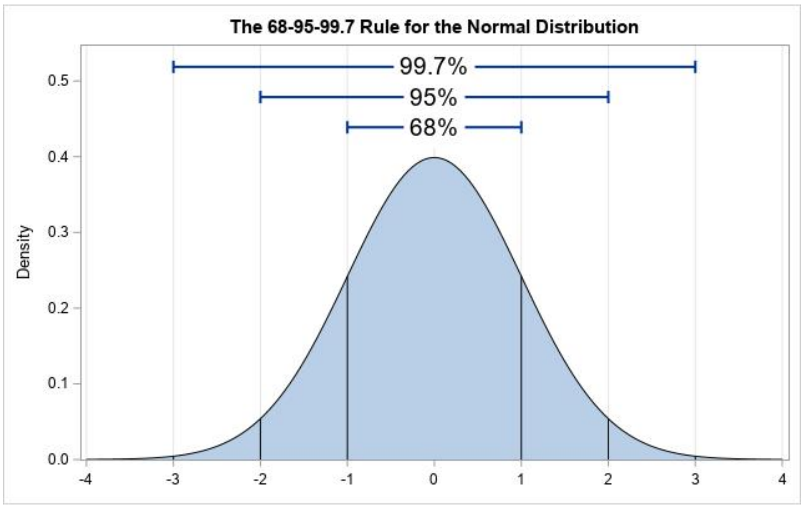
（监督学习，无监督学习，半监督学习，强化学习）

Q45. You are given a dataset, the dataset has missing values that spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why? - 给定一个数据集，该数据集具有沿中位数 1 个标准差分布的缺失值。

有多少百分比的数据不受影响？ 为什么？

Since the data is spread across the median, let’s assume that it’s a normal distribution. In normal distribution, about 68% of the data lies in 1 standard deviation from mean(or mode, median), which leaves about 32% of the data unaffected. Therefore, about 32% of the data would remain unaffected by missing values.

（因为数据分布在中位数附近，因此假设此数据集服从正态分布。在正态分布中，大约 68%的数据分布在距离平均值或是中位数附近的 1 倍标准差之间，大约 95%的数据分布在距离平均值或是中位数附近的 2 倍标准差之间，因此由题可得，大约 32%的数据落在中位数附近的 1 倍标准差之外，所以有 32%的数据不受影响。）



Q46. What are PCA, KPCA, and ICA used for? – 解释 PCA, KPCA, ICA 的用途

PCA(Principle Component Analysis), KPCA(Kernel-based Principle Component Analysis) and ICA(Independent Component Analysis) are used in feature extraction techniques for dimensionality reduction.

（PCA 主成分分析，KPCA 核主成分分析，ICA 独立成分分析用于数据降维的特征提取技术）

PCA 主成分分析：参照 **Q23**

KPCA 核主成分分析：核主成分分析是利用核函数将线性不可分的输入空间映射到线性可分的高维特征空间中，然后再按照主成分分析的方法对特征空间降维。PCA 是线性的，面对非线性数据例如人脸图像无法处理，基于核函数的主成分分析是 PCA 的非线性扩展。

ICA 独立成分分析：从混杂数据中分离出线性叠加而来的单独源数据。

Q47. What are support vector machines? – 什么是支持向量机？

Support vector machines are supervised learning algorithm used for classification and regression analysis.

（线性可分的 SVM：对于线性可分的数据集来说，找到一个超平面，使得这个超平面能将数据集当中的两个类别划分开。线性可分的 SVM 只能用于二分类问题。对于线性不可分的数据集，需要使用核函数，参照 **Q39**）

Q48. What is batch statistical learning? – 什么是批量统计学习？

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

（统计学习技术允许从一组观察到的数据中学习一个函数或预测变量，这些数据可以对看不见的或未来的数据进行预测。 这些技术基于对数据生成过程的统计假设，为学习预测器对未来看不见的数据的性能提供保证。）

Q49. What is the bias-variance decomposition of classification error in the ensemble method? – 解释在集成方法中对于分类误差的偏差方差分解

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

（偏差方差分解可以把一种学习算法的期望误差分解为偏差 bias 和方差 variance。偏差度量了某种学习算法的平均预测结果所能逼近学习目标的程度，方差度量了在面对同样规模的不同训练集时，学习算法的预测结果发生变动的程度。偏差度量了学习算法期望预测与真实结果的偏离程度，即刻画了学习算法本身的拟合能力；方差度量了同样大小的训练集的变动所导致的学习性能的变化。）

Q50. When is Ridge regression favorable over Lasso regression? – 什么时候 Ridge 回归会优于 Lasso 回归?

In the presence of few variables with medium/large sized effect, use Lasso regression. In presence of many variables with small small/medium sized effect, use Ridge regression.

Lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression (L2) only does parameter shrinkage and end up including all the coefficients in the model. In the process of correlated variables, ridge regression might be the preferred choice. Also, Ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective,

（Lasso 回归能够同时进行变量选择和参数收缩，而 Ridge 回归只能进行参数收缩。在有关变量的处理中优先选择 Ridge 回归，此外 Ridge 回归在最小二乘估计具有较高方差的情况下效果最佳。在变量很少的情况下可以选择 Lasso 回归，在变量多的情况下选择 Ridge 回归。）

Q51. You've built a random forest model with 10,000 trees. You got delighted after getting training error as 0. But the validation error is 34.23. What is going on? – 你构建了一个带有 10,000 棵树的随机森林模型，在训练过后你发现训练误差为 0 但是验证误差却高达 34.23，解释哪里出问题了？

The model has overfitting. Training error 0 means that the model has mimicked the training data patterns to an extent, and the model will perform bad on the unseen data. Hence, when the model was run the an unseen sample, it could not find the patterns and returned prediction with higher error. In random forest model, it happens when we use a larger number of trees than necessary. Hence, to avoid this situation, we should tune the number of trees using cross-validation.

