* 在卷积神经网络计算中，已知输入特征层大小为32x32x64, 使用标准卷积计算，带偏置项，卷积核大小为3\*3，输出特征层数目为64，请问卷积层的参数个数为？

每个Filter有3\*3\*64个参数（64为# Channels）。总共有64个Filter（输出特征层数为64）。共3\*3\*64\*64个参数。

详细公式：<https://blog.csdn.net/gaishi_hero/article/details/81512404>

* 什么数据集不适合用深度学习？

1. 数据集太小，数据样本不足时，深度学习相对其它机器学习算法，没有明显优势。
2. 数据集没有局部相关特性，目前深度学习表现比较好的领域主要是图像／语音／自然语言处理等领域，这些领域的一个共性是局部相关性。图像中像素组成物体，语音信号中音位组合成单词，文本数据中单词组合成句子，这些特征元素的组合一旦被打乱，表示的含义同时也被改变。对于没有这样的局部相关性的数据集，不适于使用深度学习算法进行处理。举个例子：预测一个人的健康状况，相关的参数会有年龄、职业、收入、家庭状况等各种元素，将这些元素打乱，并不会影响相关的结果。（如果特征顺序打乱并不会影响结果，则不适合用深度学习）

* 在二分类问题中，当测试集的正例和负例数量不均衡时，以下评价方案哪个是相对不合理的（ A ）（假设precision=TP/(TP+FP),recall=TP/(TP+FN)。）

A: Accuracy:(TP+TN)/all

B: F-value:2\*recall\*precision/(recall+precision)

C: G-mean:sqrt(precision\*recall)

D: AUC:ROC曲线下面积

对于分类器，主要的评价指标有precision，recall，F-score，以及ROC曲线等。

在二分类问题中，我们主要关注的是测试集的正样本能否正确分类。当样本不均衡时，比如样本中负样本数量远远多于正样本，此时如果负样本能够全部正确分类，而正样本只能部分正确分类，那么(TP+TN)可以得到很高的值，也就是Accuracy是个较大的值，但是正样本并没有取得良好的分类效果。因此A选项是不合理的。在样本不均衡时，可以采用BCD选项方法来评价。

* 9、在其它条件不变的前提下，以下哪种做法容易引起机器学习中的过拟合问题（ D ）

A 增加训练集数量

B 减少神经网络隐藏层节点数

C 删除稀疏的特征

D SVM算法中使用高斯核/RBF核代替

机器学习中发生过拟合的主要原因有：

（1）使用过于复杂的模型；

（2）数据噪声较大；

（3）训练数据少。

由此对应的降低过拟合的方法有：

（1）简化模型假设，或者使用惩罚项限制模型复杂度；

（2）进行数据清洗，减少噪声；

（3）收集更多训练数据。

本题中，A对应于增加训练数据，B为简化模型假设，C为数据清洗（稀疏特征即很大部分特征值都为0）。D选项中，高斯核的使用增加了模型复杂度，容易引起过拟合。选择合适的核函数以及软边缘参数C就是训练SVM的重要因素。一般来讲，核函数越复杂，模型越偏向于过拟合；C越大模型越偏向于过拟合，反之则拟合不足。

* 假设你需要调整超参数来最小化代价函数（cost function），会使用下列哪项技术？ D

A.穷举搜索

B.随机搜索

C.Bayesian优化

D.都可以

* 关于朴素贝叶斯分类算法，描述正确的是：A

A.它假设属性之间相互独立

B.根据先验概率计算后验概率

C.对于给定的待分类项X={a1,a2,...,an}，求解在此项出现的条件下各个类别 yi 出现的概率，哪个P(yi|X)最大，就把此待分类项归属于哪个类别。

D.有最小错误率判断规则和最小风险判断规则

B应该加一句“根据贝叶斯公式”

C的解释：简单点回复吧，这是朴素贝叶斯，p(yi|X)可能为0，甚至可能X的一组特征在训练集中没有同时出现，此时P(yi|X)如何计算。所以想要求X出现的概率还是要根据后验来。

* 模式识别中，不属于马式距离较之于欧式距离的优点的是（ ）

欧氏距离是在N维空间中两个点的真实距离；马氏距离表示数据的协方差距离。

而欧式距离的特征是：平移不变性、旋转不变性。

马式距离的特征则是：平移不变性、旋转不变性、尺度不变性、不受量纲影响、考虑了模式分布。

<https://www.zhihu.com/question/20852004> 关于协方差的解释（两个变量的同向程度）

* 统计模式分类问题中，当先验概率未知时，可以使用(A)

1. 最小最大损失准则
2. 最小误判概率准则
3. 最小损失准则
4. N-P判决

链接：https://www.nowcoder.com/questionTerminal/34017bb5168b448abf3bb92231b70ea3

来源：牛客网

A. 考虑p(wi)变化的条件下，是风险最小

B. 最小误判概率准则， 就是判断p(w1|x)和p(w2|x)哪个大，x为特征向量，w1和w2为两分类，根据贝叶斯公式，需要用到先验知识

C. 最小损失准则，在B的基础之上，还要求出p(w1|x)和p(w2|x)的期望损失，因为B需要先验概率，所以C也需要先验概率

D. N-P判决，即限定一类错误率条件下使另一类错误率为最小的两类别决策，即在一类错误率固定的条件下，求另一类错误率的极小值的问题，直接计算p(x|w1)和p(x|w2)的比值，不需要用到贝叶斯公式\_

简单概括：所以可以理解为：先验概率未知，其实就是说不能用生成模型，只能用判别模型。

* 4、如果以特征向量的相关系数作为模式相似性测度，则影响聚类算法结果的主要因素有（B ）

A已知类别样本质量

B分类准则

C量纲

A选项，类别已知暂且不说，样本质量是不会影响聚类结果的。因为聚类的任务只是把数据按照相似性原则进行划分，不存在分类问题中由于训练集样本存在噪声数据，从而影响分类结果的情况。此外，在类别已知的情况下，直接按照样本的类别标签进行聚类就可以了，用不到复杂的聚类算法，所以也就不存在影响聚类算法结果这么一说了；

