机器学习解决问题的步骤：

收集数据

数据预处理

构建特征

处理特征

选择模型

数据集划分

训练模型

交叉验证调参得到正确率

1. 数据预处理

收集到数据之后，把数据标准化，然后去掉在预设范围内的极端值。如果数据中有需要数据也有不需要的数据，就要把不需要的数据剔除之后再把这些数据对有用数据的影响去掉。简单的做法是对不需要的数据做一个回归，得到一个新的数据A，再用每个有用数据减去A就得到了去掉无用数据影响后的有用数据。

1. 构建特征

用处理过的有用数据构建特征向量。对特征向量还要进一步进行处理，比如对缺省值进行处理、对名称型数据的处理、对数值型数据的处理、对连续型数据的处理等。

对缺省值处理的方式有几种：如果缺省值的比例比较大，可以直接去掉这个特征；如果缺省值的占比很小而且是有限个类别的离散型数据，则可以用数量最多的类别去标注缺省值，也可以把缺省当作一个类别；而如果是连续性数据，则可以用先把该特征当作一个目标向量，用其他的特征去回归一个曲线，这样把缺省的值填上。

对名称性数据的处理是对名称进行编码，比如one hot编码，把第一类、第二类、第三类编码成001、010、100等。也可以直接用1、2、3。

如果数据是连续的，则要设定几个阈值把数据变成离散的。设定阈值时一定要根据数据的本身物理意义进行设定。

总之，对特征的处理过程就是：连续型->离散型->编码。名称型也算作离散型。也可以不进行编码，这要看后面的效果怎么样。

1. 处理特征（特征选择）

对特征的处理就是对特征的筛选，要筛选出和目标值关系比较大的特征，也要去掉能用其他特征表示的特征。去掉无用特征的方法可以是在其他的特征中找到和该特征相似度最大的，如果相似度超过了一个阈值，则两个特征只能留一个，其中相似度的定义可以是MSE、也可以是向量余弦等等。

特征选择的方法有很多。有的是直接操作特征矩阵和目标向量，有的是要进行模型预测之后看结果在筛选。前一种和上面的描述很相似，找特征和目标向量的关系（线性、非线性），如果关系超过了一定阈值就留下，反之就抛弃。但是这个阈值的设定也是要看具体问题，有时候也是要走一遍模型才能知道留下的特征的好坏。

对特征的处理，无论是去掉相似特征还是进行特征选择，都要进行交叉验证，尤其是走模型的时候，不进行交叉验证就不能确定去掉某特征的操作是不是准确。

有时候对特征把握不准或者得不到很有用的特征的时候，可以先用聚类算法，让数据去表征一些特征，然后再用得到的特征进行下一步。

1. 选择模型

对模型的选择要看该问题的目标，是分类问题，还是回归问题。

分类问题有KNN、SVM、Logistic Regression、naive bayesian、贝叶斯网络、决策树、RF等等...

回归问题有Linear Regression、LASSO、RLS等等...

分类问题有一个通用的算法模型，就是神经网络。三个隐层的神经网络可以区分任意非线性的关系，深度越高的神经网络对复杂的情况解决的越好，但是要注意神经网络的过拟合现象。神经网络也要注意调参，尤其是BP网络的学习速率是重点调整的，过小了可能训练的很慢，过大了可能越过了最优值，得不到最优解。而且神经网络的各个连接的参数的初始值一定不能相同，一定要是一个很小的随机值。它的输出也要用one hot类型来表示，不能用1、2、3 。

1. 数据集划分

对进过处理后的特征矩阵进行划分，要划分出三部分，训练集、交叉验证集和测试集。训练集不宜过大，过大可能会引起训练时间过长和过拟合现象。如果问题的时间特性很强，则划分数据集时一定要按照时间去划分。

1. 训练模型和调参

用训练集去训练模型，用交叉验证集进行交叉验证进行调参。例如学习速率、迭代次数、特征选择、RLS中的遗忘因子等等参数。经过一系列的交叉验证调参之后得到最好的各项参数在用测试集去进行测试，观察最后的正确率等信息，如果不理想则还要继续调整算法，调整特征，调整各种，然后再去训练和调参。

1. Tips

拿到数据之后第一件必须要做的事是观察数据的分布，最好把数据可视化出来。根据数据的走势、分布来确定特征的选取、特征的处理、要不要进行聚类、用什么模型比较合适等等问题。