（这个模型出现了过拟合的现象。训练误差为 0 意味着此模型与训练数据过于契合从而导致此模型遇到新的数据时无法识别新数据的特征，因此预测结果会存在更大的误差。在随机森林模型中，过拟合通常是因为模型使用了过多的树，因此只要使用交叉验证将树的数量进行

分割调整就能避免过拟合。）

Q52. What is convex hull? – 什么是凸包？

In the case of linearly separable data, the convex hull represents the outer boundaries of the two groups of data points. Once the convex hull is created, we get maximum margin hyperplane as a perpendicular bisector between two convex hulls. The maximum margin hyperplane is the line which attempts to create the greatest separation between two groups.

（在线性可分的数据集中，凸包表示两组数据点的外边界。一旦创建了凸包，就可以得到最大边距超平面作为两个凸包之间的垂直分界线。此算法即是求取使得两组凸包之间距离最大的最大边距超平面。）

Q53. In k-means or KNN, we use Euclidean distance to calculate the distance between nearest neighbors. Why not Manhattan distance? – 为什么使用欧几里得距离来计算 K-means 或是 KNN 当中的最邻近数据而不是曼哈顿距离？

Manhattan distance only calculates distance horizontally or vertically, which means that it calculates the distance through the projection on the axis. It has dimension restrictions. On the other hand, the Euclidean distance focuses on the relative distance between two points. In KNN or K-means, we focus on the absolute difference between two points rather than the projection, and the Euclidean distance can be used in any space, which has no dimension restriction.

（曼哈顿距离计算的是两点间的距离在坐标轴上的投影，并且曼哈顿距离的计算只能水平或是垂直移动，有维度限制。而欧几里得距离计算的是两点之间的相对距离，并且欧几里得距离能在任意空间当中使用，没有维度限制。在 K-means 或是 KNN 当中，我们更关注两个数据（两点）之间的绝对差异而不是投影差异，所以使用欧几里得距离）

欧几里得距离：$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}=\sqrt{\sum_{k=1}^{n}(x_i-y_i)^2}$

曼哈顿距离：$d_{12}=\sum_{k=1}^{n}|x_{1k}-x_{2k}|$

Q54. Do you suggest that treating a categorical variable as a continuous variable would result in a better predictive model? – 将分类变量视为连续变量能否得出更好的预测模型？

For better prediction, the categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

（只用当该变量是有序变量时，将分类变量视为连续变量才会得出更好的预测模型）

Q55. OLS is to linear regression. The maximum likelihood is logistic regression. Explain the statement. – 解释最小二乘法是线性回归，最大似然估计是逻辑回归

Ordinary Least Square is a method to approximate the unknown parameters in the minimum distance between the actual value and the predicted value. Maximum likelihood is a method to choose the values of the parameters that maximizes the likelihood that the parameters are mostly likely to produce the observed data.

（线性回归是求取一条直线来拟合所给的数据点，最小二乘法的目的就是求取用来拟合这条直线的未知参数。假设 $y=wx+b$ 可以拟合一组给定的数据 $\{(x_1,y_1),(x_2,y_2),(x_3,y_3)\ldots(x_n,y_n)\}$，得到拟合程度最大的方程，即$\hat{y}=wx+b$。最小二乘

法的目的即是求取能使得$|y_i - \hat{y_i}|$最小，等价于使得$(y_i - \hat{y_i})^2$最小，等价于使得$\sum_{i=1}^{m}(y_i - (wx_i + b))^2$最小的 w 和 b。$\sum_{i=1}^{m}(y_i - (wx_i + b))^2$就是损失函数。

令损失函数为$J(w,b) = \sum_{i=1}^{m}(y_i - (wx_i + b))^2$, 原问题先转化为求取损失函数的最小值问题, 现对损失函数中的两个参数分别求取偏导数即可得。

补充：梯度下降法求取损失函数的最小值。对于损失函数$J(w,b)$，现有：

$$w = w - \alpha\frac{\partial J(w,b)}{\partial w}$$

$$b = b - \alpha\frac{\partial J(w,b)}{\partial b}$$

梯度下降和最小二乘法都用于计算损失函数的最小值, 区别在于梯度下降法需要在最开始时先对初始 w 和 b 取任意估计值，并且在后续迭代中还需要学习率α的参与。而最小二乘法则无需给出初始 w 和 b 的估计值，而是直接进行求导计算。