B选项，分类准则是指选取什么特征将该特征相似的数据聚为一类，这个会直接影响到样本聚类的结果；

C选项，两个特征向量的相关系数与其量纲无关，故不选。

* 6、以下( )不属于线性分类器最佳准则？

A感知准则函数

B贝叶斯分类

C支持向量机

DFisher准则

何为线性分类器？ <https://www.zhihu.com/question/30633734>

线性分类器有三大类：感知器准则函数、SVM、Fisher准则，而贝叶斯分类器不是线性分类器。

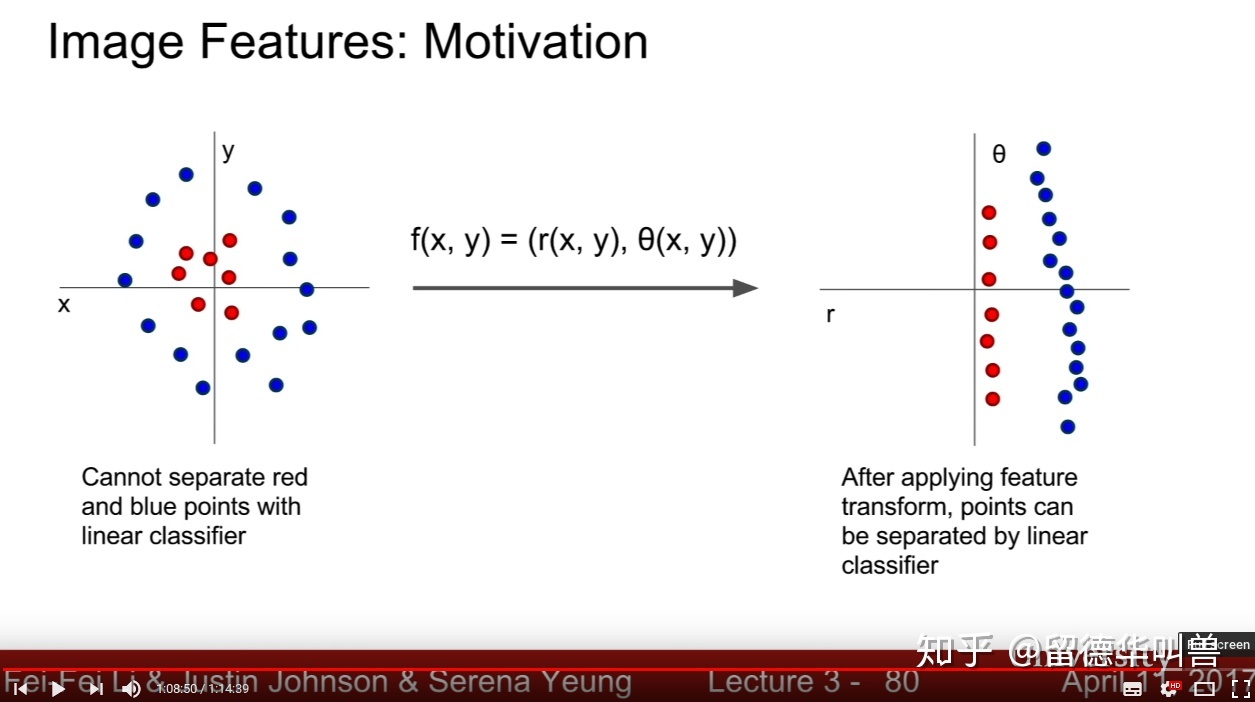
感知准则函数 ：准则函数以使错分类样本到分界面距离之和最小为原则。其优点是通过错分类样本提供的信息对分类器函数进行修正，这种准则是人工神经元网络多层感知器的基础。

支持向量机 ：基本思想是在两类线性可分条件下，所设计的分类器界面使两类之间的间隔为最大，它的基本出发点是使期望泛化风险尽可能小。（使用核函数可解决非线性问题）

Fisher 准则 ：更广泛的称呼是线性判别分析（LDA），将所有样本投影到一条远点出发的直线，使得同类样本距离尽可能小，不同类样本距离尽可能大，具体为最大化“广义瑞利商”。

根据两类样本一般类内密集，类间分离的特点，寻找线性分类器最佳的法线向量方向，使两类样本在该方向上的投影满足类内尽可能密集，类间尽可能分开。这种度量通过类内离散矩阵 Sw 和类间离散矩阵 Sb 实现。

使用线性分类器分类一组线性不可分数据：高维映射



* 

注意最大间隔就是垂直平分线

* 
* 1、以下几种模型方法属于判别式模型(Discriminative Model)的有( 2,3)

1)混合高斯模型

2)条件随机场模型

3)区分度训练

4)隐马尔科夫模型

**常见的判别式模型有：**

Logistic regression（logistical 回归）

Linear discriminant analysis（线性判别分析）

Supportvector machines（支持向量机）

Boosting（集成学习）

Conditional random fields（条件随机场）

Linear regression（线性回归）

Neural networks（神经网络）

**常见的生成式模型有:**

Gaussian mixture model and othertypes of mixture model（高斯混合及其他类型混合模型）

Hidden Markov model（隐马尔可夫）

NaiveBayes（朴素贝叶斯）

AODE（平均单依赖估计）

Latent Dirichlet allocation（LDA主题模型）

Restricted Boltzmann Machine（限制波兹曼机）

生成式模型是根据概率乘出结果，而判别式模型是给出输入，计算出结果。

* 在Logistic Regression 中,如果同时加入L1和L2范数,不会产生什么效果(D)

A以做特征选择,并在一定程度上防止过拟合

B能解决维度灾难问题

C能加快计算速度

D可以获得更准确的结果

L1范数是向量中各个元素的绝对值之和，又叫稀疏规则算子。L1正则化通过向代价函数中添加权重向量的L1范数（即正则化项），使得优化后的模型中无用特征对应的权值变为0，相当于减少了特征维数，实现了特征的自动选择，所以LR中加入L1范数可以进行特征选择、解决维度灾难问题、加快计算速度；

L2范数是向量中各个元素平方和的1/2次方。L2正则化通过向代价函数中添加权重向量的L2范数，使得优化后的模型中所有的权值w尽可能趋于0但不为0，通过L2范数，可以实现对模型空间的限制，从而在一定程度上避免了过拟合；

过拟合的时候，拟合函数需要顾忌每一个点，最终形成的拟合函数波动很大， 在某些小区间里，函数值的变化很大，也就是w非常大。所以LR中加入L2范数可以在一定程度上防止过拟合；

在LR中同时加入L1和L2范数不会产生结果更准确的效果。

* 6、机器学习中L1正则化和L2正则化的区别是？ A

A使用L1可以得到稀疏的权值

B使用L1可以得到平滑的权值

C使用L2可以得到稀疏的权值

L1正则化偏向于稀疏，它会自动进行特征选择，去掉一些没用的特征，也就是将这些特征对应的权重置为0.

