**第1章 特征工程**

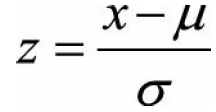
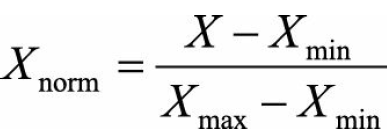
结构化数据：清晰定义的数据，类似关系型数据库的一张表

非结构化数据：文本，图像，音频，视频等，没有清晰的类别定义。

**归一化**：消除数据特征之间的量纲影响。可将所有特征都统一到一个大致相同的数据区间。

（1） min-max scaling: 对数据进行线性变化，结果映射到[0,1]范围，对原始数据等比缩放

（2）Z-score Normalization: 将数据映射到正态分布上。



注意：归一化不是万能的。一般通过梯度下降求解的模型都需要归一化，如线性回归，逻辑回归，SVM，CNN。但是**决策树模型**并不适用，因为决策树在进行节点分裂时主要依据数据集关于特征x的信息增益比，而信息增益比跟是否经过归一化时无关的，因为归一化并不会改变样本在特征x上的信息增益，归一化改变的只是样本的分布区间。

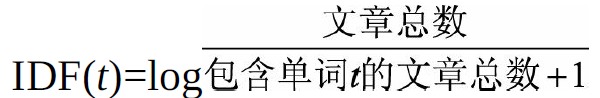
**Categorical Feature**：除了决策树等少数模型能够直接处理字符串形式的输入，对于逻辑回归，SVM，这种特别必须经过**处理转换成数值型特征**才能正确工作。

不同encoding：（1）ordinal encoding：处理具有大小顺序的数据；（2）one-hot encoding: 处理不具有大小关系的特征；用sparse vector表示可节省空间。配合特征选择来降低维度（3）binary encoding。

高维度特征问题：（1）K近邻算法中，高维空间下两点之间的距离很难得到有效的衡量；（2）逻辑回归模型中，参数量随着维度增高而增加，容易过拟合；（3）只有部分维度对分类，预测有帮助，因此可以考虑配合特征选择降低维度。

组合特征：为了提供复杂关系的拟合能力，把一阶离散特征两两组合成高阶组合特征。可基于决策树的特征组合寻找，每一条从根节点到叶节点的路径都可以看成一种特征组合方式。即，路径🡪特征组合。

**文本模型**：(1) bag of words; (2) TF-IDF (Term Frequency-Inverse Document Frequency); (3) Topic Model; (4) Word Embedding;

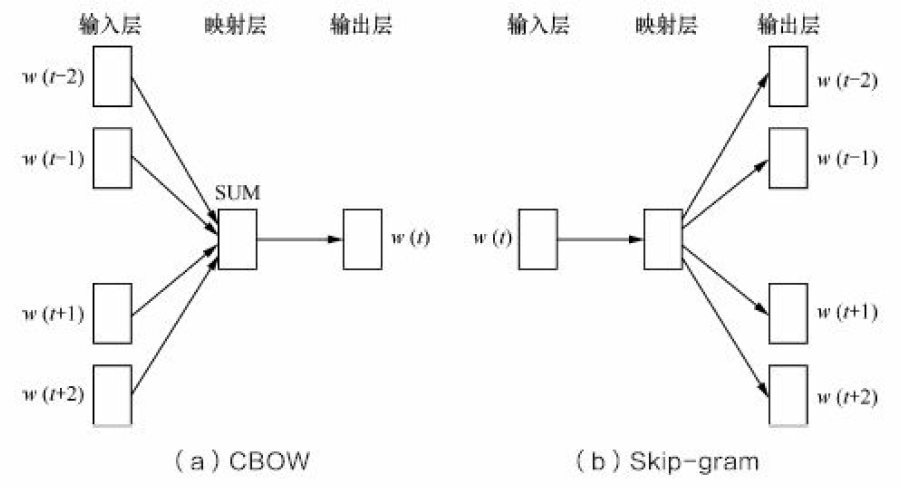
Bag of words: 最基础的文本模型。将每篇文章看成一袋子词，忽略单词出现顺序。即，每篇文章表示成一个常相良，向量中的每一维表示一个单词，该维度对应的权重表示其重要性。常用TF-IDF计算权重。，。TF(t,d)为单词t在文档d中出现的频率，IDF(t)是逆文档频率，用来衡量单词t对表达语义的重要性。通俗理解：如果一个单词在非常多的文章里面都出现，那么它可能是一个比较通用的词汇，对区分某篇文章特殊语义的贡献较小，因此对权重进行惩罚。