逻辑回归是在线性回归的基础上加了一层激活函数，逻辑回归用于解决二分类问题。现有线性回归模型$z = wx + b$，加上激活函数 sigmoid 函数可得：$g(z) = \frac{1}{1+e^{-z}}$。Sigmoid 函数能将 z 转化成一个接近 1 或是 0 的值。由于 sigmoid 函数的取值在[0,1]之间，所以可以将其视为类 1 的后验概率估计$P = (y = 1|x)$。因此，给定一个阈值$threshold = 0.5$，把通过 sigmoid 函数计算得到的值大于 0.5 的归到类 1，小于 0.5 的归到类 0。接着推导逻辑回归模型的损失函数$J(w) = \frac{1}{2}\sum_{i=1}^{m}(g(z_i) - \hat{y_i})^2$，将$g(z) = \frac{1}{1+e^{-z}}$代入发现损失函数为非凸函数，存在多个极值点不利于求解。因此，先将$g(z)$转化为类 1 的后验概率：

$$p(y = 1|x;w) = g(wx + b) = g(z)$$

$$p(y = 0|x;w) = 1 - g(z)$$

其中$p(y = 1|x;w)$表示给定 w，在 x 处 y=1 的概率。

上面两式写成一般形式为：$p(y|x;w) = g(z)^y(1 - g(z))^{1-y}$

接着，就要用最大似然估计法来根据给定的训练集估计出参数 w 的值。即最大似然估计法就是通过已知的结果取反推最大概率导致该结果的参数。使用最大似然估计法后得：

$$L(w) = \prod_{i=1}^{n} p(y_i|x_i;w) = \prod_{i=1}^{n}(g(z_i))^{y_i}(1 - g(z_i))^{1-y_i}(g(z_i))^{y_i}$$

化简得：

$$J(w) = J(g(z),y;w) = \{\begin{matrix} -\ln g(z) & \text{if } y = 1 \\ -\ln(1 - g(z)) & \text{if } y = 0 \end{matrix}$$

再使用梯度下降算法对损失函数进行求导即可求得参数。

Q56. When does regularization becomes necessary in machine learning? – 在机器学习中什么时候应该使用正则化?

Regularization becomes necessary when the model begins to overfit or underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce the cost term. This helps to reduce model complexity so that the model can become better a t predicting.

（当模型出现过拟合或是欠拟合的时候需要使用正则化。正则化会在目标函数中引入一项一个惩罚项，惩罚模型的复杂度。这个惩罚项可以将许多变量的系数降至接近 0 从而降低模型的复杂度防止过拟合。）

Q57. What is Linear Regression? – 什么是线性回归?

Linear regression is a supervised learning algorithm. It is used to find the linear relationship between the dependent and the independent variables for predictive analysis.

（线性回归是一种监督学习算法。线性回归被用来寻找因变量与自变量之间的线性关系，并且以此来进行预测分析。）

Q58. What is Variance Inflation Factor? – 什么是方差膨胀因子？

Variance Inflation Factor (VIF) is the estimate of the volume of multicollinearity in a collection of many regression variables.

VIF=Variance of the model/Variance of the model with a single independent variable.

When the variables are linearly independent, the VIF is 1. If VIF is high, then it shows the high collinearity of the independent variables, we can deleting those variables with high VIF or combing these variables into one single variable.

（方差膨胀因子是一种度量回归变量之间的多重共线性的方法。VIF 等于多元模型中该系数的方差与只有一个变量的模型中该系数的方差的商。当各变量之间线性无关时方差膨胀因子为 1.如果方差膨胀因子过大，则说明自变量之间具有较强的相关性，可以去掉这些变量或是将这些变量组合成一个单一的变量。）

Q59. We know that one hot encoding increases the dimensionality of a dataset, but label encoding does not. How? – 为什么 one-hot 编码会增加数据的维度而标签编码不会?

One-hot encoding creates a new variable for each level I the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

（One-hot 编码：对于每一个特征，如果此特征有 m 个可能值，那么经过 one-hot 编码后，这些可能值就会变成 m 个二元特征，并且这些特征互斥。例如西瓜数据集中特征瓜蒂有 4 个可能特征值：圆形，卷形，线形和无瓜蒂，那么经过 one-hot 编码后，这 4 个可能特征值就会变成 1000，0100，0010 和 0001 四个二元特征值。）

Q60. What is a Decision Tree? – 什么是决策树?

A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.

（决策树是一种树结构的分类算法。其每个非叶节点表示一个特征属性的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，知道到达叶子节点，将叶子节点存放的类别作为决策结果。）

Q61. What is the Binarizing of data? How to Binarize? – 什么是数据二值化？如何数据二值化？

Converting data into binary values on the basis of threshold values is known as the binarizing of data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when we have to perform feature engineering, and we can also use it for adding unique features.

（根据给定的阈值将矩阵数据转化为二元数值的方法就是数据二值化。有些业务不需要分析矩阵的完整数据，可以根据给定的阈值，将矩阵中的数值低于阈值的变成 0，将高于阈值的变成 1，二值化后矩阵中的数据只有 0 和 1，达到简化模型的目的。数据二值化经常在特征工程中使用，并且也可以用来添加特定的特征。）

Q62. What is cross-validation? – 什么是交叉验证？

Cross-validation is essential a technique used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data into two groups: training data and testing data, where you use the training data to build the model and the testing data to test the model.