L2主要功能是为了防止过拟合，当要求参数越小时，说明模型越简单，而模型越简单则，越趋向于平滑，从而防止过拟合。

L1正则化/Lasso

L1正则化将系数w的l1范数作为惩罚项加到损失函数上，由于正则项非零，这就迫使那些弱的特征所对应的系数变成0。因此L1正则化往往会使学到的模型很稀疏（系数w经常为0），这个特性使得L1正则化成为一种很好的特征选择方法。

L2正则化/Ridge regression

L2正则化将系数向量的L2范数添加到了损失函数中。由于L2惩罚项中系数是二次方的，这使得L2和L1有着诸多差异，最明显的一点就是，L2正则化会让系数的取值变得平均。

对于关联特征，这意味着他们能够获得更相近的对应系数。还是以Y=X1+X2为例，假设X1和X2具有很强的关联，如果用L1正则化，不论学到的模型是Y=X1+X2还是Y=2X1，惩罚都是一样的，都是2alpha。但是对于L2来说，第一个模型的惩罚项是2alpha，但第二个模型的是4\*alpha。可以看出，系数之和为常数时，各系数相等时惩罚是最小的，所以才有了L2会让各个系数趋于相同的特点。

可以看出，L2正则化对于特征选择来说一种稳定的模型，不像L1正则化那样，系数会因为细微的数据变化而波动。所以L2正则化和L1正则化提供的价值是不同的，L2正则化对于特征理解来说更加有用：表示能力强的特征对应的系数是非零。

因此，一句话总结就是：L1会趋向于产生少量的特征，而其他的特征都是0，而L2会选择更多的特征，这些特征都会接近于0。Lasso在特征选择时候非常有用，而Ridge就只是一种规则化而已。

* 3、在 k-均值算法中，以下哪个选项可用于获得全局最小？（D）

A尝试为不同的质心（centroid）初始化运行算法

B调整迭代的次数

C找到集群的最佳数量

D以上所有

传统Ｋ均值算法随机选取初始聚类中心，往往会造成聚类结果陷入局部最优解，改进初始类中心的选取方法可以提升Ｋ均值算法的聚类效果，获得全局最优解。

A选项，尝试为不同的质心初始化实际就是在寻找最佳的初始类中心以便达到全局最优；

B选项，迭代的次数太少无法获得最优解，同样也无法获得全局最优解，所以需要通过调整迭代次数来获得全局最优解；

C选项，集群的最佳数量也就是K值是人为定义的，事先不知道多大的K值能够得到全局最优，所以需要调试K值，以达到全局最优。

综上所述，D选项为正确答案。

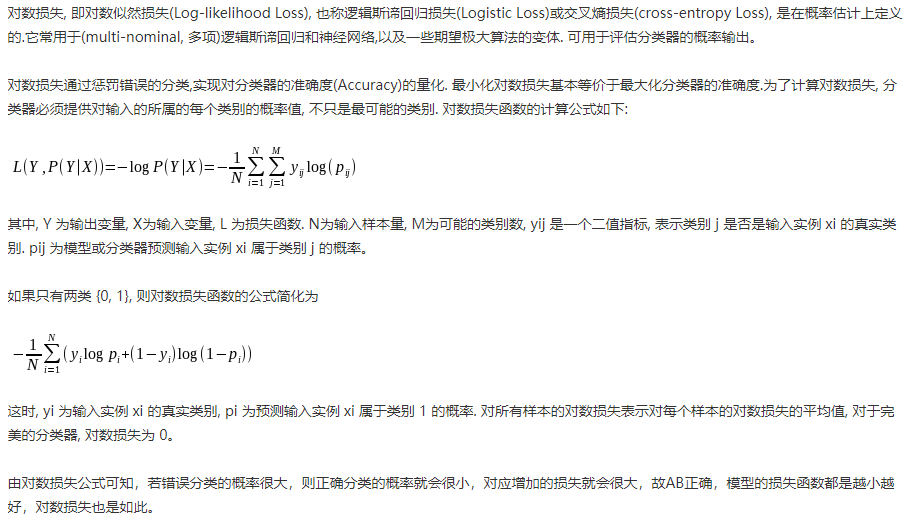
* 4、假设你使用 log-loss 函数作为评估标准。下面这些选项，哪些是对作为评估标准的 log-loss 的正确解释。 D

A如果一个分类器对不正确的分类很自信，log-loss 会严重的批评它

B对一个特别的观察而言，分类器为正确的类别分配非常小的概率，然后对 log-loss 的相应分布会非常大

Clog-loss 越低，模型越好

D以上都是



* 5、下面哪个选项中哪一项属于确定性算法？ A

A：PCA

B：K-Means

C：以上都不是

确定性算法表明在不同运行中，算法输出并不会改变。如果我们再一次运行算法，PCA 会得出相同的结果，而 k-means 不会。

相关课程

* 6、两个变量的 Pearson 相关性系数为零，但这两个变量的值同样可以相关。这句描述是正确还是错误？A

A正确

B错误

Pearson相关系数只能衡量线性相关性，但无法衡量非线性关系。如y=x^2，x和y有很强的非线性关系。









* 下列哪些不适合用来对高维数据进行降维 C

A LASSO

B 主成分分析法

C 聚类分析

D 小波分析法

E 线性判别法

F 拉普拉斯特征映射

LASSO通过参数缩减达到降维的目的；

主成分分析法（PCA）通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维；

线性鉴别法（LDA）通过降维找到一个类内距离最小、类间距离最大的空间实现分类；

小波分析有一些变换的操作降低其他干扰，可以看做是降维；

拉普拉斯特征映射将处于流形上的数据，在尽量保留原数据间相似度的情况下，映射到低维下表示，实现降维；

聚类分析不能用来对高维数据进行降维。

* 5、下列不是SVM核函数的是 B

A 多项式核函数

B logistic核函数

C径向基核函数

经常使用的核函数

核函数的定义并不困难，根据泛函的有关理论，只要一种函数 K ( x i , x j ) 满足Mercer条件，它就对应某一变换空间的内积．对于判断哪些函数是核函数到目前为止也取得了重要的突破，得到Mercer定理和以下常用的核函数类型：

