分类问题以离散随机变量或者离散随机变量的概率分布作为输出，回归问题以连续变量作为预测输出。

分类模型：图像分类、视频分类、文本分类、机器翻译、语音识别

回归模型：银行信贷评分、人脸/人体关键点估计、年龄估计、股市预测

回归问题与分类问题之间可以相互转化。

根据机器学习是否可用于生成新数据，可以将机器学习模型分为生成模型和判别模型。

生成模型：朴素贝叶斯分类器、高斯混合模型、隐马尔可夫模型、生成对抗网络

判别模型：逻辑回归、决策树、支持向量机、神经网络、k-近邻算法

首先将数据划分为训练集、验证集和测试集三个部分。验证集与训练集的划分是所有项目都需要使用的。

对于分类问题，KNN从所有训练样本集合中找到与x最近的k个样本，然后通过投票法选择这k个样本中出现次数最多的类别作为x的预测结果；对回归问题，找到与x最近的k个样本并对其标签求平均值，得到x的预测结果。

朴素贝叶斯分类器是一种有监督的统计学过滤器，在垃圾邮件过滤、信息检索等领域十分常用。

生成式对抗网络（GAN）是近些年计算机视觉领域非常常见的一种方法。它可以从已有数据集中生成新数据，能力强大到令人惊叹，甚至可以完成连人眼都无法进行分辨的一些任务。

监督任务是有明确的输入和输出来对模型进行优化调整，而这一项任务是基于已有的数据集生成新的与原有数据集相似的新的数据。这是一个典型的生成式任务，最重要的核心是如何获取这个分布。基于图像现有的生成式框架有VAE和GAN两大分支。

反卷积网络是GAN网络非常重要的一个部件。大多数卷积层会使特征图的尺寸不断变小，但反卷积层是为了使得特征图逐渐变大，甚至与最初的输入图片一致。

卷积神经网络中引入了卷积层进行卷积操作，引入了池化操作进行特征的聚合统计，引入了Dropout层防止过拟合。

工程数据的提取聚合和存储

在实际工程中，原始数据存储在MySQL业务库中。需要将业务库的数据通过Sqoop同步到基于Hadoop的Hive数据仓库中。基于Hive的数据仓库会对数据进行清洗聚合，进而实现信息的分层存储。ODS层又名临时存储层，这一层所做的工作是贴源。DW层又名数据仓库层，是将一些数据关联的日期进行拆分，将其更具体地分类，一般拆分成年、月、日。App层又名引用层，在实际线上生产中，训练模型所需要的数据就存储在这一层里面。

Sqoop是一款开源的工具，主要用于在Hadoop与传统的数据库间进行数据的传递。Hive是基于Hadoop的一个数据仓库工具。

采样分为上采样和下采样。上采样的原理是把数据量小的那一类样本重复多次，直到两类样本的数据量达到均衡，此时总体样本量变大。下采样的原理是利用随机采样，从数据量大和数据量少的样本集选出同样数量的样本作为进行模型训练的数据。