知识点：算法评价

问题：对算法的评价包含哪些方面：

答案：

1. 泛化误差
2. 时间开销
3. 存储开销
4. 可解释性

知识点：泛化误差

问题：为什么测试误差可以近似泛化误差？

答案：

我们可通过实验测试来对学习器的泛化误差进行评估并进而做出选择，为此， 需使用一个"测试集" (testing set)来测试学习器对新样本的判别能力，然后以测试集上的"测试误差" (testing error)作为泛化误差的近似。通常我们假设测试样本也是从样本真实分布中独立同分布采样而得，但需注意的是，测试集应该尽可能与训练集互斥，即测试样本尽量不在训练集中出现、未在训练过程中使用过.

知识点：假设检验

问题：假设检验对于学习器性能评价的作用？

统计假设检验(hypothesis test)为我们进行学习器t性能比较提供了重要依据.基于假设检验结果我们可推断出，若在测试集上观察到学习器A 比B 好，则A 的泛化性能是否在统计意义上优于B ，以及这个结论的把握有多大.

假设检验中的"假设"是对学习器泛化错误率分布的某种判断或猜想，现实任务中我们并不知道学习器的泛化错误率，只能获知其测试错误率，泛化错误率与测试错误率未必相同，但直观上二者接近的可能性应比较大，相差很远的可能性比较小.因此， 可根据测试错误率估推出泛化错误率的分布.

知识点：偏差与方差

问题：如何理解泛化误差的分解？

答案：泛化误差可分解为偏差、方差与噪声之和.偏差度量了学习算法的期望预测与真实结果的偏离程度，即刻画了学习算法本身的拟合能力;方差度量了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响;噪声则表达了在当前任务上任何学习算法所能达到的期望泛化误

差的下界，即刻画了学习问题本身的难度.偏差一方差分解说明，泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度所共同决定的.给定学习任务?为了取得好的泛化性能，则需使偏差较小，即能够充分拟合数据，并且使方差较小，即使得数据扰动产生的影响小.一般来说偏差与方差是有冲突的，这称为偏差-方差窘境(bias-variance dilemma).给定学习任务，假定我们能控制学习算法的训练程度，则在训练不足时?学习器的拟合能力不够强，训练数据的扰动不足以便学习器产生显著变化，此时偏差主导了泛化错误率;随着训练程度的加深，

学习器的拟合能力逐渐增强，训练数据发生的扰动渐渐能被学习器学到，方差逐渐主导了泛化错误率;在训练程度充足后，学习器的拟合能力已非常强，训练数据发生的轻微扰动都会导致学习器发生显著变化，若训练数据自身的、非全局的特性被学习器学到了，则将发生过拟合.

知识点：综合

问题：简述没有免费午餐定理和意义

答案：

没有免费午餐定理（NFL），对于一个学习算法A，若它在某些问题上比学习算法B好，则必然存在另一些问题，在那里B比A好.，这个结论对任何算法均成立. 也就是说，无论学习算法A多聪明、学习算法B多笨拙，它们的期望性能是相同的。

NFL定理是让我们清楚地认识到，脱离具体问题，空泛地谈论"什么学习算法更好"毫无意义，因为若考虑所有潜在的问题，贝。所有学习算法都一样好.要谈论算法的相对优劣，必须要针对具体的学习问题;在某些问题上表现好的学习算法，在另一些问题上却可能不尽如人意，学习算法自身的归纳偏好与问题是否相配，往往会起到决定性的作用.

知识点：机器学习

问题：机器学习的研究内容：

参考答案：?机器学习所研究的主要内容，是关于在计算机上从数据中产生"模型" (model) 的算法，即"学习算法" (learning algorithm). 有了学习算法，我们把经验数据提供给它，它就能基于这些数据产生模型;在面对新的情况时(例如看到一个没剖开的西瓜)，模型会给我们提供相应的判断(例如好瓜) .如果说计算机科学是研究关于"算法"的学问，那么类似的，可以说机器学习是研究关于"学习算法"的学问.