(1)线性核函数

K ( x , x i ) = x ⋅ x i

(2)多项式核

K ( x , x i ) = ( ( x ⋅ x i ) + 1 ) d

(3)径向基核（RBF）

K ( x , x i ) = exp ( − ∥ x − x i ∥ 2 σ 2 )

Gauss径向基函数则是局部性强的核函数，其外推能力随着参数 σ 的增大而减弱。多项式形式的核函数具有良好的全局性质。局部性较差。

(4)傅里叶核

K ( x , x i ) = 1 − q 2 2 ( 1 − 2 q cos ( x − x i ) + q 2 )

(5)样条核

K ( x , x i ) = B 2 n + 1 ( x − x i )

(6)Sigmoid核函数

K ( x , x i ) = tanh ( κ ( x , x i ) − δ )

采用Sigmoid函数作为核函数时，支持向量机实现的就是一种多层感知器神经网络，应用SVM方法，隐含层节点数目(它确定神经网络的结构)、隐含层节点对输入节点的权值都是在设计(训练)的过程中自动确定的。而且支持向量机的理论基础决定了它最终求得的是全局最优值而不是局部最小值，也保证了它对于未知样本的良好泛化能力而不会出现过学习现象。

* 一般，k-NN最近邻方法在（B）的情况下效果较好

A 样本较多但典型性不好

B 样本较少但典型性好

C 样本呈团状分布

D 样本呈链状分布

K近邻算法主要依靠的是周围的点，因此如果样本过多，那肯定是区分不出来的。因此应当选择B

样本呈团状颇有迷惑性，这里应该指的是整个样本都是呈团状分布，这样kNN就发挥不出其求近邻的优势了，整体样本应该具有典型性好，样本较少，比较适宜。

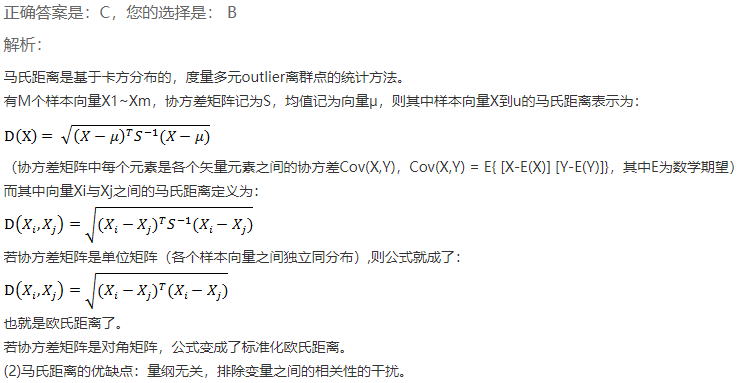
* 在一个n维的空间中， 最好的检测outlier(离群点)的方法是（B）

A 作正态分布概率图

B 作盒形图

C 马氏距离

D 作散点图



* “过拟合”只在监督学习中出现，在非监督学习中，没有“过拟合”，这是（B）

A对的

B错的

聚类中，如果一个样本为一个类就是过拟合。

<http://sofasofa.io/forum_main_post.php?postid=1000490>

<https://www.zhihu.com/question/327484489>

* 对于k折交叉验证, 以下对k的说法正确的是（D）

A：k越大, 不一定越好, 选择大的k会加大评估时间

B：选择更大的k, 就会有更小的bias (因为训练集更加接近总数据集)

C：在选择k时, 要最小化数据集之间的方差

D：以上所有

k越大, bias越小, 训练时间越长. 在训练时, 也要考虑数据集间方差差别不大的原则. 比如, 对于二类分类问题, 使用2-折交叉验证, 如果测试集里的数据都是A类的, 而训练集中数据都是B类的, 显然, 测试效果会很差.

如果不明白bias和variance的概念, 务必参考下面链接:

Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning

http://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/

Understanding the Bias-Variance Tradeoff

<http://scott.fortmann-roe.com/docs/BiasVariance.html>

* 回归模型中存在多重共线性, 你如何解决这个问题？

1 去除这两个共线性变量

2 我们可以先去除一个共线性变量

3 计算VIF(方差膨胀因子), 采取相应措施

4 为了避免损失信息, 我们可以使用一些正则化方法, 比如, 岭回归和lasso回归

解决多重公线性, 可以使用相关矩阵去去除相关性高于75%的变量 (有主观成分). 也可以VIF, 如果VIF值<=4说明相关性不是很高, VIF值>=10说明相关性较高.

我们也可以用 岭回归和lasso回归的带有惩罚正则项的方法. 我们也可以在一些变量上加随机噪声, 使得变量之间变得不同, 但是这个方法要小心使用, 可能会影响预测效果

* 模型的高bias是什么意思, 我们如何降低它 ？B

A 在特征空间中减少特征

B 在特征空间中增加特征

C 增加数据点

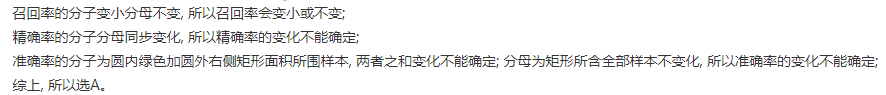
D： B和C

E 以上所有

bias表示模型预测值的均值与样本实际值的差距，它反映了模型对样本数据的拟合能力。bias越低，说明模型越复杂，参数越多，对样本数据的拟合效果越好，但是容易过拟合；bias越高，说明模型越简单，参数太少，对样本数据的拟合效果不好，这就是欠拟合。