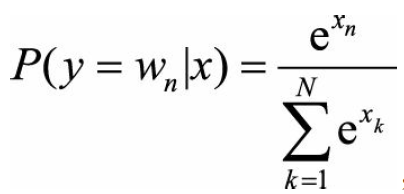
注：单独划分单词并不是好的做好，因为单独拆开单词有时候会产生完全不同的语义。另一种方法是采用N-gram模型，即n(n<=N)个词组成一个单独特征。在实际英中中，一般会对单词进行词干抽取（word stemming），将不同词性的单词统一成同一词干的形式。

Word embedding：将词向量化的模型统称，即将每个词都映射成低维空间(k=50~300)上的一个dense vector，每个词都是一个k维向量。假设一篇文本有N个词，则尺度为N\*K。直接用这个矩阵训练的结果是不好的，需要在这个基础上加工出更高层的特征，可以采用DL来实现高级特征提取。

Word2Vec：谷歌在13年提出的最常用word embedding模型之一。有两种网络结果，分别是CBOW(continuous bag of words)和skip-gram。

Word2Vec工作原理：（1）CBOW:根据上下文出现的词语来预测当前词的生成概率，上下文🡪当前词；（2）skip-gram：根据当前词来预测上下文中各词的生成概率，当前词🡪上下文。上下文是一个滑窗，可设窗口大小，下图滑窗大小为2。



输入层每个单词由one-hot encoding表示，所有词均表示为一个N维向量，N为词汇表中单词的总数。映射层可通过输入层乘以N\*K权重矩阵得到K个隐层单元。映射层到输出层是另一个K\*N矩阵，得到N维的输出层。然后对这个输出层应用softmax得到每个单词的概率。

Word2Vec与LDA（隐狄利克模型）的区别和联系：（1）LDA利用文档中单词的共现关系对单词按主题聚类，理解为对**文档-单词**矩阵进行分解，得到**文档-主题**和**主题-单词**两个概率分布。（2）Word2Vec主要利用**上下文-单词**矩阵进行学习，更多融入了上下文共现的特征。（3）上述两个只是一般的不同点，但最主要的不同点是模型本身：主题模型是一种基于概率模型的生成式模型，其似然函数可以写成若干条件概率连乘的方式，其中包括需要推测的隐含变量，即主题；而word embedding一般表达为神经网络的形式，似然函数定义在网络的输出智商，需要通过学习网络的权重以得到单词的dense vector表示。

***在图像分类任务中，训练数据不足会带来什么问题？如果解决数据量不足的问题？***

一个模型所能提供的信息有两种：（1）训练数据本身包含的信息；（2）模型过程中人类提供的先验信息。当数据少，则信息少，因此需要从先验信息方面进行补充。如何操作先验？可以添加约束或者假设；也可以直接操作数据集，如数据增广。数据增广除了操作原始数据变化，也可以操作特征变化。

数据不足会带来过拟合的问题。常用解决过拟合：正则化，ensembling, dropout，数据增广。

**第2章 模型评估**

模型评估：offline evaluation和online evaluation。

准确率accuracy：分类正确的样本占总样本个数的比例

**准确率是分类问题**中最简单直观的metric，但也有明显缺陷，即负样本占99%时，分类器把所有样本预测为负样本也可以得到99%的精确率。所以，当样本比例非常不均衡时，比例大的类别往往成为影响准确率的最主要因素。可以使用平均准确率mean accuracy，每个类别下的样本准确率的算数平均值。

精确率（查准率）precision：分类正确的正样本数占分类器判定为正样本数的比例。即tp/(tp+fp)。

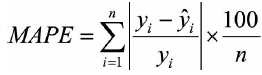
召回率（查全率）recall：分类正确的正样本数占总样本正样本数的比例。即tp/(tp+fn)