**知识点：泛化误差的偏差方差分解**

问题: 什么是偏差（bias）和方差（variable）均衡？

答案：偏差指的是模型预测值与真实值的差异，是由使用的学习算法的某些错误或过于简单的假设造成的误差。它会导致模型欠拟合，很难有高的预测准确率。  
方差指的是不同训练数据训练的模型的预测值之间的差异，它是由于使用的算法模型过于复杂，导致对训练数据的变化十分敏感，这样会导致模型过拟合，使得模型带入了过多的噪音。  
任何算法的学习误差都可以分解成偏差、方差和噪音导致的固定误差。模型越复杂，会降低偏倚增加方差。为了降低整体的误差，我们需要对偏差方差均衡，使得模型中不会由高偏差或高方差。

**知识点：监督学习与非监督学习**

问题：监督学习和非监督学习有什么不同？

答案：监督学习需要具有标签（label）的训练数据，比如做分类，你需要先对训练数据做标记，然后才能训练模型将数据分成你说需要的标记类。  
而非监督学习则不需要。

**知识点：综合**

问题: KNN和k-means聚类由什么不同？

答案：k-Nearest Neighbors 是一种监督学习算法，而k-means 是非监督的。这两种算法看起来很相似，都需要计算样本之间的距离。knn算法需要事先已有标注好的数据，当你需要对未标注的数据进行分类时，统计它附近最近的k个样本，将其划分为样本数最多的类别中。k-means聚类只需要一些未分类的数据点和阀值，算法会逐渐将样本点进行分成族类。

**知识点：**ROC曲线，本题可以拓展为计算题

解释一下ROC曲线的原理

ROC曲线是真正率和假正率在不同的阀值下之间的图形表示关系。通常用作权衡模型的敏感度与模型对一个错误分类报警的概率。  
真正率表示：表示正的样本被预测为正占所有正样本的比例。  
假正率表示：表示负的样本被预测为正占所有负样本的比例。  
（0，0）点表示所有样本都被预测为负，此时阀值很大。  
（1，1）点表示所有样本都被预测为正，此时阀值很小。

**知识点：准确率，查准率，查全率，真正例率，假正例率**

问题：准确率，查准率，查全率，真正例率，假正例率的定义

什么是贝叶斯定理，它是如何使用在机器学习中的？

贝叶斯定理会根据一件事发生的先验知识告诉你它后验概率。数学上，它表示为：一个条件样本发生的真正率占真正率和假正率之和的比例，即：

举个例子： 已知某疾病的患病概率为5%，现用某检验方法进行诊断，若患有该病，则有90%的几率检验结果呈阳性。但即使正常人使用该检验方法，也有10%的几率误诊而呈阳性。某人检验结果为阳性，求此人患病的概率。

贝叶斯定理使一些机器学习算法如：朴素贝叶斯等的理论基础。

**知识点：朴素贝叶斯**

问题：为什么我们要称朴素贝叶斯？

答案：因为我们在用到它的时候，有一个很强的假设，现实数据中几乎不会出现的：我们假设特征之间是相互独立，也就是我们计算条件概率时可以简化成它的组件的条件概率乘积。

**知识点：hL1、L2正则之间有什么不同？**

答案：L2正则 对应的是加入2范数，使得对权重进行衰减，从而达到惩罚损失函数的目的，防止模型过拟合。保留显著减小损失函数方向上的权重，而对于那些对函数值影响不大的权重使其衰减接近于0。相当于加入一个gaussian prior。  
L1正则 对应得失加入1范数，同样可以防止过拟合。它会产生更稀疏的解，即会使得部分权重变为0，达到特征选择的效果。相当于加入了一个laplacean prior。

**知识点：综合**

问题：你最喜欢的算法是什么？把它解释一下。

开放答案

**知识点：极大似然法则**

问题：概率和似然有什么区别？

答案：概率和似然都是指可能性，但在统计学中，概率和似然有截然不同的用法。概率描述了已知参数时的随机变量的输出结果；似然则用来描述已知随机变量输出结果时，未知参数的可能取值。例如，对于“一枚正反对称的硬币上抛十次”这种事件，我们可以问硬币落地时十次都是正面向上的“概率”是多少；而对于“一枚硬币上抛十次，我们则可以问，这枚硬币正反面对称的“似然”程度是多少。  
概率(密度)表达给定θ下样本随机向量X=x的可能性，而似然表达了给定样本X=x下参数θ1(相对于另外的参数θ2)为真实值的可能性。我们总是对随机变量的取值谈概率，而在非贝叶斯统计的角度下，参数是一个实数而非随机变量，所以我们一般不谈一个参数的概率，而说似然。