降低bias的方法是增加数据的特征维数，从而实现模型参数的增加，提高模型复杂度，增强模型对样本数据的拟合能力，拟合能力越高bias越低。

增加样本数量并没有增加模型训练时的参数，所以不会提高模型复杂度，也就无法降低bias，C错误。在特征空间中增加特征就是增加样本数据的输入特征维数，所以A错误，B正确。



* 假如我们使用非线性可分的SVM目标函数作为最优化对象, 我们怎么保证模型线性可分？ C

A 设C=1

B 设C=0

C 设C=无穷大

D 以上都不对





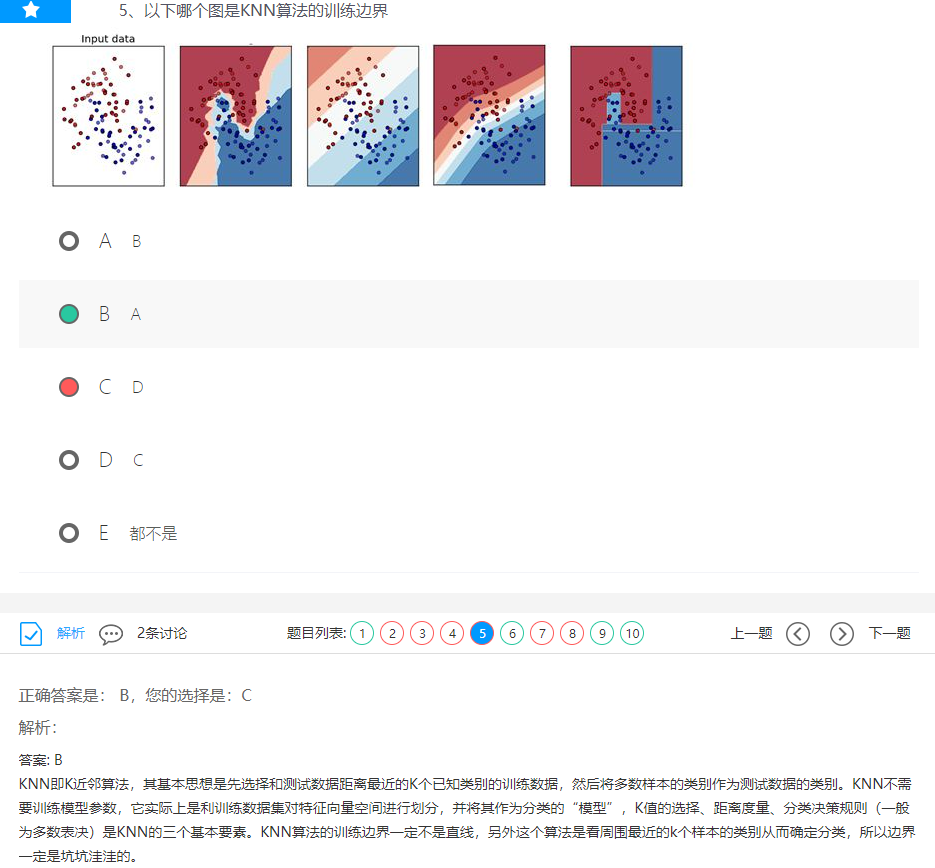








叶子处样本越少，就是分类越细，越容易过拟合。所以增加最小样本数防止过拟合。 减少单个样本会增加偏差，所以会降低variance





重复两次是指只运行两次，总共有有5 fold，只取两个fold来运行

，

* 常见损失函数：<https://zhuanlan.zhihu.com/p/58883095>
* KMeans算法步骤，KMeans ++算法步骤



















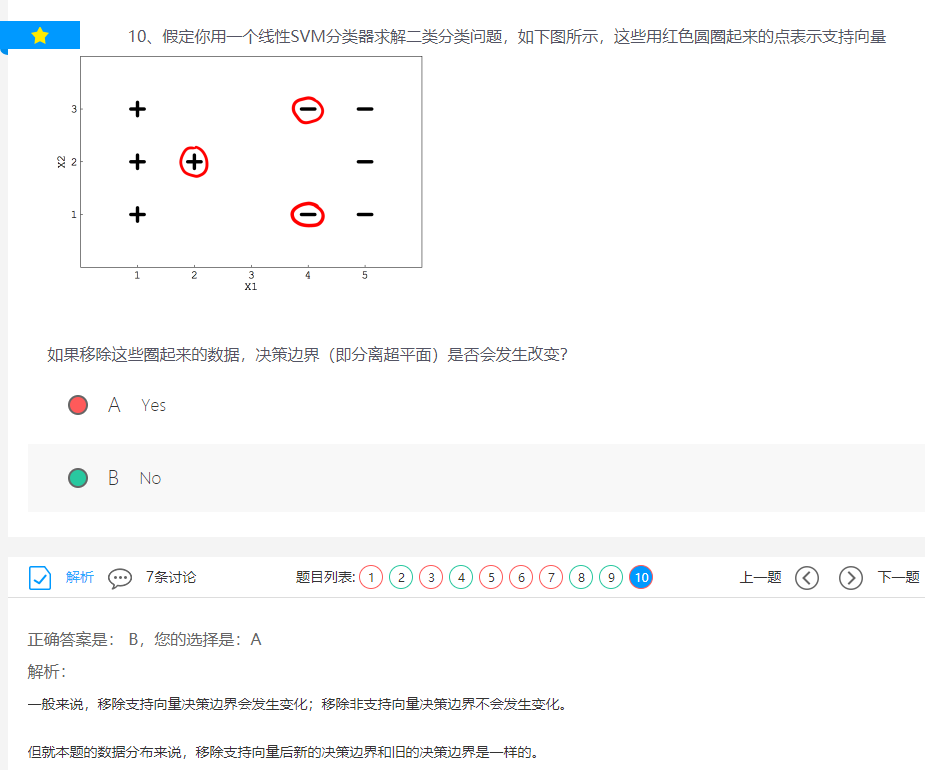
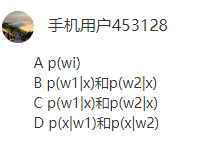
卡方检验进行特征选择