（保留交叉验证 hand-out cross validation：首先随机地将数据分为训练集和测试集，然后进行训练，在第一次训练完成后随机打乱数据集重新进行分割再进行训练，最后在得到的不同的模型上评估测试误差，选择测试误差最小的模型。

K 折交叉验证 k-fold cross validation：首先随地将数据集分割为 k 个互不相交的大小相同的子集，然后将 k-1 个子集当成训练集训练模型，剩下的一个子集当成测试集测试模型，接着对上一步可能的 k 种选择重复进行，每次都挑选不同的子集作为测试集，这样就训练了 k 个模型，每个模型都在相应的测试集上计算测试误差，得到了 k 个测试误差，对这个 k 次的测试误差取平均值便得到一个交叉验证误差。

留一交叉验证 leave-one-out cross validation：k 折交叉验证的特殊情况，k=N，N 是数据集的样本数量。）

补充. Bootstrapping：bootstrapping 和交叉验证一样使用重采样（resampling）的方法，是一种有放回的抽样方法，具体过程为：假设数据分成 10 组，接着设置一个采样比例，比如采样比例 70%，则 10 组数据是每次从原始数据集中随机采样总数 70%的数据构成训练集 1，没有选中的样本作为测试集 1，然后将数据放回，再随机采样 70%数据构成训练集 2，没有被选中的作为测试集 2，以此类推直到满足 10 组数据。然后训练生成 10 个模型，计算平均误差来评估当前参属下的模型性能。

Q63. When would you use random forests Vs SVM and why? – 对比 SVM，什么时候适合使用随机森林算法？为什么？

The reasons why a random forest is a better choice of the model than a SVM:

1. Random forests allow you to determine the feature importance. SVM cannot do it.
2. Random forests are much quicker and simpler to build than an SVM
3. For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

Q64. What are the drawbacks of a linear model? – 线性模型的缺点是什么？

1. A linear model holds some strong assumptions that may not be true in the application. It assumes a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity.
2. A linear model cannot be used for discrete or binary outcomes.
3. You cannot vary the model flexibility of a linear model.

（1. 线性模型包含有些在应用中可能不正确的假设。它假设线性关系，多元正态性，没有或很少多重共线性，没有自相关性和同方差性。

2. 线性模型不能处理离散的或是二元的结果。
3. 无法改变线性模型的灵活性）

Q65. Do you think 50 small decision trees are better than a large one? Why? – 具有 50 个分支的决策树是否要比更大的决策树更好？为什么？

Another way of asking this question is “Is a random forest a better model than a decision tree?” And the answer is yes because a random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

（另一种问法是随机森林模型是否比决策树更好？答案：随机森林模型比决策树更好，因为随机森林是一种集成方法，随机森林可以将许多弱决策树集成为一个强学习器。随机森林更精准，鲁棒性更强并且更不会导致过拟合。）

Q66. What is a kernel? Explain the kernel trick – 什么是核函数？

A kernel is a way of computing the dot product of two vectors in some (possibly very high dimensional) feature space, which is why kernel functions are sometimes called generalized dot product.

The kernel trick is a method of using a linear classifier to solve a non-linear problem by transforming linearly inseparable data to linearly separable data in a higher dimension.

（核函数是将原始空间中的向量作为输入向量，并返回特征空间（很有可能是高维）中向量的点积的函数。核函数能够将线性不可分的数据转化高维空间中的线性可分数据，并且将数据用在线性分类模型中。）

Q67. State the difference between causality and correlation? – 说明因果关系和相关性之间的区别

Causality refers to situations where one action X causes an outcome Y, whereas correlation is just relating one action X to another action Y, but X does not necessarily cause Y.

（因果关系指的是一个行为 X 导致了一个结果 Y，而相关性仅仅说明了行为 X 与行为 Y 有关系，而不能说明行为 X 导致了行为 Y。）

Q68. What is the exploding gradient problem while using the backpropagation technique? – 什么是梯度爆炸?

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the vanishing gradient problem.

（梯度在神经网络训练时被用来得到网络参数更新的方向和幅度, 进而在正确的方向上以合适的幅度更新网络参数。在深层网络中，梯度在更新中累积到一个非常大的梯度，这种梯度会大幅度更新网络参数，进而导致网络不稳定，在极端情况下权重的值会大到溢出，变成 NaN 值。当梯度爆炸发生时，网络层之间反复乘以大于 1 的梯度值会使得梯度成倍增长。）

补充：

1. 梯度消失：在深层网络中，梯度值过小会导致梯度消失。

2. 防止梯度爆炸：
   梯度裁剪：设置阈值，若梯度超过这个阈值则直接裁剪或是将阈值设为梯度。
   使用修正线性激活函数 ReLU 函数。
   Batch normalization：规范神经网络输出，控制过拟合。
3. 防止梯度消失：将激活函数改为 ReLU 函数

Q69. What do you mean by Associative Rule Mining? – 什么是关联规则挖掘?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features which are correlated.
（关联规则挖掘是一种发现数据模式的技术，例如同时出现的特征和相关的特征。）

Q70. What is Marginalisation? Explain the process. – 什么是边缘化?

Marginalisation is summing the probability of a random variable X given the joint probability distribution of X with other variables. It is an application of the law of total probability.
（边际化是在给定 X 与其他变量的联合概率分布的情况下对随机变量 X 的概率求和。 它是全概率法则的应用。）

Q71. Why is the rotation of components so important in PCA? – 为什么旋转变换在 PCA 中很重要?

Rotation in PCA is very important as it maximizes the separation within the variance obtained by all the components because of which interpretation of components would become easier. If the components are not rotated, then we need extend components to describe the variance of the components.
（在 PCA 中旋转是必要的，因为旋转能把由主成分捕获的方差之间的差异最大化，这使得主成分更容易解释。通过旋转，各主成分的相对位置不发生变化，它只能改变点的实际坐标。如果我们没有旋转主成分，PCA 的效果会减弱，那样我们会不得不选择更多个主成分来解释数据集里的方差。）

Q72. What is difference between regularization and normalization? – 正则化和归一化有什么区别?