**知识点：深度学习**

问题：什么是深度学习，它与机器学习算法之间有什么联系？

答案：深度学习是机器学习的一个子领域，它关心的是参照神经学科的理论构建神经网络，使用反向传播对大量未标注或半结构化的数据进行建模。从这个角度看，深度学习可以看成一种非监督学习算法，通过使用神经网络学习数据的表示。

**知识点：生成模型与判别模型**

问题：生成模型与判别模型有什么区别？

答案：生成模型会学习数据的分布；判别模型学习的是不同类型数据之间的区别，不学习数据内部特点。在分类问题上，判别模型会优于生成模型。  
判别模型求解的思路是：条件分布------>模型参数后验概率最大------->最大似然  
生成模型的求解思路是：联合分布------->求解类别先验概率和类别条件概率  
常见的生成方法有混合高斯模型、朴素贝叶斯法和隐形马尔科夫模型等，常见的判别方法有SVM、LR等

**知识点：决策树**

问题：如何对决策树进行剪枝？

答案：剪枝是决策树发生过拟合后，为了降低模型复杂度，提高模型准确率的一种做法。可以分为自上而下和自下而上两种。常见的方法有：误差降低剪枝（REP）和代价复杂度剪枝（CCP）。  
REP简单的来说就是对树的每一个结点进行剪枝，如果剪掉某个结点不会降低模型准确率，那么将其剪掉。这种启发式的做法实际上就是为了最大化准确率。

知识点：决策树

问题：在没有剪枝的决策树中，有哪几种叶子结点？这些叶子结点对应的类别分别是什么？

答案：

1. 属性集合为空，样本子集非空，叶子结点对应的类别是子集中多数样本的类别；
2. 属性集合非空，样本子集为空，叶子结点对应的类别是父结点的类别；
3. 属性集合不必考虑，如果样本子集的类别是一样的，叶子结点对应的类别就是类别。

知识点：决策树

问题：试证明对于不含冲突数据(即特征向量完全相同但标记不同)的训练集，必存在与训练集一致(即训练误差为 0) 的决策树

答案：

我们对某次判断作终止条件的分析：

（1）当前结点包含样本全属于同一类别；

（2）当前属性集为空，或所有样本在所有属性上取值相同；

（3）当前结点包含样本集为空。

显然，若判断终止，（2）时，若当前属性集为空则取标记多的为分类结果，无误差，若样本在所有属性上取值相同，因数据不冲突，故必为同类，等价于（1），（1）时，分类完毕，无误差，（3）回溯上一次判断的情况，必然属于（1）（2）（3）中一种。

故不含冲突数据的训练集必存在无训练误差决策树。

问题：如果不对决策树进行剪枝，可能会出现什么结果？

答案：

模型泛化能力较差。类似于多元回归模型，有些变量应该被剔除，决策树也存在过拟合问题，即不仅学习了数据的泛化特征，也学习了训练集本身特点。

**知识点：泛化能力**

许多机器学习的模型会有高的精度，但是预测能力也就是泛化能力较低。精度只是模型性能的一部分，有可能是会产生误导的那个。对于具有倾斜的数据集，比如要从大量的金融数据中识别出少量的诈骗数据，一个精度高的模型可能会告诉你没有诈骗，然而这样的模型预测是没有意义的。所以，不要总是把精度当作模型最重要的事。

**知识点：查准率，查全率和F1指标**

问题：什么是F1数，怎么使用它？

F1数是衡量模型性能的一个指标。它是模型精准率和召回率的加权平均，1表示最好，0表示最差。在分类问题中有时精准率和召回率不会同时都高，那么我们可以使用F1数。

**知识点：不平衡样本**

问题：如何处理一个不平衡的数据集？

答案：不平衡的数据集：比如二分类问题中，一类数据有90%，而另一类只有10%。我们可以轻易的得到90%准确率的模型，但是它对第二类的预测值为0。那么我们需要对这样的数据进行处理：

