**Machine Learning**

**1. 机器学习中防止过拟合的方法有哪些？**

过拟合的原因是算法的学习能力过强，一些假设条件（如样本独立同分布）可能是不成立的；训练样本过少不能对整个空间进行分布估计。

处理方法有：

①早停止：如在训练中多次迭代后发现模型性能没有显著提高就停止训练；

②数据集扩增：原有数据增加、原有数据加随机噪声、重采样；

③正则化；

④交叉验证；

⑤特征选择/特征降维。

**2.谈谈判别式模型和生成式模型？**

判别方法：由数据直接学习决策函数 Y = f（X），或者由条件分布概率 P（Y|X）作为预测模型，即判别模型。常见的判别模型有：K近邻、SVM、决策树、感知机、线性判别分析（LDA）、线性回归、传统的神经网络、逻辑斯蒂回归、boosting、条件随机场。

生成方法：由数据学习联合概率密度分布函数P（X,Y）,然后求出条件概率分布P(Y|X)作为预测的模型，即生成模型。 由生成模型可以得到判别模型，但由判别模型得不到生成模型。 常见的生成模型有：朴素贝叶斯、隐马尔可夫模型、高斯混合模型、文档主题生成模型（LDA）、限制玻尔兹曼机。

**3. L1和L2的异同点在哪里？**

L1范数：指向量中各个元素绝对值之和，也有个美称叫“稀疏规则算子” 。

L2范数：为x向量各个元素平方和的1/2次方，L2范数又称Euclidean范数或Frobenius范数。

相同点：都用于避免过拟合。

不同点：L1可以让一部分特征的系数缩小到0，从而间接实现特征选择。所以L1适用于特征之间有关联的情况。L2让所有特征的系数都缩小，但是不会减为0，它会使优化求解稳定快速。所以L2适用于特征之间没有关联的情况。L1范数可以使权值稀疏，方便特征提取。L2范数可以防止过拟合，提升模型的泛化能力。

**4. L1和L2正则先验分别服从什么分布**

L1是拉普拉斯分布，L2是高斯分布。

**5. 归一化的作用。**

归一化：对不同特征维度的伸缩变换的目的是使各个特征维度对目标函数的影响权重是一致的，即使得那些扁平分布的数据伸缩变换成类圆形。这也就改变了原始数据的一个分布。

好处：

①提高迭代求解的收敛速度；

②提高迭代求解的精度；

③深度学习中数据归一化可以防止模型梯度爆炸。

**6. 常见的归一化方法有哪些？**

①min-max标准化：

②z-score 标准化：

**7. 哪些机器学习算法不需要做归一化处理？**

概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、随机森林。而像AdaBoost、SVM、LR、KNN、K-Means之类的最优化问题就需要归一化。

**8. 线性分类器与非线性分类器的区别以及优劣。**

区别：如果模型是参数的线性函数，并且存在线性分类面，那么就是线性分类器，否则不是。

常见的线性分类器有：LR,贝叶斯分类，单层感知机、线性回归；

常见的非线性分类器：决策树、RF、GBDT、多层感知机；

SVM两种都有(看线性核还是高斯核)

优劣势：线性分类器速度快、编程方便，但是可能拟合效果不会很好；非线性分类器编程复杂，但是效果拟合能力强。

**9. 常见的损失函数有哪些？**

①0-1损失函数和绝对值损失函数

0-1损失函数为：



绝对值损失函数为：



②log对数损失函数



③平方损失函数



④指数损失函数



⑤hinge损失函数



**10. 标准化与归一化的区别。**

简单来说，标准化是依照特征矩阵的列处理数据，其通过求z-score的方法，将样本的特征值转换到同一量纲下。归一化是依照特征矩阵的行处理数据，其目的在于样本向量在点乘运算或其他核函数计算相似性时，拥有统一的标准，也就是说都转化为“单位向量”。

