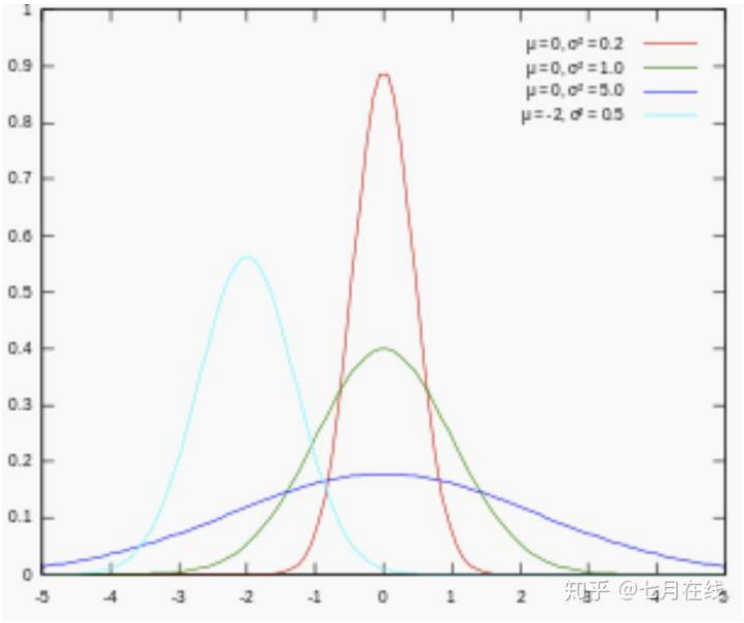
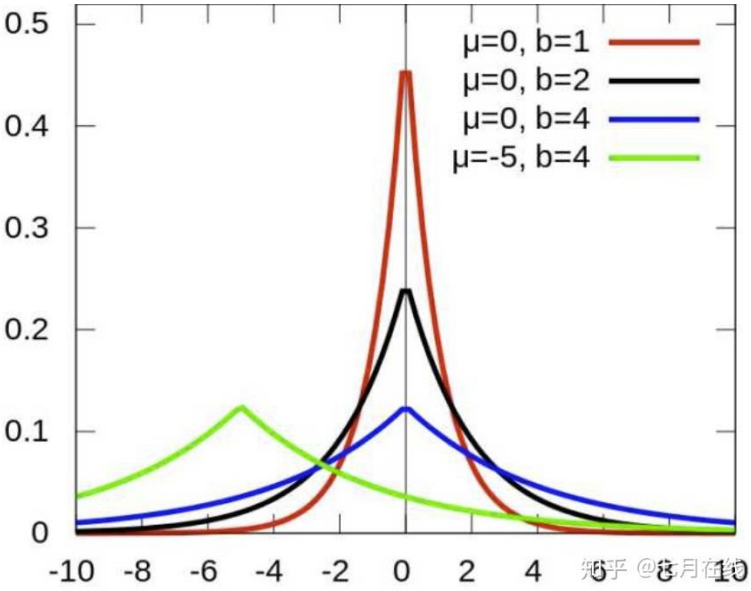
1	在模型评估过程中，过拟合和欠拟合具体指什么现象 **过拟合**是指模型在训练数据拟合呈过当的情况，反应到评估指标上，就是**模型在训练集上的表现很好，但在测试集和新数据上的表现很差**。 **欠拟合**指的是模型在**训练和预测时都不好的情况**。	如何描述这两个现象
2	降低过拟合和欠拟合的方法 **降低过拟合**风险的方法： **1.从数据入手，获得更多的训练数据。**使用更多的训练数据是解决过拟合问题最有效的手段，因为更多的样本能够让模型学习到更多更有效的特征，减少噪声的影响。当然，直接增加实验数据一般是很困难的，但是可以通过一定的规则来扩充训练数据。比如在图像分类问题上，可以通过图像的平移，旋转，缩放等方式扩充数据，更进一步地，可以使用生成式对抗网络来合成大量新训练数据。 **2.降低模型的复杂度。**在数据较少时，模型过于复杂是产生过拟合的主要因素，适当降低模型复杂化度可以避免模型拟合过多的采样噪声。例如在神经网络模型中减少网络层数，神经元个数等，在决策树模型中降低树的深度，进行剪枝。 **3.正则化方法。**给模型参数加上一定正则约束，比如将权值大小加入到损失函数总。 **4.集成学习方法。**集成学习是吧多个模型集成在一起，来降低单一模型过拟合风险，如 bagging 方法。 **降低欠拟合**风险的方法： **1.添加新特性。**当特征不足或者现有特征与样本标签的相关性不强时，模型容易出现欠拟合。通过挖掘上下文特征 ID 类特征、组合特征等新的特征，往往能够取得很好的效果。 **2.增加模型复杂度。**简单模型的学习能力较差，通过增加模型的复杂度可以使模型拥有更强的拟合能力。例如在线性模型中增加高次项，在神经网络模型中增加网络层数和神经元个数。 **3.减小正则化系数。**正则化是用来防止过拟合的，但当模型出现欠拟合现象时，则需要有针对性的减小正则化系数。	从多个维度来考虑，比如数据，特征，模型，目标函数等等

3	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**。对参数引入高斯正态先验分布相当于 L2 正则化。 $$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$  对参数引入拉普拉斯先验等价于 L1 正则化。 $$f(x\mid\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$  从上面两图可以看出，L2 先验趋向零周围，L1 先验趋向零本身。	可根据 L1 和 L2 正则项的数学表达式的形式来分析

4	对于树形结构为什么不需要归一化？ **因为数值缩放不影响分裂点位置，**对树模型的结构不造成影响。按照特征值进行排序的，排序的**顺序不变，那么所属的分支以及分裂点就不会有不同**。而且，**树模型是不能进行梯度下降的**，因为构建树模型（回归树）寻找最优点时是通过寻找最优分裂点完成的，因此树模型是阶跃的，阶跃点是不可导的，并且求导没意义，也就不需要归一化。	理解清楚特征归一化所适用的模型场景
5	什么是数据不平衡，如何解决？ **数据不平衡又称样本比例失衡，**比如二分类问题，如果标签为 1 的样本占总数的 99%，标签为 0 的样本占比 1%则会导致判断**「失误严重」**，准确率虚高。 常见的**解决不平衡问题的方法**如下。 1.**「数据采样」**数据采样分为上采样和下采样，**上采样是将少量的数据通过重复复制使得各类别比例均衡**，不过很容易导致过拟合问题，所以需要在新生成的数据中加入随机扰动。下采样则相反，**下采样是从多数类别中筛选出一部分从而使得各类别数据比例维持在正常水平**，但容易丢失比较重要的信息，所以应该多次随机下采样。 2.**「数据合成」**是利用已有样本的特征相似性**生成更多的样本**。 3.**「加权」**是通过不同类别的错误**施加不同的权重惩罚**，使得 ML 时更侧重样本较少并容易出错的样本。 4.**「一分类」**当正负样本比例失衡时候，可以利用 One-class SVM，该算法利用**「高斯核函数」**将样本空间映射到**「核空间」**，在核空间找到一个包含**「所有数据」**的高维球体。如果测试数据位于这个高维球体之**「中」**，则归为多数类，否则为少数类。	理解数据不平衡会给模型训练带来什么影响