收集更多的数据，使其达到平衡

1. 使用重复采样
2. 使用不同的算法
3. 重要的是：你注意到了数据的不平衡导致的问题，以及如何去解决它们。

**知识点：分类和回归**

问题：什么时候你应该使用分类而不是回归？

分类会产生离散的数值，使得数据严格的分为不同类。回归会得到连续的值，使你更好的区分独立点之间的区别。当你需要知道你的数据明确的属于那些类时你可以用分类。

**知识点：集成学习**

问题：举例说明使用集成学习会很有用。

集成学习通过组合一些基学习算法来优化得到更好的预测性能，通常可以防止模型的过拟合使模型更具有鲁棒性。  
可以从bagging、boosting、stacking等算法中选一个例子，说明他们是如何增加模型预测能力的。

**知识点：过拟合**

问题：你如何确保你的模型没有过拟合？

答案：过度拟合的训练数据以及数据携带的噪音，对于测试数据会带来不确定的推测。有如下三种方法避免过拟合：

保持模型尽可能地简单：通过考量较少的变量和参数来减少方差，达到数据中消除部分噪音的效果。

使用交叉检验的手段如：k-folds cross-validation。

使用正则化的技术如：例如LASSO方法来惩罚模型中可能导致过拟合的参数。

**知识点：模型的评价标准**

问题：如何评估你的机器学习模型的有效性？

首先你需要将数据分成训练集和测试集，或者使用给交叉验证方法分割。然后你需要选择度量模型表现的metrics，如F1数、准确率、混淆矩阵等。更重要的是，根据实际情况你需要理解模型度量的轻微差别，以便于选择正确的度量标准。

**知识点：核技巧和核函数**

问题：什么是核技巧，有什么用处？

答案：核技巧使用核函数，确保在高维空间不需要明确计算点的坐标，而是计算数据的特征空间中的内积。这使其具有一个很有用的属性：更容易的计算高维空间中点的坐标。许多算法都可以表示称这样的内积形式，使用核技巧可以保证低维数据在高维空间中运用算法进行计算。

**知识点：缺失值处理**

问题：如何处理数据中的缺失值？

**知识点：误差理论**

问题：经验误差(empirical error)与泛化误差(generalization error)分别指？

答案：

经验误差: 也叫训练误差(training error)，模型在训练集上的误差。

泛化误差: 模型在新样本集(测试集)上的误差。

## 知识点：类别不平衡

问题：如何处理数据中的“类别不平衡”？例如在一个二分类问题中，训练集中80%分类为A，20%分类为B。

答案

简单方法

数据多的欠采样(under-sampling)，舍弃一部分数据，使其与较少类别的数据相

数据少的过采样(over-sampling)，即重复使用一部分数据，使其与较多类别的数据相当。

阈值调整（threshold moving，例如数据均衡时，阈值为0.5，那么可以按比例，例如调整到 0.8。

复杂方法

数据采样过程中，生成并插样“少数类别”数据，代表算法 SMOTE 和 ADASYN。

数据先聚类，“多数类别”随机欠采样，“少数类别”数据生成。

随机欠采样容易丢失重要信息，可结合集成学习欠采样，代表算法：EasyEnsemble。利用集成学习机制，将大类划分为若干个集合供不同的学习器使用。相当于对每个学习器欠采样，避免全局丢失重要信息。

**知识点：数据归一化**

问题：数据预处理阶段，对数值特征归一化或标准化，理论上不会对哪种模型产生很大影响。

A 决策树

B k-means

C kNN

​

答案

A k-means和kNN(k-NearestNeighbor)都需要使用距离。而决策树对于数值特征，只在乎其大小排序，而非绝对大小。不管是标准化或者归一化，都不会影响数值之间的相对大小。

## 知识点：神经网络的激活函数

某个神经网络的激活函数是 ReLU，若使用线性激活函数代替 ReLU，那么该神经网络还能表征异或(XNOR)函数吗？

A 可以

B 不可以

答案

B 线性激活函数不能解决非线性问题，异或(XNOR)关系是非线性的。

常用的非线性激活函数有：sigmoid，tanh，softplus，Relu。

其中 Relu 的函数表达式为：y = max(0, x)

## 知识点：曼哈顿距离&欧式距离

问题：计算 (1, 3), (4, 7) 的曼哈顿距离(L1距离)和欧氏距离(L2距离)