Normalization adjusts data; regularization adjusts the prediction function. If your data is on very different scales (especially low or high), you would want to normalize the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting. Regularization imposes some control o this by providing simpler fitting functions over complex ones.
（归一化是将数据映射到指定的范围，当数据之间差别非常大的时候，可以使用归一化。常见的归一化映射范围有[0,1]和[-1,1]。正则化主要用于避免过拟合和减少网络误差。正则化是在损失函数中加入一个惩罚项λ，来减少模型的复杂度。）

Q73. When does linear regression line stop rotating or finds an optimal spot where it is fitted on data?
A place where highest RSquared value is found, is the place where the line comes to rest.

RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

Q74. How does SVM deal with self-learning? – SVM 如何处理自学习?

SVM has a learning rate and expansion rate which take care of this. The learning rate compensates or penalizes the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

（SVM 有一个学习率和扩展率来处理这个问题。学习率补偿或惩罚超平面做出所有错误的移动，而扩展率处理寻找类之间的最大分离区域。

补充：self-learning：

假设我们有标签数据 X1,y1 和无标签数据 X2;

1)首先利用有标签数据训练一个模型 Model1；

2）利用模型对无标签数据预测，得到无标签数据的概率；

3）设定一个概率阈值（如 0.8），将标签为 1 的概率大于阈值的无标签样本打上 1 的标签，同理打上 0 的标签，并从无标签的数据中剔除；

4）将 3 中打上标签的无标签样本加入训练集，重新训练得到模型；

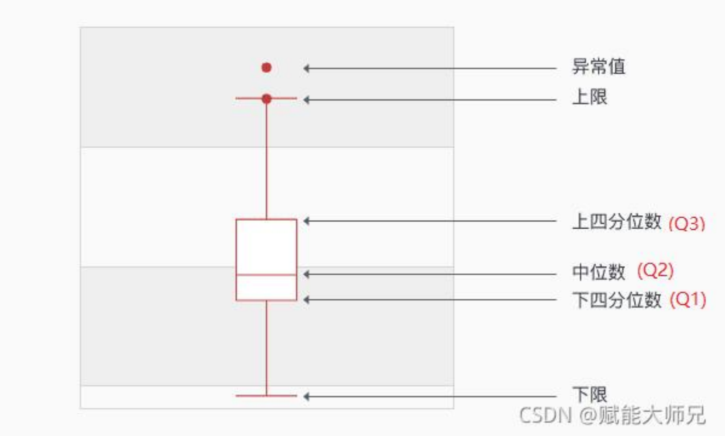
5）重复步骤 2-4，知道数据集为空或是达到迭代阈值。

Q75. How do you handle outliers in the data? – 如何处理异常值?

Outlier is an observation in the dataset that is far away from other observations in the dataset. We can discover outliers using tools and functions like box plot, scatter, Z-score, IQR score etc. and then handle them based on the visualization we have got. To handle outliers, we can cap at some threshold, use transformations to reduce skewness of the data and remove outliers if they are anomalies or errors.

（判别是否存在异常值的方法:

1. 箱型图 box plot：



)

下四分位数 Q1：等于所有样本按由小到大排列后第 25%的数字，即 Q1 线以下的数据为总数据的 25%

中位数 Q2：等于所有样本按由小到大排列后第 50%的数字，Q2 线上下各含有总数居 50%的数据

上四分位数 Q3：等于所有样本按由小到大排列后第 75%的数字，Q3 线段、以下含有总数据的 75%。

四分位距 IQR：上四分位数与下四分位数的差 Q3-Q1。

上限：Q3+1.5IQR，指的是非异常范围的最大值

下限：Q1-1.5IQR，指的是非异常范围内的最小值，大于 Q3+1.5IQR 或是小于 Q1-1.5IQR 的被称为内限，而在 Q3+3IQR 或是 Q1-3IQR 则被称为外限。

异常值：超出上限和下限的数据被称为异常值，在内限与外限之间的是温和异常值，而超出外限的数据被称为极端异常值。

处理异常值的方法：

1. 删除：直接将含有异常值的记录删除。
2. 视为缺失值，利用处理缺失值的方法来处理
3. 平均值修正：可以用前后连续观察的平均值来修正该异常值
4. 缩尾盖帽法

Q76. Name and define techniques used to find similarities in the recommendation system – I 列出在推荐系统中寻找相似性的方法

Person correlation and Cosine correlation are techniques used to find similarities in recommendation system.

Q77. Why would you prune your tree? – 决策树为何需要剪枝？

Pruning refers to the process of reducing redundant branches of a decision tree. Decision trees are prone to overfitting, pruning the tree helps to reduce the size and minimize the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the trade off.

（剪枝是为了减少一些属性分类实际上并不存在关联关系的情况，可以有效防止决策树出现过拟合现象。）

Q78. Mention some of EDA techniques? – 列举探索性数据分析的方法

1. Visualization:
   a) Univariate visualization
   b) Bivariate visualization
   c) Multivariate visualization
2. Missing value treatment – replace missing values with either mean/median
3. Outlier detection – use boxplot to identify the distribution of outliers, then apply IQR to set the boundary for IQR.

（1. 单变量/双变量/多变量可视化

（2. 缺失值处理：使用平均值或是中位数替换缺失值

（3. 异常值检测：使用厢式图画出异常值分布，再使用 IQR 设置上下界）

Q79. What is Inductive Logic Programming in machine learning?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logic programming representing background knowledge and examples.

Q80. What is data augmentation? – 什么是数据增强？

Data augmentation is a technique for synthesizing new data by modifying existing data in such a way that the target is not changed, or it is changed in a known way.

