**Machine Learning**

**1. 如何理解判别式模型和生成式模型？**

1. 判别方法：由数据直接学习决策函数，或者由条件分布概率作为预测模型，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。
2. 生成方法：由数据学习联合概率密度分布函数,然后求出条件概率分布作为预测的模型，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.

**2. 有监督学习和无监督学习分别是什么？**

1. 监督学习的训练集要求包括输入输出，也可以说是特征和目标。训练集中的目标是由人标注的。
2. 输入数据没有被标记，也没有确定的结果，样本数据类别未知，需要根据样本间的相似性对样本集进行分类。

**3. 归一化是什么？有什么作用？常见的归一化方法有哪些？哪些机器学习模型需要进行归一化？**

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. 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。
2. 作用：
3. 提高迭代求解的收敛速度；
4. 提高迭代求解的精度；
5. 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：
7. min-max标准化：
8. z-score 标准化：
9. 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。

**4. 线性分类器与非线性分类器的区别以及优劣。**

1. 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。
2. 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）。

1. 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。

**5. 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. 删除共线性变量：皮尔逊相关系数，去除高度相关的列。

**6. 数据不平衡问题是什么？如何解决？**

数据不平衡是指数据集中，每个类别下的样本数目相差很大。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。

**7. 多重共线性是什么？如何减小多重共线性对模型效果的影响？**

**8. 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。

1. 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。

**9. 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。

**10. 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。

**11. 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。

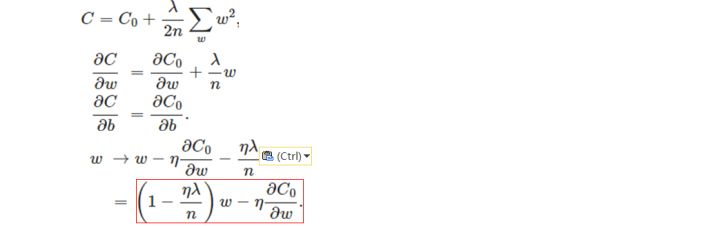
不同点：L1可以让一部分特征的系数缩小到0，从而间接实现特征选择。所以L1适用于特征之间有关联的情况。L2让所有特征的系数都缩小，但是不会减为0，它会使优化求解稳定快速。所以L2适用于特征之间没有关联的情况。L1范数可以使权值稀疏，方便特征提取。L2范数可以防止过拟合，提升模型的泛化能力。

**12. 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。

**13. 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。

**14. 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。

**15. 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。