**排序问题采用P-R曲线**。排序问题中通常没有一个确定的阈值把结果直接判定为正/负样本，而是采用TOP N返回结果的P-R值来衡量模型的性能，即TOP N结果就是正样本，然后计算前N个位置P-R值。

P-R曲线的含义：曲线上的一个点代表在某一阈值下，模型将大于该阈值的结果判断为positive，小于该阈值的为negative，此时返回对应的P和R。只用某个点对应的PR值是不能全面地衡量模型性能，只有通过PR曲线的整体表现，才能够对模型进行更全面的评估。

F1 score：P和R的调和平均值，F1=2\*P\*R/(P+R)

均方根误差RMSE：gt值与预测值之差的平方和，求平均值后开方。

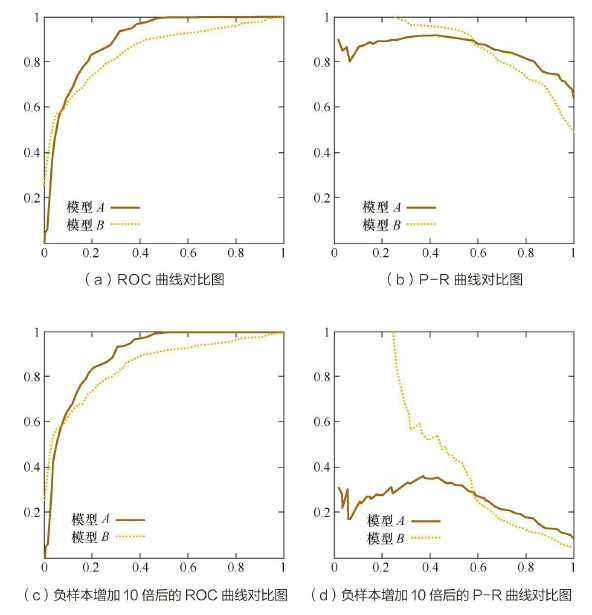
**回归模型采用RMSE**.一般情况RMSE可以很好返回模型预测精度，但是RMSE容易受到outlier影响，即使outlier数量很少，也会让RMSE变差。比RMSE鲁棒性更好的metric时mean absolute percent error (MAPE): 。它相当于把每个点的误差进行了归一化，降低outlier带来的绝对误差的影响。

ROC曲线（Receiver Operating Characteristic）。横坐标为假阳性率FPR，纵坐标为真阳性率TPR。FPR=FP/N, TPR=TP/P, 其中N为GT的negative数量，P为GT的positive数量。曲线的绘制：对于二元分类来说，预测结果会输出预测概率，对于正负样本的预测，我们会采取一个阈值，大于阈值的为正，小于的为负。当这个阈值从正无穷大一直减小时，会产生不同的点对(FPR, TPR)，把这些点连起来就可以形成ROC曲线。

AUC：ROC曲线下的面积大小。能够量化地基于ROC衡量模型性能。在曲线下面求积分就可以得到面积值。AUC一般在0.5~1之间。AUC越大，说明分类器越可能把真正的正样本排在前面，分类性能越好。

***ROC曲线和P-R曲线比较***：

当正负样本的分布发生变化时，ROC曲线的形状基本能够保持不变，P-R曲线的形状一般会发生较剧烈的变化。ROC曲线的这个特点能够尽量降低**不同测试集**带来的干扰，因为很多问题中正负样本数量往往很不均衡。ROC曲线的适用场景更多，常用于排序，推荐，广告领域。但是这两个曲线的选择因问题而异，如果希望更多地看到模型在特定数据集上的表现，P-R曲线则更直观。



余弦相似度：分析两个特征向量之间的相似性，即向量夹角的余弦值。范围[-1, 1]。相似度为1表示相同，反之。

余弦距离：将1减去余弦相似度，范围[0, 2]。距离为0表示相同。

***为什么有些场景要使用余弦相似度而不是欧式距离***？

余弦相似度更多关注的是两个向量的夹角，而不关心它们的绝对大小。如果一对文本内容相近，但长度差距很大，采用词频或者词向量作为特征，它们在特征空间中的欧式距离通常很大。而是用余弦相似度更能表征它们的相似情况。另外，文本，图像和视频中，对象的特征维度很高，采用余弦相似度仍然可以限制在[-1,1]范围内。而欧式距离的数值受到维度的影响，范围不固定。如果欧式距离是归一化的，如Word2Vec中，选择距离最小（相似度最大）的近邻，使用两者是相同的结果。