Usually used techniques:

1. Resize
2. Horizontal/Vertical Flip
3. Rotate
4. Add noise
5. Deform
6. Modify colors

（数据增强就是在不实质增加数据的情况下，让有限的数据产生等价于更多数据的价值。常用的方法有：调整大小，水平/垂直翻转，旋转，添加噪音，变形，修改颜色）

Q81. What is the difference between inductive machine learning and deductive machine learning? –

The difference between inductive machine learning and deductive machine learning are as follows: machine learning where the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn.

Q82. Difference between machine learning and deep learning? – 机器学习和深度学习区别?

Machine learning is a branch of computer science and a method to implement artificial intelligence. This technique provides the ability to automatically learn and improve from experience without being explicitly programmed.

Deep learning can be said as a subset of machine learning. It is mainly based on the artificial neural network where data is taken as an input and the technique makes intuitive decisions using the artificial neural network.

（机器学习是计算机科学的一个分支，是实现人工智能的一种方法。 这种技术提供了自动学习和从经验中改进的能力，而无需明确编程。

深度学习可以说是机器学习的一个子集。 它主要基于人工神经网络，其中数据作为输入，该技术使用人工神经网络做出直观的决策。）

Q83. What are the steps involved in machine learning project?

As you plan for doing a machine learning project. There are several important steps you must follow to achieve a good working model and they are data collection, data preparation, choosing a machine learning model, training the model, model evaluation, parameter tuning and lastly prediction.

Q84. Difference between Artificial Intelligence and machine learning? – 人工智能和机器学习区别?

Artificial Intelligence is a broader prospect than machine learning. Artificial intelligence mimics the cognitive function of the human brain. The purpose of AI is to carry out a task in an intelligent manner based on algorithms. On the other hand, machine learning is a subclass of artificial intelligence. To develop an autonomous machine in such a way so that it can learn without being explicitly programmed is the goal of machine learning.

（人工智能的定义范围比机器学习更大。 人工智能模仿人脑的认知功能。 人工智能的目的

是基于算法以智能方式执行任务。 而机器学习是人工智能的一个子类。 以这种方式开发自主机器，使其无需明确编程即可学习是机器学习的目标。）

Q85. Steps needed to choose the appropriate machine learning algorithm for your classification problem.

Firstly, you need to have a clear picture of your data, your constraints and your problems before heading towards different machine learning algorithms. Secondly, you have to understand which type and kind of data you have because it plays a primary role in deciding which algorithm you have to use.

Following this step is that data categorization step, which is a two-step process – categorization by input and categorization by output. The next step is to understand your constraints; that is, what is your data storage capacity? How fast the prediction has to be? Etc.

Finally, find the available machine learning algorithm and implement them wisely. Along with that, also try to optimize the hyperparameters which can be done in three ways – grid search, random search and Bayesian optimization.

Q86. Explain Backpropagation in machine learning – 解释反向传播算法

Backpropagation is the algorithm for computing artificial neural network. It is used by the gradient descent optimization that exploits the chain rule. By calculating the gradient of the loss function, the weight of the neurons is adjusted to a certain value. To train a multi-layered neural network is the prime motivation of backpropagation so that it can learn the appropriate internal demonstrations. This will help them learn to map any input to its respective output arbitrarily.

（反向传播算法是一种与最优化方法（如梯度下降法）结合使用的，用来训练人工神经网络的常见方法，该方法对网络中所有权重计算损失函数的梯度，这个梯度会反馈给最优化方法，用来更新权值以最小化损失函数。）

Q87. What is convex function? – 什么是凸函数？

A convex function is a continuous function, and the value of the midpoint at every interval in its given domain is less than the numerical mean of the values at the two ends of the internal.

（对于任意 t∈[0,1]，均满足$f(tx_1 + (1-t)x_2) \le f(tx_1) + f((1-t)x_2)$)，则称$f(x)$为凸函数。几何图像如下：

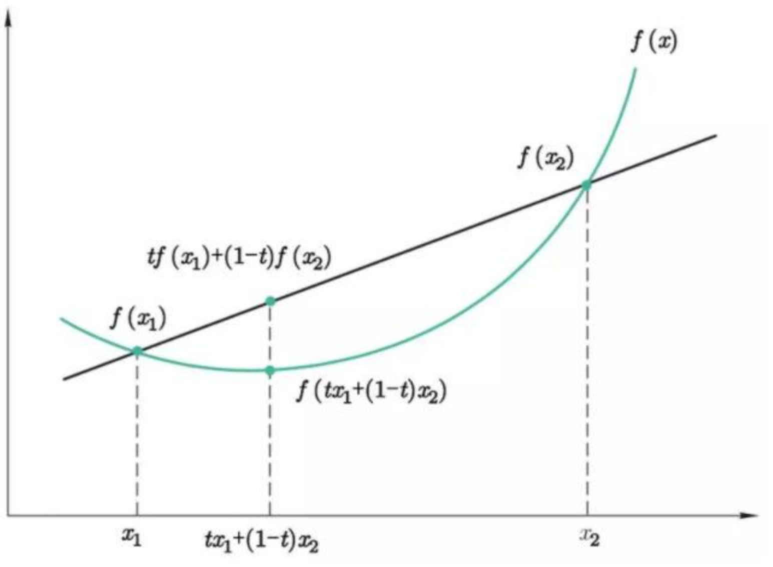


图1： 凸函数的割线在函数曲线上方

在机器学习的模型训练中，如果优化的目标函数为凸函数，则局部极小值就是全局最小值，即全局最优解，不会陷入到局部最优值当中。）

Q88. What is the relationship between True Positive Rate and Recall? – 真阳率和召回率之间的关系？

The True Positive Rate in machine learning is the percentage of the positives that have been acknowledge properly, and the recall is the count of the results that have been correctly identified. Therefore, TPR and Recall are the same things.