**11. 数据预处理的方法有哪些？**

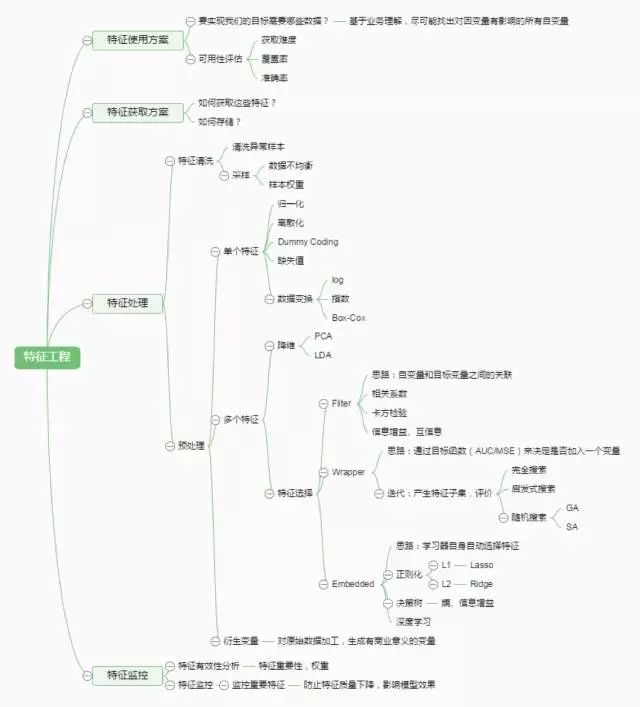
①缺失值，填充缺失值：离散，None；连续，均值；缺失值太多，则直接去除该列；

②连续值：离散化。有的模型（如决策树）需要离散值

③对定量特征二值化。核心在于设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0。

④皮尔逊相关系数，去除高度相关的列。

**12. 你知道有哪些数据处理和特征工程的处理**



**13. 数据不平衡问题是什么？如何解决？**

这主要是由于数据分布不平衡造成的。解决方法如下：

①进行特殊的加权，如在Adaboost中或者SVM中；

②采用对不平衡数据集不敏感的算法；

③改变评价标准：用AUC/ROC来进行评价；

④采用Bagging/Boosting/ensemble等方法；

⑤采样，对小样本加噪声采样，对大样本进行下采样；

⑥考虑数据的先验分布。

**14. 常见的分类算法有哪些？**

SVM、神经网络、随机森林、逻辑回归、KNN、贝叶斯。

**15. 常见的监督学习算法有哪些？常见的无监督学习算法有哪些？**

①监督学习算法：感知机、svm、人工神经网络、决策树、逻辑回归。

②无监督学习：K-Means、PCA。

**16. 欠拟合和过拟合的原因分别有哪些？如何避免？**

①欠拟合的原因：模型复杂度过低，不能很好的拟合所有的数据，训练误差大；

避免欠拟合：增加模型复杂度，如采用高阶模型（预测）或者引入更多特征（分类）等。

②过拟合的原因：模型复杂度过高，训练数据过少，训练误差小，测试误差大；

避免过拟合：降低模型复杂度，如加上正则惩罚项，如L1，L2，增加训练数据等。

**17. 正则化为什么能防止过拟合？**

过拟合表现在训练数据上的误差非常小，而在测试数据上误差反而增大。其原因一般是模型过于复杂，过分得去拟合数据的噪声. 正则化则是对模型参数添加先验，使得模型复杂度较小，对于噪声的输入扰动相对较小。

正则化时，相当于是给模型参数w 添加了一个协方差为1/lambda 的零均值高斯分布先验。 对于lambda =0，也就是不添加正则化约束，则相当于参数的高斯先验分布有着无穷大的协方差，那么这个先验约束则会非常弱，模型为了拟合所有的训练数据，w可以变得任意大不稳定。lambda越大，表明先验的高斯协方差越小，模型约稳定， 相对的variance(方差)也越小。

**18. 特征选择方法有哪些？**

①Filter方法

其主要思想是：对每一维的特征“打分”，即给每一维的特征赋予权重，这样的权重就代表着该维特征的重要性，然后依据权重排序。

主要的方法有：

Chi-squared test(卡方检验)；

information gain(信息增益)；

correlation coefficient scores(相关系数)。

②Wrapper方法

其主要思想是：将子集的选择看作是一个搜索寻优问题，生成不同的组合，对组合进行评价，再与其他的组合进行比较。这样就将子集的选择看作是一个是一个优化问题，这里有很多的优化算法可以解决，尤其是一些启发式的优化算法，如GA，PSO，DE，ABC等，详见“优化算法——人工蜂群算法(ABC)”，“优化算法——粒子群算法(PSO)”。

主要方法有：recursive feature elimination algorithm(递归特征消除算法)。

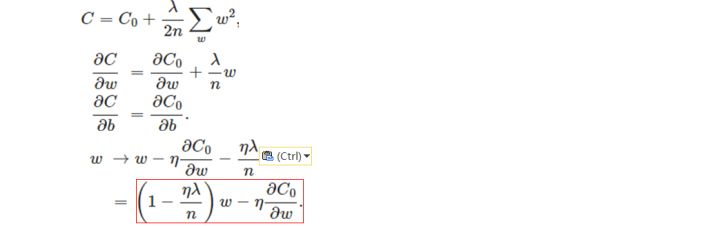
③Embedded方法

其主要思想是：在模型既定的情况下学习出对提高模型准确性最好的属性。这句话并不是很好理解，其实是讲在确定模型的过程中，挑选出那些对模型的训练有重要意义的属性。

主要方法：正则化，可以见“简单易学的机器学习算法——岭回归(Ridge Regression)”，岭回归就是在基本线性回归的过程中加入了正则项。

**19. L1为什么可以保证稀疏？**

假设原先损失函数是C0，那么在L2和L1正则条件下对参数求导分别是：



可以想象用梯度下降的方法，当w小于1的时候，L2正则项的惩罚效果越来越小，L1正则项惩罚效果依然很大，L1可以惩罚到0，而L2很难。

假设有两个完全一样的特征，使用L2正则项的话，两个特征权重相等的时候惩罚最小，所以L2具有权重平均分配的效果。

**20. 模型评价指标，解释AUC，准确率和召回率。**

True Positive(真正, TP)：将正类预测为正类数.