答案：

​L1：|1-4| + |3-7| = 7

L2: sqrt((1-4)^2 + (3-7)^2) = 5

## 知识点：召回率和准确率

问题：简述召回率(Recall)和准确率 (Precision)的区别，并举例说明。

​答案

准确率和召回率是广泛用于信息检索和统计学分类领域的两个度量值，用来评价结果的质量。其中精度是检索出相关文档数与检索出的文档总数的比率，衡量的是检索系统的查准率；召回率是指检索出的相关文档数和文档库中所有的相关文档数的比率，衡量的是检索系统的查全率。

简而言之：

召回率 (Recall)：正样本有多少被找出来了（召回了多少）。 准确率 (Precision)：你认为的正样本，有多少猜对了（猜的准确性如何）。

举个例子：1000辆卡车，5辆有质量问题。选取了其中10辆，其中有3辆有质量问题。那么找到有质量问题的卡车的准确率为 3/10，召回率为 3/5。

**知识点：集成错误率**

问题：为什么随着集成中个体分类器数目T的增大，集成的错误率将指数级下降？

答案：对于二分类问题，假设分类错误率是，即对每个基分类器有：

假设集成通过简单投票法结合T 个基分类器，若有超过半数的基分类器正确，则集成分类就正确:

假设基分类器的错误率相互独立，则集成的错误率为：

上式显示出，随着集成中个体分类器数目T 的增大，集成的错误率将指数级下降，最终趋向于零.

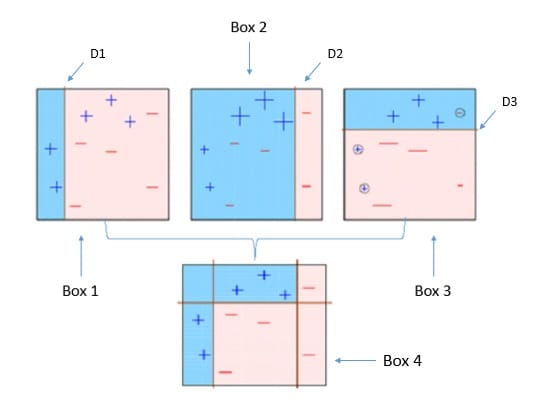
## 知识点：集成学习研究

问题：集成学习的关键假设是什么？研究的核心问题是什么？

回答：集成学习关键假设是基学习器的误差相互独立.在现实任务中，个体学习器是为解决同一个问题训练出来的，它们显然不可能相互独立!事实上，个体学习器的"准确性"和"多样性"本身就存在冲突.一般的，准确性很高之后，要增加多样性就需牺牲准确性.事实上，如何产生并结合"好而不同"的个体学习器，恰是集成学习研究的核心.

## 知识点：Boosting

问题：简单描述Boosting方法的原理。



答案：

boosting算法是一类将弱学习器提升为强学习器的集成学习算法，它通过改变训练样本的权值，学习多个分类器，并将这些分类器进行线性组合，提高泛化性能。

先介绍一下“强学习”和“弱学习”的概念：一个分类，如果存在一个多项式算法能够学习他，并得到很高的正确率，那么这个算法称为强学习器，反之如果正确率只是稍大于随机猜测（50%），则称为弱学习器。在实际情况中，我们往往会发现弱学习器比强学习器更容易获得，所以就有了能否把弱学习器提升（boosting）为强学习器的疑问。于是提升类方法应运而生，它代表了一类从弱学习器出发，反复训练，得到一系列弱学习器，然后组合这些弱学习器，构成一个强学习器的算法。大多数boost方法会改变数据的概率分布（改变数据权值），具体而言就是提高前一轮训练中被错分类的数据的权值，降低正确分类数据的权值，使得被错误分类的数据在下轮的训练中更受关注；然后根据不同分布调用弱学习算法得到一系列弱学习器实现的，再将这些学习器线性组合，具体组合方法是误差率小的学习器会被增大权值，误差率大的学习器会被减小权值，典型代表adaboost算法。

## 知识点：随机森林

问题：随机森林的随机性体现在哪里？

​答案：

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。随机森林的随机性体现在每颗树的训练样本是随机的，树中每个节点的分裂属性集合也是随机选择确定的。有了这2个随机的保证，随机森林就不会产生过拟合的现象了。