严格意义上的距离定义需要满足三个条件：（1）正定性；（2）对称性；（3）三角不等式。余弦距离不是严格意义上的距离，因为它不满足三角不等式。

A/B测试：验证模型最终效果的手段。

***如何进行线性A/B测试***？

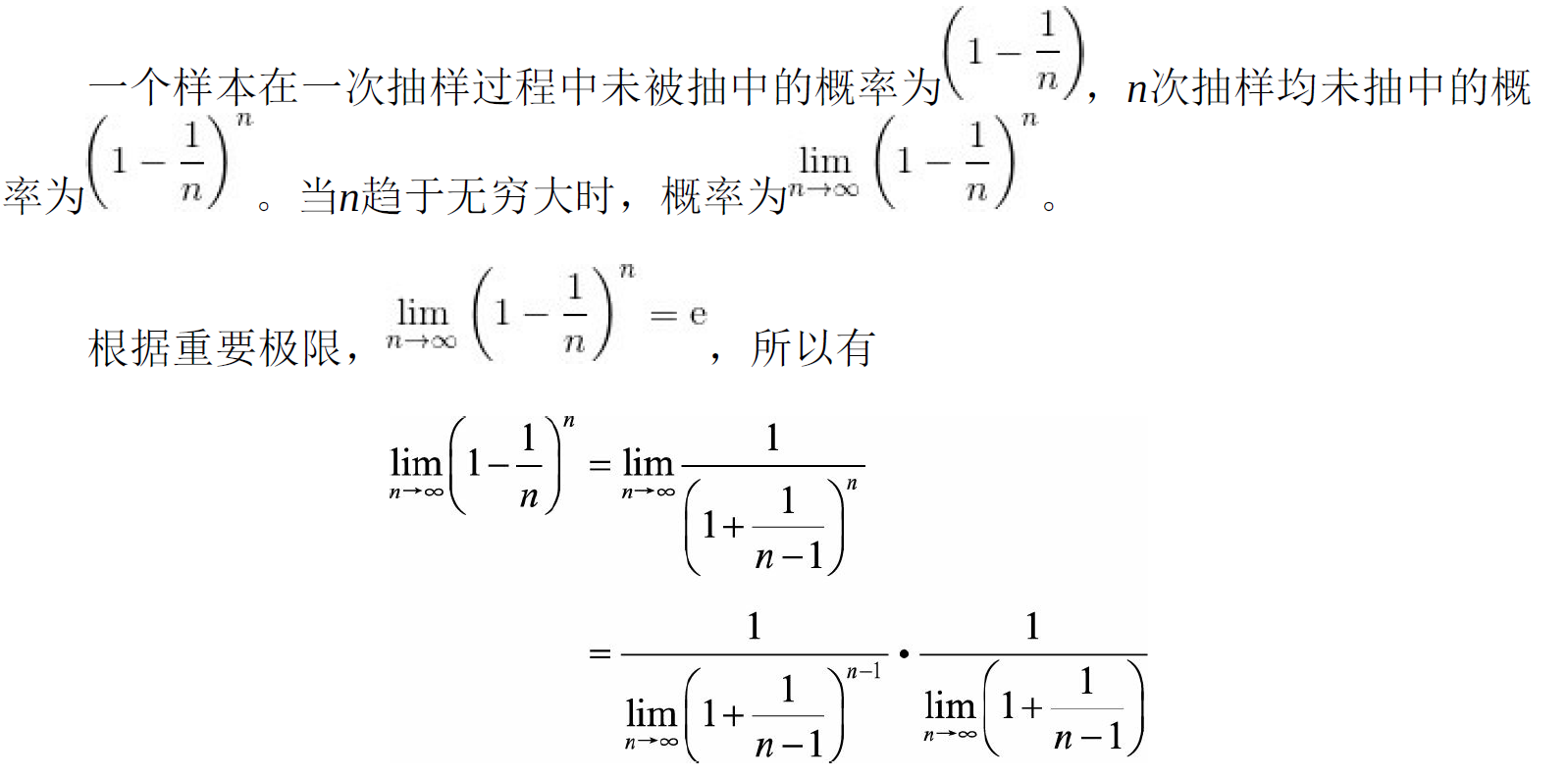
进行用户分桶，将用户分成实验组和对照组。对实验组用户使用新模型，对对照组用户使用旧模型。在分桶过程中，要注意样本的独立性和采样方式的无偏性，确保同意给用户每次只能分到一个桶中，在分桶过程中所选取的user\_id需要是一个随机数，这样才能保证样本无偏。