<https://segmentfault.com/a/1190000003719712>

<https://cloud.tencent.com/developer/article/1178904>







移出后，新的决策边界还是X=3的超平面





硬间隔容易造成过拟合



<https://zhuanlan.zhihu.com/p/86844493>



核函数SVM可以扩展到聚类问题

A是过拟合解决手段



归一化并不一定有用

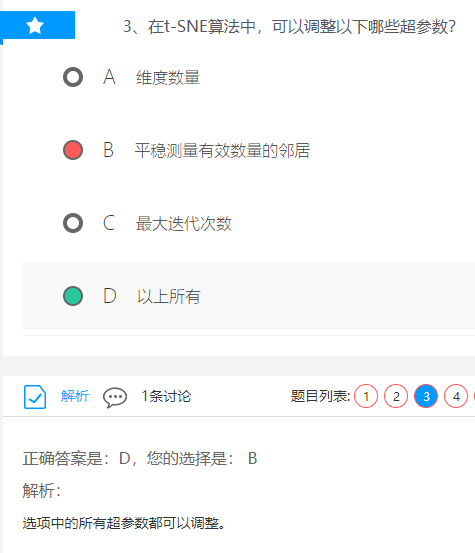


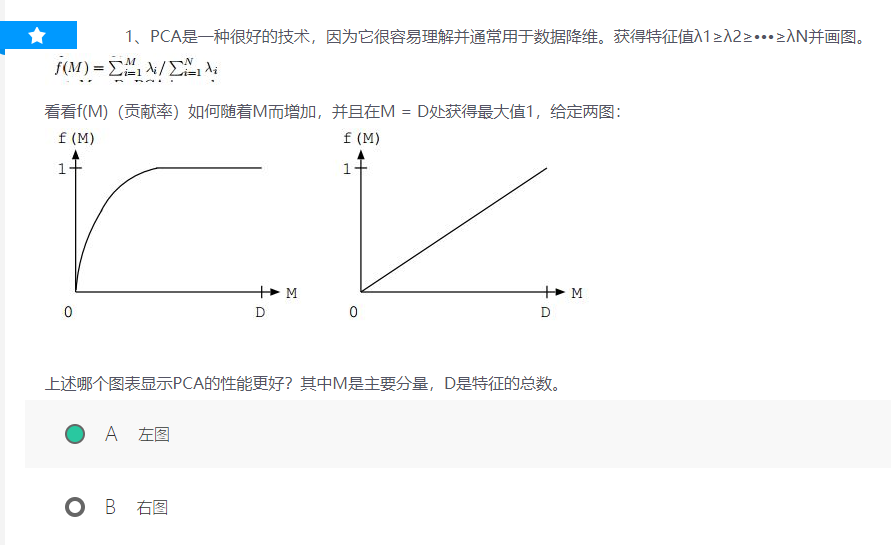
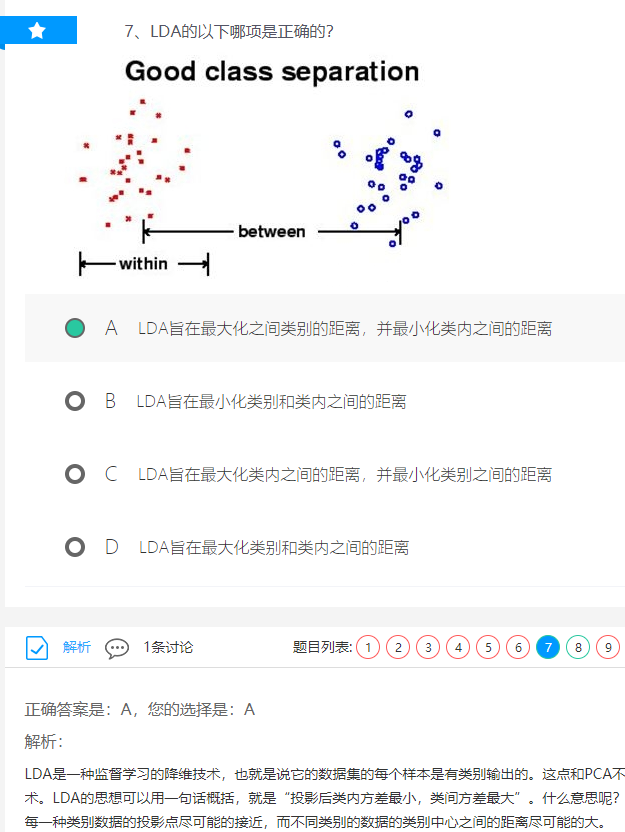
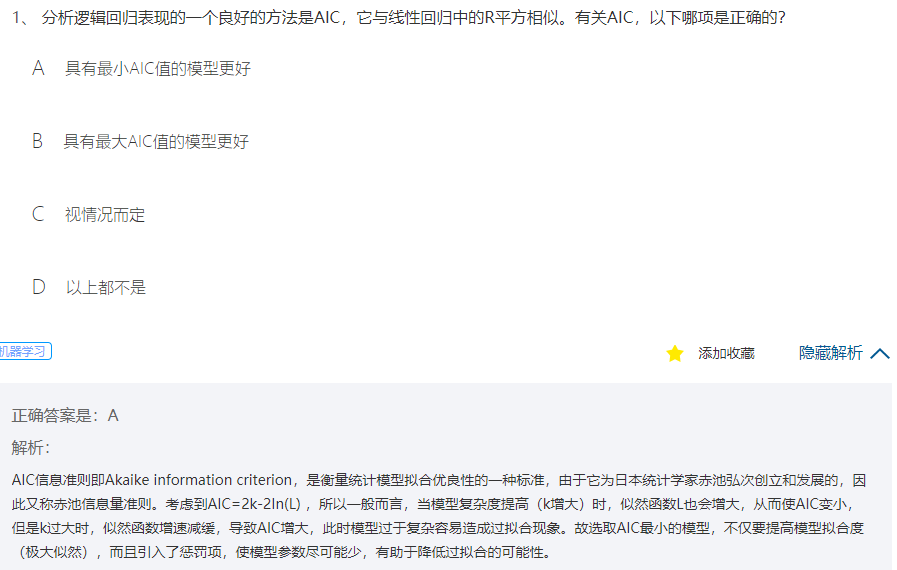
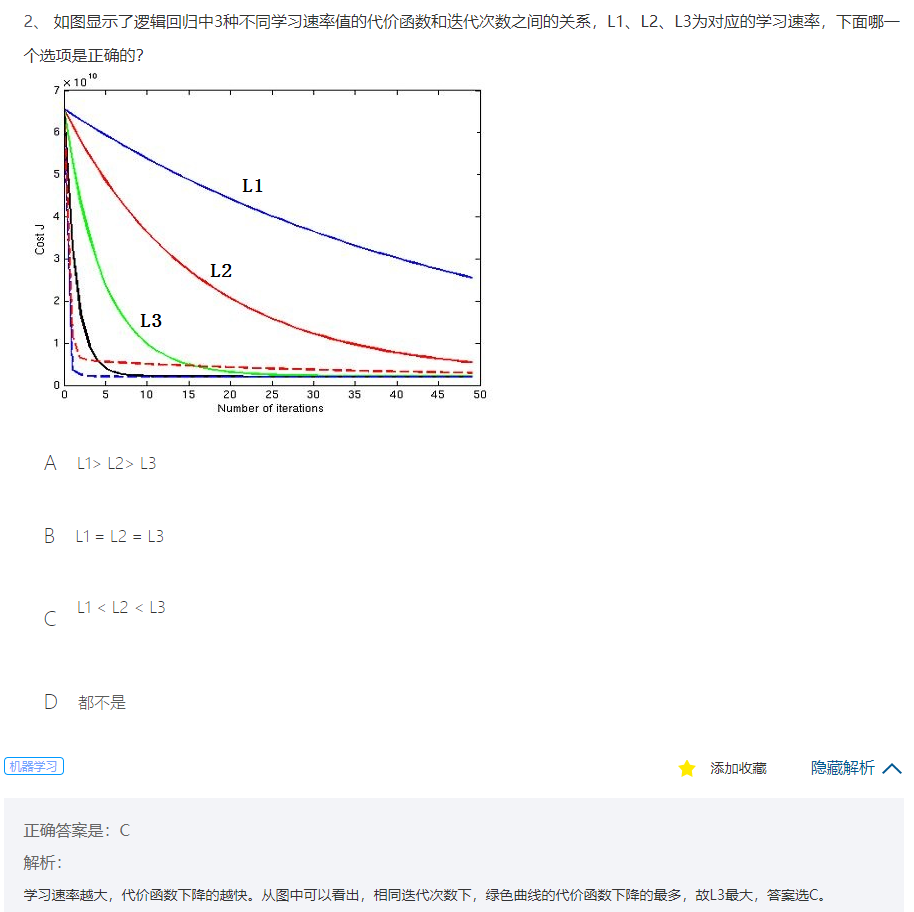