（真阳率和召回率是同一种概念，都是样本中被正确预测的正类占原本的正类的比率，即 $\frac{TP}{TP+FN}$）

Q89. List some tools for parallelizing machine learning algorithms

Almost all machine learning algorithms are easy to serialize. Some of the basic tools for parallelizing are Matlab, Weka, R, Python-based scikit learn.

Q90. What do you mean by genetic programming? – 什么是遗传编程？

Genetic Programming is almost similar to an Evolutionary Algorithm, a subset of machine learning. Genetic programming software systems implement an algorithm that uses random mutation, a fitness function, crossover, and multiple generations of evolution to resolve a user-defined task. The genetic programming model based on testing and choosing the best option among a set of results.

（遗传编程几乎类似于进化算法，是机器学习的一个子集。 遗传编程软件系统实现了一种算法，该算法使用随机变异、适应度函数、交叉和多代进化来解决用户定义的任务。 基于测试并从一组结果中选择最佳选项的遗传编程模型。）

Q91. What do you know about Bayesian networks?

Bayesian networks also referred to as belief network or casual network, are used to represent the graphical model for probability relationship among a set of variables. Efficient algorithms can perform inference or learning in Bayesian networks. Bayesian networks which relate the variables are called dynamic Bayesian networks.

Q92. Which are the two components of the Bayesian logic program?

A Bayesian logic program consists of two components:

1. Logical: it contains a set of Bayesian Clauses, which capture the qualitative structure of the domain
2. Quantitative: it is used to encode quantitative information about the domain.

Q93. How is machine learning used in day-to-day life?

Most of the people are already using machine learning in their everyday life. Assume that you are engaging with the Internet, you are actually expressing your preferences, likes, dislikes through your searches. All these things are picked up by cookies coming on your computer, from this, the behavior of a user is evaluated. It helps to increase the progress of a user through the Internet and provide similar suggestions.

The navigation system can also be considered as one of the examples where we are using machine learning to calculate a distance between two places using optimization techniques. Surely, people are going to more engage with machine learning in the near future.

Q94. Define Sampling. Why do we need it? – 什么是采样?

Sampling is a process of choosing a subset from a target population that would serve as its representative. We use the data from the sample to understand the pattern in the community as a whole. Sampling is necessary because often, we can not gather or process the complete data within a reasonable time.

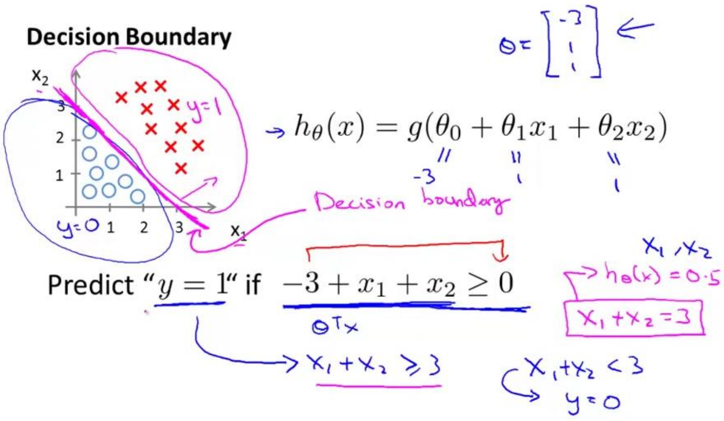
（抽样是从目标人群中选择一个子集作为其代表的过程。 我们使用样本中的数据来了解整个社区的模式。采样是必要的，因为我们通常无法在合理的时间内收集或处理完整的数据。）

Q95. What does the term decision boundary mean? – 什么是决策边界?

A decision boundary is a hypersurface which divides the underlying feature space into two subspaces, one for each class. If the decision boundary is a hyperplane, then the classes are linearly separable.

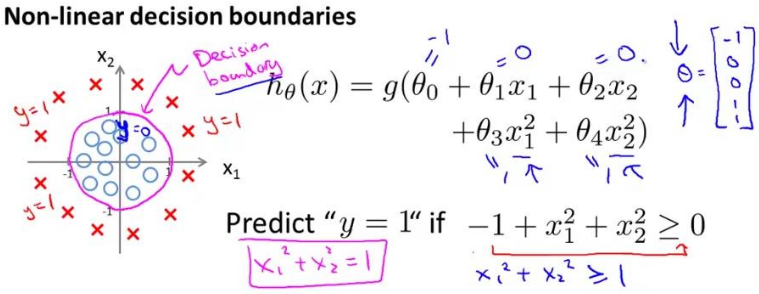
（决策边界是一个超曲面，它将基础特征空间划分为两个子空间，每个子空间对应一个类别。如果决策边界是超平面，则类是线性可分的。决策边界主要有线性决策边界和非线性决策边界。

线性决策边界：


Decision Boundary
x2
x1
y=1
y=0
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$
-3
1
1
$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
Decision boundary
Predict "y = 1" if $-3 + x_1 + x_2 \geq 0$
$\theta^T x$
$x_1 + x_2 \geq 3$
x_1, x_2
$h_\theta(x) = 0.5$
$x_1 + x_2 = 3$
$x_1 + x_2 < 3$
$y = 0$

)

非线性决策边界：


Non-linear decision boundaries
x2
x1
y=1
y=0
Decision boundary
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$
-1
=0
=0
1
1
$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$
Predict "y = 1" if $-1 + x_1^2 + x_2^2 \geq 0$
$x_1^2 + x_2^2 = 1$
$x_1^2 + x_2^2 \geq 1$

Q96. Define entropy – 定义熵

Entropy is the measure of uncertainty associated with random variable Y. It is the expected number of bits required to communicate the value of variable.