***在模型评估过程中，有哪些主要验证方法，优缺点是什么***？

（1）Holdout检验：最简单直接。样本集合随机分成训练集和验证集。然后用其它metric评估。缺点是验证机上计算出来的最后评估指标与原始分组有很大关系，也就是说随机性对结果影响比较大。

（2）交叉检验：k-fold，holdout的改善版，可消除随机性的影响。最好把k次评估结果的均值作为最终评估指标。留p验证：每次留下p个样本作为验证机，其余所有样本作为测试集。从n个元素中选择p个元素有C(p.n )可能，计算成本非常高，所以很少在实际中应用。

（3）自助法（bootstrap）：上述两种验证法对于样本量小的情况不是很适用，因为样本量本来就少，再细分下去，训练样本太少容易导致过拟合。因此可以使用bootstrap的方法，它属于自助采样法。对于总数为n的样本集合，进行n次有放回的随机抽样，得到大小为n的训练集。在n次采样过程中，有的样本会被重复采样，有的样本没有被抽出过，将这些没有被抽出的样本作为验证集，进行模型验证。

***上述采样过程中，当n趋于无穷大时，最终有多少数据从未被选择过***？ 结果为1/e，上面的重要极限有误。

**超参调优**：grid search，random search，Bayes optimization。超参搜索算法要素：（1）目标函数；（2）搜索范围，如上下限；（3）算法其它参数，如步长。

网格搜索：最简单应用最广。查找搜索范围内的所有点来确定最优值。如果采用较大的搜索范围以及较小的步长，大概率能找到全局最优值。但是这种方法耗时耗资源。因此，实际中会先用较大范围及较大步长先找出全局最优的大概位置，然后再减小范围和步长进行精确搜索。但是由于目标函数一般是非凸的，很可能会错过全局最优。

随机搜索：没有上界和下界，在搜索范围中随机选取样本点。理论依据是，如果样本点集足够大，则通过随机采样也能大概率找到全局最优或近似值。比网格搜索快一点。但结果也是无法保证的。

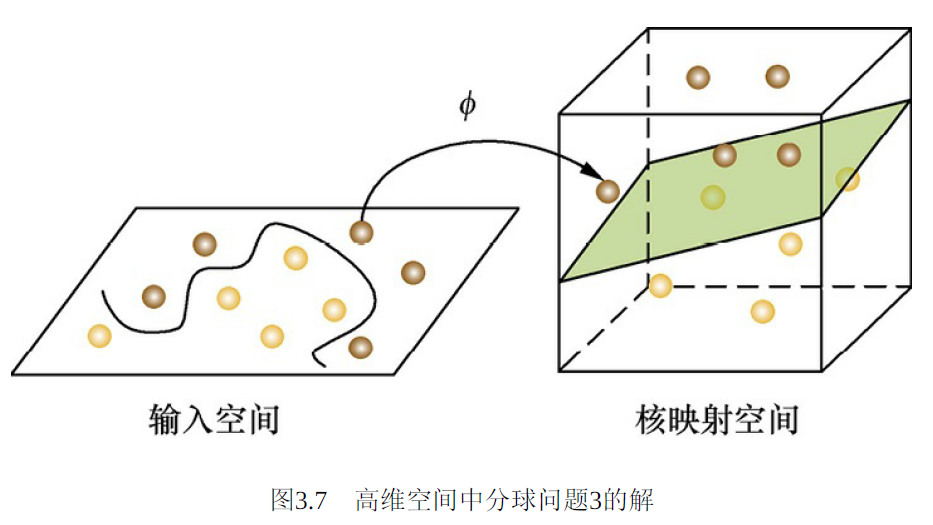
贝叶斯优化算法：与上述两种搜索完全不同，它们会忽略前一个搜索样本的信息，贝叶斯法则充分利用了之前的信息。具体过程：根据先验分布，假设一个搜集函数；每次使用新的采样点测试目标函数，利用这个信息更新目标函数的先验分布；最后，由厚颜分布给出全局最优可能出现的位置点。缺点：一旦找到了一个局部最优，它会在该区域不断采样，然后陷入局部最优。为了弥补这个缺陷，贝叶斯法会选择在未取样的区域采样，然后根据后验分布在最可能出现全局最优的区域采样。