 随机森林是用一种随机的方式建立的一个森林，森林是由很多棵决策树组成的，每棵树所分配的训练样本是随机的，树中每个节点的分裂属性集合也是随机选择确定的。

**知识点：监督学习和非监督学习**

问题：什么是监督学习和非监督学习，请说明它们的区别，并各举一个例子。说明分类和回归问题的区别，并各举一个例子。

答案：

（1）有监督学习：对具有标记的训练样本进行学习来建立从样本特征到标记的映射。例如：支持向量机

无监督学习：对没有标记的训练样本进行学习，以发现训练样本集中的结构性知识。聚类就是典型的无监督学习。比如：K-means等。

（2）回归是监督学习的一种，它的标记是连续取值，有大小区别，可以计算标记间的距离。比如linear回归。

分类问题是监督学习的一种，它的标记是若干个离散取值，没有大小区别，不能计算标记间的距离。针对的是离散型结果。比如，朴素贝叶斯，SVM等。

**知识点：过拟合**

问题：什么是过学习，什么情况下可能发生过学习，采取什么措施有助于消除过学习。

答案:

(1) 过学习是指训练误差比较小，而测试误差大得多的情况。

(2) 模型过于复杂，参数过多；数据集相对于模型复杂度太小。

(3) 1.搜集大量的训练样本；2.用一部分样本构造验证集；3.引入正则项惩罚模型复杂度

**知识点：损失函数**

问题：给出线性回归（Regression）和逻辑回归（Logistic Regression）所对应的损失函数。

解：

线性回归的损失函数：

* 线性模型： 通常采用均方误差损失函数：



逻辑回归的损失函数：

逻辑回归是最大化对数几率和：



等价于最小化：



其中：，

**知识点：多层感知器**

问题：简述多层感知机（MLP）中BP算法的流程。

答案：

BP神经网络的流程：

对所有的输入和标签对循环（1-3）：

1.对某个输入从输入层沿网络传播到输出层；

2.从输出层开始从后向前计算每个单元的delta；

3.计算该输入对每个权值梯度的贡献

4.所有输入对梯度的贡献求和得到总梯度，根据学习率来改变原来的权值，完成了权值的一次修改。

**知识点：朴素贝叶斯分类器**

问题：有数据集D1，其中样本的特征是离散取值（可以简单地考虑取二值），数据集D2和D1基本一样，唯一的区别是D2中每个样本的某个特征被重复了100次，请问在这两个数据集上训练的朴素贝叶斯分类器是否一样，请给出具体分析。

答案：

分类器是不一样的。因为朴素贝叶斯方法假设了特征间的独立性，但D2中的100个特征彼此不独立，因此不在适用，如果用了两者的结果不等。在D2上训练，被重复的特征的概率会被乘100次，放大了它的影响。

**知识点：KNN**

问题.考虑下面样本特征为二维欧式空间点的两分类问题的训练集，分别用最近邻法和三近邻法给出测试样本点(1,1)的类别。

正例样本：(0,1)(1,0)(1,2)(2,3)

反例样本：(-1,1) (0,2)(1,-1)(2,2)

答案：计算(1,1) 到各样本的距离：

(-1,1) √((-1-1)^2+(1-1)^2 )=2 -

(0,1) √((0-1)^2+(1-1)^2 )=1 +

(0,2) √((0-1)^2+(2-1)^2 )=√2 -

(1,-1) √((1-1)^2+(-1-1)^2 )=2 -

(1,0) √((1-1)^2+(0-1)^2 )=1 +

(1,2) √((1-1)^2+(2-1)^2 )=1 +

(2,2) √((2-1)^2+(2-1)^2 )=√2 -

(2,3) √((2-1)^2+(3-1)^2 )=√5 +

最近邻法: (0,1) +, (1,0) +, (1,2) + -----> +

三近邻法: (0,1) +, (1,0) +, (1,2) + -----> +

**知识点：最大似然法则**

问题：用两个硬币玩抛硬币的游戏，硬币1得到正面的概率为θ，硬币2得到正面的概率为2θ，你一共抛了五次，得到的结果是这样的（硬币1，正面）（硬币2，反面）（硬币2，反面）（硬币2，反面）（硬币2，正面），用极大似然法求参数θ。