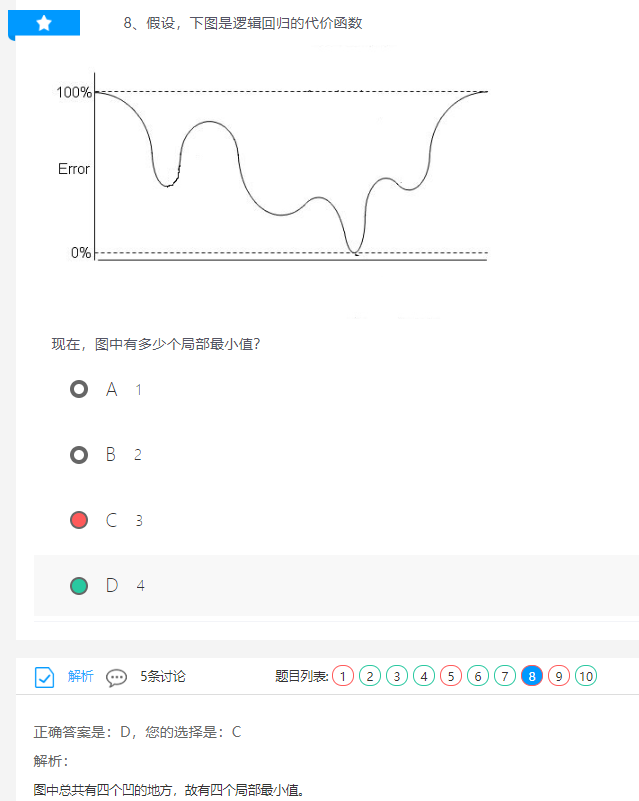
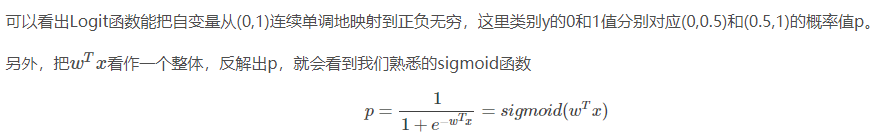
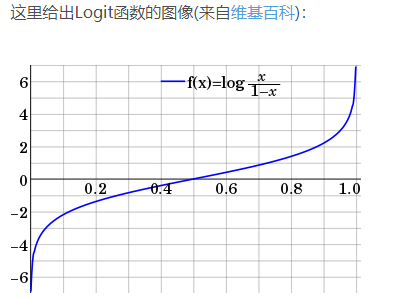
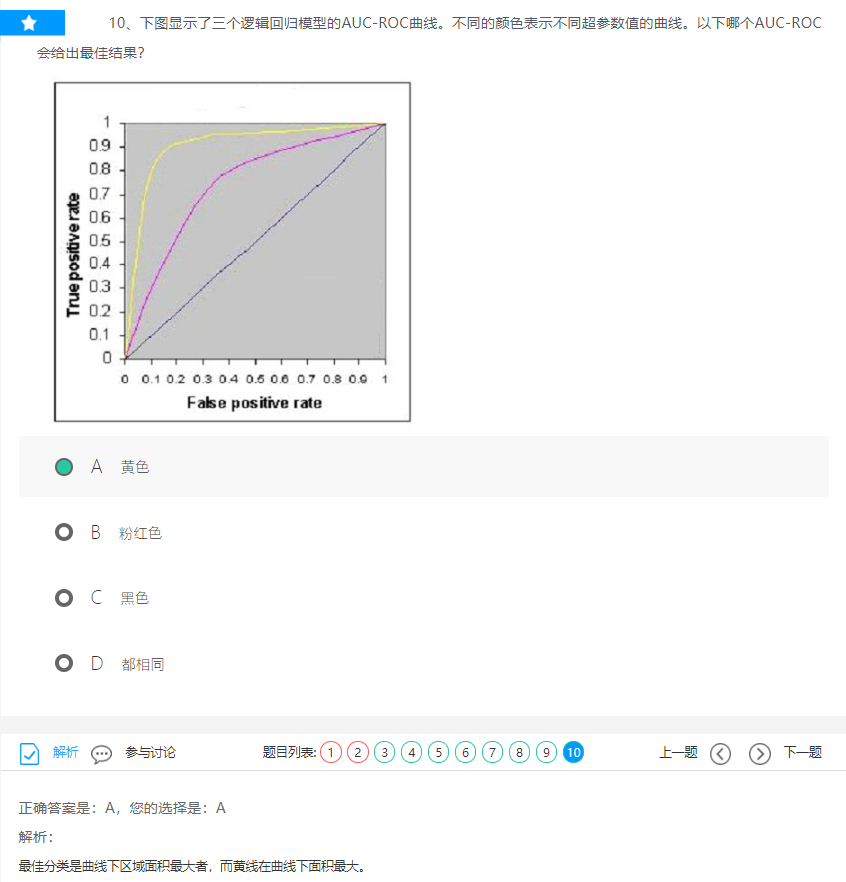
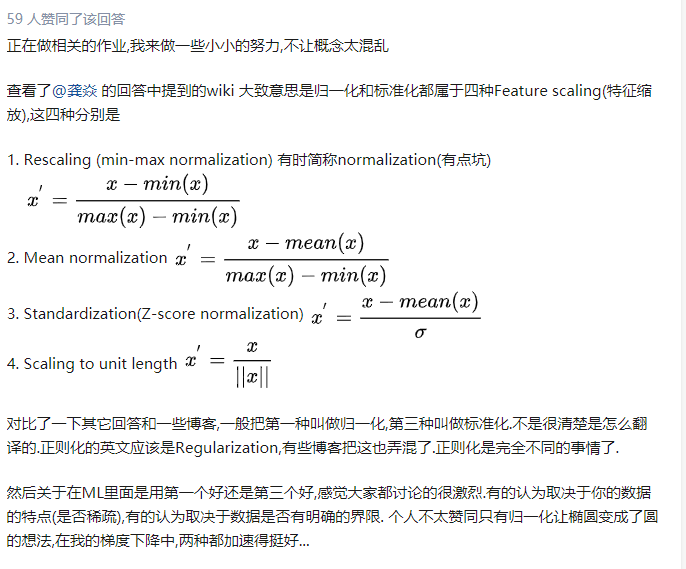