**过拟合优化**：（1）数据增广；（2）降低模型复杂度：减少网络层数，神经元个数，例如dropout；在决策树中降低树的深度，进行剪枝等。（3）正则化。（4）集成学习。

欠拟合优化：（1）添加新特征；（2）增加模型复杂度；（3）减小正则化系数。

**第3章 经典算法**

**SVM：**



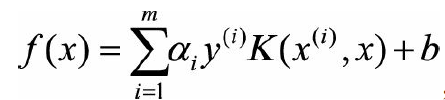
***在空间上线性可分的两类点，分别向SVM分类的超平面上做投影，这些点在超平面上的投影仍然是线性可分的吗***？

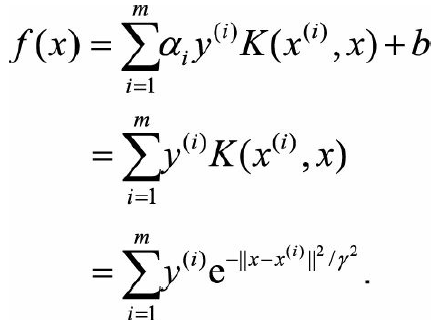
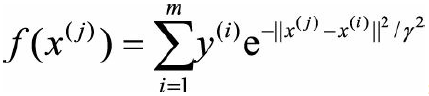
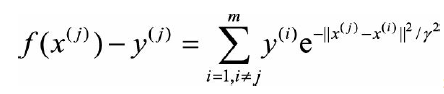
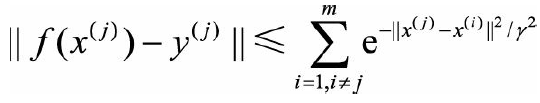
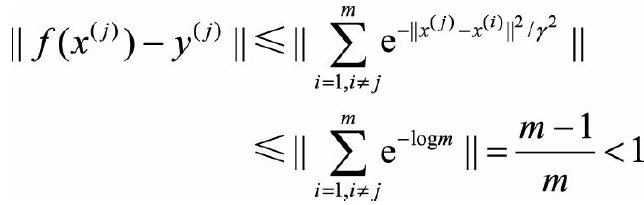
不是。反例：假设有两类，各自有一个点，则SVM超平面就是这两个点连线的中垂线，当两个点往SVM超平面（线）投影时，就会得到相同的点。