解：

P1(x=正面)=θ

P2(x=正面)=2\*θ

F(θ)=P1(x=正面)\*(1-P2(x=正面))^3\*P2(x=正面)=θ\*(1-2\*θ)^3\*2\*θ=6\*θ^2\*(1-2\*θ)^3

按照极大似然法则，当F(θ)达到极大值时：θ=1/5

**知识点：SVM**

问题：有一个训练集，其样本为二维空间的点，正样本(1,1) 、(−1,−1)，负样本 (1,−1)、(−1, 1)。

（1）正负样本在原空间是否线性可分？

（2）考虑一个特征变换φ(x) = [1, x1, x2, x1x2], 其中x1和x2为某样本x的两个坐标，在特征空间的预测函数为y(x) = wTφ(x)，利用最大间隔方法求预测函数的参数w。（可以通过观察求解）

（3）特征映射函数φ(x)对应的核函数K(x, x’ )是什么样的？

解：（1）原问题是异或问题，正负样本线性不可分；

（2）容易写出特征映射之后的样本点的值，正样本（1，1，1，1），（1，-1，-1，1），负样本（1，1，-1，-1），（1，-1，1，-1），观察易得w取（0，0，0，1）.最大间隔为2。

（3）

知识点：BP算法

**问题：什么是标准BP算法和累积BP算法？它们之间的区别和联系？**

答案：

标准BP算法每次仅针对一个训练样例更新连接权和阔值，也就是说算法的更新规则是基于单个样例的误差推导而得.如果基于批量样例的累积误差最小化的更新规则，就是累积误差逆传播(accumulated error backpropagation)。算法累积BP算法与标准BP算法都很常用.一般来说，标准BP 算法每次更新只针对单个样例，参数更新得非常频繁，而且对不同样例进行更新的效果可能出现"抵消"现象.因此，为了达到同样的累积误差极小点，标准BP 算法往往需进行更多次数的法代.累积BP 算法直接针对累积误差最小化，它在读取整个训练集D 一遍后才对参数进行更新，其参数更新的频率低得多.但在很多任务中，累积误差下降到一定程度之后，进一步下降会非常缓慢，这时标准BP 往往会更快获得较好的解，尤其是在训练集D 非常大时更明显.

**问题：如何解决人工神经网络的过拟合问题？**

答案：

BP 神经网络经常遭遇过拟合，其训练误差持续降低，但测试误差却可能上升。有两种策略常用来缓解BP网络的过拟合.第一种策略是"早停" (early stopping): 将数据分成训练集和验证集，训练集用来计算梯度、更新连接权和阈值。验证集用来估计误差，若训练集误差降低但验证集误差升高，则停止训练，同时返回具有最小验证集误差的连接权和阑值.第二种策略是"正则化" (regularization) ，基本思想是在误差目标函数中增加一个用于描述网络复杂度的部分。

**问题：简述基于梯度的搜索的局部最小与全局最小之间的区别与联系？**

答案：

基于梯度的搜索足使用最为广泛的参数寻优方法.在此类方法中，我们从某些初始解出发?迭代寻找最优参数值.每次迭代中，我们先计算误差函数在当前点的梯度，然后根据梯度确定搜索方向.例如，由于负梯度方向是函数值下降

，因此梯度下降法就是沿着负梯度方向搜索最优解.若误差函数在当前点的梯度为零，则已达到局部极小，更新量将为零，这意味着参数的迭代更新将在此停止.显然，如果误差函数仅有一个局部极小，那么此时找到的局部极小就是全局最小;然而，如果误差函数具有多个局部极小?则不能保证找到的解是全局最小.对后一种情形?我们称参数寻优陷入了局部极小， 这显然不是我们所希望的.

**问题：常用的“跳出”局部极小的方法有哪些？**

答案：

（1）以多组不同参数值初始化多个神经网络，按标准方法训练后，取其中误差最小的解作为最终参数.这相当于从多个不同的初始点开始搜索， 这样就可能陷入不同的局部极小从中进行选择有可能获得更接近全局最小的结果.

（2）使用模拟退火算法。模拟退火在每一步都以二定的概率接受比当前解更差的结果，从而有助于"跳出"局部极小. 在每步主代过程中7，接受"次优解"的概率要随着时间的推移而逐渐降低。从而保证算法稳定.

（3）使用随机梯度下降.与标准梯度下降法精确计算梯度不同， 随机梯度下降法在计算梯度时加入了随机因素.于是?即便陷入局部极小点? 它计算出的梯度仍可能不为零3 这样就有机会跳出局部极小继续搜索.

（4）遗传算法。