True Negative(真负 , TN)：将负类预测为负类数.

False Positive(假正, FP)：将负类预测为正类数 误报 (Type I error).

False Negative(假负 , FN)：将正类预测为负类数 漏报 (Type II error).

①ROC曲线：接收者操作特征，ROC曲线上每个点反映着对同一信号刺激的感受性。

横轴：负正类率(FPR)特异度，划分实例中所有负例占所有负例的比例；FPR越大，预测正类中实际负类越多。



纵轴：真正类率(TPR)灵敏度;TPR越大，预测正类中实际正类越多。



AUC：ROC曲线下的面积，介于0.1和1之间。AUC作为数值可以直观的评价分类器的好坏，值越大越好。

首先AUC值是一个概率值，当你随机挑选一个正样本以及负样本，当前的分类算法根据计算得到的Score值将这个正样本排在负样本前面的概率就是AUC值，AUC值越大，当前分类算法越有可能将正样本排在负样本前面，从而能够更好地分类。

②准确率：样本中预测正确的比例。



③召回率：样本中的正例有多少被预测正确了。



④精确率：预测为正的样本中有多少是真正的正样本。



**21. 聚类算法的类型有哪些？**

①层次化聚类算法：连续不断的将最为相似的两个群组合并，来构造一个新的群组。这样从最开始的单一元素开始，最后我们能够把所有的元素都聚在一个大类里。典型的有BIRCH算法、CURE算法、CHAMELEON算法、Sequence data rough clustering算法、Between groups average算法、Furthest neighbor算法、Neares neighbor算法等。

算法流程：

将每个对象看作一类，计算两两之间的最小距离;

将距离最小的两个类合并成一个新类;

重新计算新类与所有类之间的距离;

重复2、3，直到所有类最后合并成一类。

②划分式聚类算法：预先指定聚类数目或聚类中心，反复迭代逐步降低目标函数误差值直至收敛，得到最终结果。K-means、K-modes-Huang、K-means-CP、MDS\_CLUSTER、Feature weighted fuzzy clustering、CLARANS等

经典K-means算法流程：

随机地选择k个对象，每个对象初始地代表了一个簇的中心;

对剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇;

重新计算每个簇的平均值，更新为新的簇中心;

不断重复2、3，直到准则函数收敛。

③基于模型的聚类算法：为每簇假定了一个模型，寻找数据对给定模型的最佳拟合，同一”类“的数据属于同一种概率分布，即假设数据是根据潜在的概率分布生成的。主要有基于统计学模型的方法和基于神经网络模型的方法，尤其以基于概率模型的方法居多。一个基于模型的算法可能通过构建反应数据点空间分布的密度函数来定位聚类。基于模型的聚类试图优化给定的数据和某些数据模型之间的适应性。

SOM神经网络算法：

该算法假设在输入对象中存在一些拓扑结构或顺序，可以实现从输入空间(n维)到输出平面(2维)的降维映射，其映射具有拓扑特征保持性质,与实际的大脑处理有很强的理论联系。

SOM网络包含输入层和输出层。输入层对应一个高维的输入向量，输出层由一系列组织在2维网格上的有序节点构成，输入节点与输出节点通过权重向量连接。学习过程中，找到与之距离最短的输出层单元，即获胜单元，对其更新。同时，将邻近区域的权值更新，使输出节点保持输入向量的拓扑特征。

算法流程：

网络初始化，对输出层每个节点权重赋初值;

将输入样本中随机选取输入向量，找到与输入向量距离最小的权重向量;

定义获胜单元，在获胜单元的邻近区域调整权重使其向输入向量靠拢;

提供新样本、进行训练;

收缩邻域半径、减小学习率、重复，直到小于允许值，输出聚类结果。

④基于密度聚类算法：只要邻近区域的密度(对象或数据点的数目)超过某个阈值，就继续聚类。擅于解决不规则形状的聚类问题，广泛应用于空间信息处理,SGC,GCHL，DBSCAN算法、OPTICS算法、DENCLUE算法。

DBSCAN：

对于集中区域效果较好，为了发现任意形状的簇，这类方法将簇看做是数据空间中被低密度区域分割开的稠密对象区域;一种基于高密度连通区域的基于密度的聚类方法，该算法将具有足够高密度的区域划分为簇，并在具有噪声的空间数据中发现任意形状的簇。

⑤基于网格的聚类算法：把对象空间量化为有限数目的单元，形成一个网格结构。所有的聚类操作都在这个网格结构(即量化空间)上进行。这种方法的主要优点是它的处理 速度很快，其处理速度独立于数据对象的数目，只与量化空间中每一维的单元数目有关。但这种算法效率的提高是以聚类结果的精确性为代价的。经常与基于密度的算法结合使用。代表算法有STING算法、CLIQUE算法、WAVE-CLUSTER算法等。