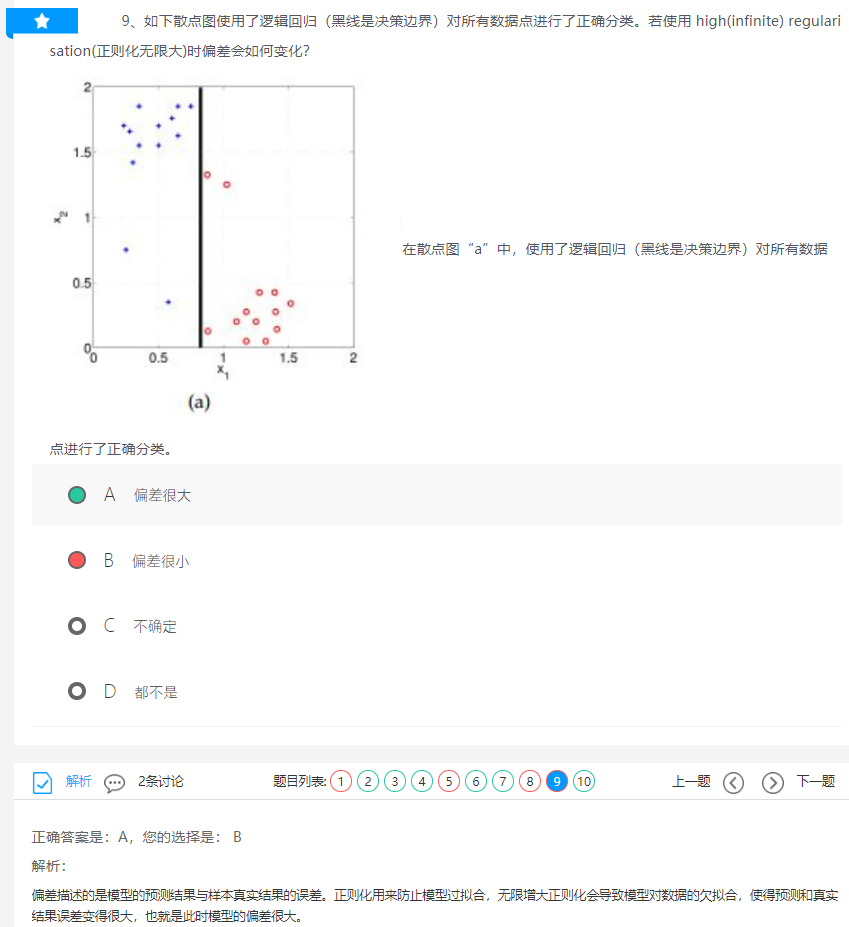
从字面意思应该是事先模型假定了参数和目标函数，那么就是学习参数映射，如果没有目标函数，就是学习非参数映射。





全局最小一定是局部最小，我觉得答案没毛病













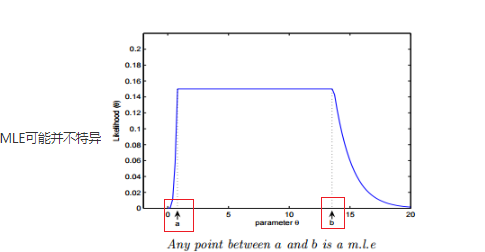


解析对各个指标的解释有误





如果极大似然函数 L(θ) 在极大值处不连续，一阶导数不存在，则 MLE 不存在









http://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/