1. 信息熵 information entropy： 信息熵表示随机变量不确定性的度量，是对所有可能发生的事件产生的信息量的期望。信息熵的定义如下：

$$H(X) = -\sum_{x} p(x)logp(x) = -\sum_{i=1}^{n} p(x_i)logp(x_i)$$

x 为随机变量，$H(X)$为信息熵。随机变量的取值越多，信息熵就越大。

2. 条件熵 conditional entropy：条件熵表示在已知随机变量 X 的条件下随机变量 Y 的不确定性。条件熵的定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望：

$$H(Y|X) = -\sum_{x} p(x)H(Y|X = x) = -\sum_{x,y} p(x,y)logp(y|x)$$

3. 相对熵 relative entropy：相对熵又称为 KL 散度(Kullback-Leibler divergence)。相对熵用于同一个随机变量 x 的两个分布 p(x)和 q(x)之间的差异。在机器学习中，p(x)常用于描述样本的真实分布，例如[1,0,0,0]表示样本属于第一类，而 q(x)则用于表示预测的分布 ，例如[0.7,0.1,0.1,0.1]。显然使用 q(x 来描述样本不如 p(x)来得准确，q(x)需要不断学习来拟合准确的分布 p(x)。KL 散度的定义如下：

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i)log(\frac{p(x_i)}{q(x_i)})$$

n 表示事件可能发生的情况总数，KL 散度越小表示两个分布越接近。

4. 交叉熵 cross entropy：将 KL 散度公式进行变形，得到：

$$D_{KL}(p||q)=\sum_{i=1}^{n}p(x_i)log(p(x_i))-\sum_{i=1}^{n}p(x_i)log(q(x_i))$$

$$=-H(p(x))+[-\sum_{i=1}^{n}p(x_i)log(q(x_i))]$$

前半部分为 p(x)的熵，后半部分就是交叉熵：

$$H(p,q)=-\sum_{i=1}^{n}p(x_i)log(q(x_i))$$

在机器学习中，我们常常使用 KL 散度来评估 predict 和 label 之间的差异，但由于 KL 散度前半部分即 p(x)的熵是一个常量，所以常常将后半部分的交叉熵作为损失函数。

5. 交叉熵作为损失函数的原因

在回归问题中，均方误差(MSE)作为损失函数，其定义如下：

$$loss=\frac{1}{2m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

因为回归问题要求拟合实际的值，通过 MSE 能够衡量真实值和预测值之间的误差，再通过梯度下降算法来优化。而分类问题则需要一系列激活函数来将预测值映射到[0,1]之间，而激活函数的均方误差会变成一个非凸函数，存在多个局部极值点，要找出全局最优解效率很低。另一个原因就是激活函数的导数有一部分为 0，如果该激活函数的输入区域对应的导数为 0，则出现梯度消失问题，梯度无法更新。而使用交叉熵作为损失函数的话，梯度中含有的是激活函数的值与实际值之间的差，是一种线性函数，梯度的更新强度和绝对误差值正比，优化过程高效。）

Q97. Indicate the top intents of machine learning

1. The system gets information from the already established computations to give well-founded decisions and outputs.
2. It locates certain patterns in the data and then makes certain predictions on it to provide answers on matters.

Q98. Highlight the differences between the Generative model and the Discriminative model – 生成模型和判别模型之间的区别

The aim of the Generative model is to generate new sample from the same distribution and new data instances, whereas, the Discriminative model highlights the difference between different kinds of data instances. It tries to learn directly from the data and then classifies the data.

（生成模型的目的是从相同的分布和新的数据实例中生成新的样本，而判别模型强调不同种类的数据实例之间的差异。 它试图直接从数据中学习，然后对数据进行分类。 **参考 Q17**）

Q99. Identify the most important aptitudes of a machine learning engineer?

Machine learning allows the computer to learn itself without being decidedly programmed. It helps the system to learn from experience and then improve from its mistakes. The intelligence system, which is based on machine learning, can learn from recorded data and past incidents. In-depth knowledge of statistics, probability, data modeling, programming language, as well as CS, Application of ML Libraries and algorithms, and software design is required to become a successful machine learning engineer.

Q100. What is feature engineering? How do you apply it in the process of modeling? – 什么是特征工程?

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data

（特征工程是将原始数据转换为特征的过程，这些特征可以更好地代表预测模型的潜在问题，从而提高模型在新的数据上的准确性）

Q101. How can learning curves help create a better model? – 学习曲线如何帮助更好地构筑模型?

Learning curves give the indication of the presence of overfitting or underfitting. In a learning curve, the training error and cross-validating error are plotted against the number of training data points.

（学习曲线指示是否存在过拟合或欠拟合。 在学习曲线中，训练误差和交叉验证误差根据训练数据点的数量绘制。）