SVM的分类结果仅依赖于支持向量，这也是SVM拥有极高运行效率的关键之一。

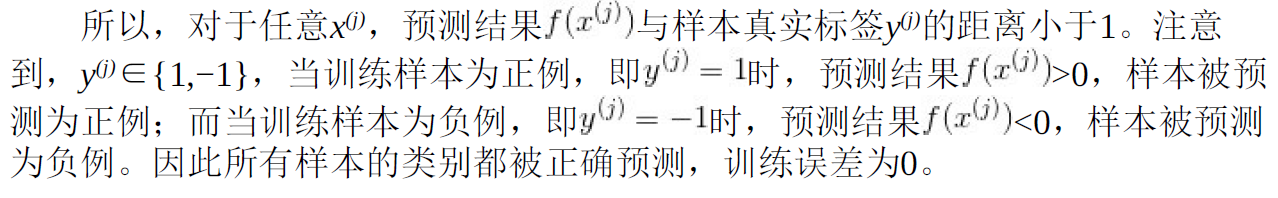
**是否存在一组参数使SVM训练误差为0**？

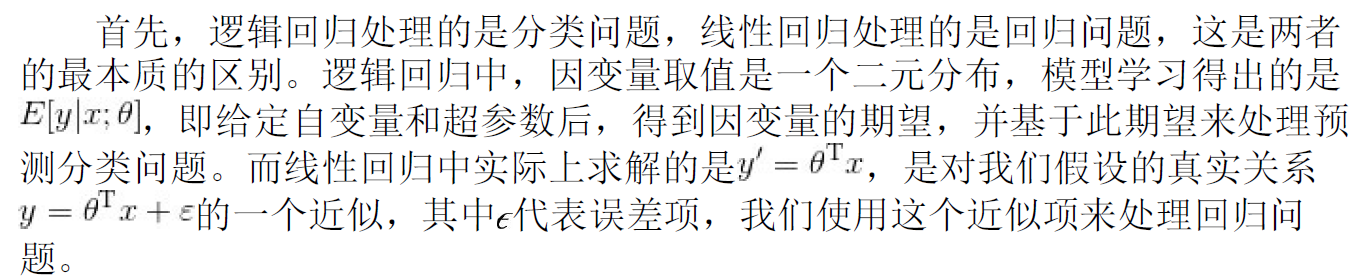
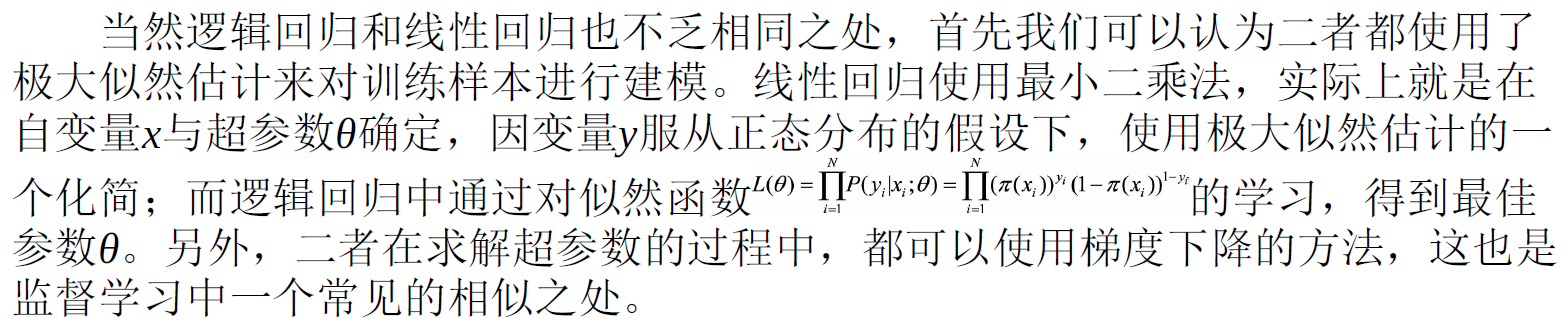
若给定训练集中不存在两个点在同一位置，则存在参数α，γ，b使得SVM的训练误差为0.

证明：SVM预测公式为，高斯核为。由于不存在两个点在同一位置，因此对于任意的i≠j，xi-xj>=ε。对任意i，固定α=1和b=0，只保留γ。则有，

代入任意xj🡪🡪 🡪 🡪令🡪 

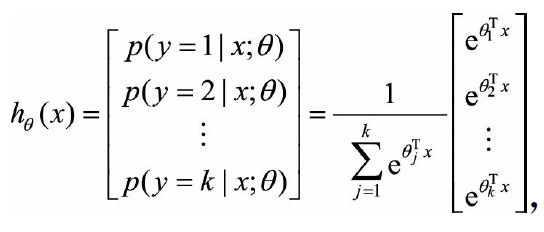
所以对任意x，预测结果于整数label的距离小于1。



***当使用逻辑回归处理多标签的分类问题时，有哪些常见做法和应用场景***？

Depends.对于一个样本只对应要给标签，可以假设每个样本属于不同标签的概率服从集合分布，使用softmax regression来分类。当存在样本可能属于多个标签的情况时，可以训练k个二分类的逻辑回归分类器。第i个分类器用于区分每个样本是否可以归为第i类，训练该分类器时，需要把标签重新整理为第i类标签与非第i类标签两类。

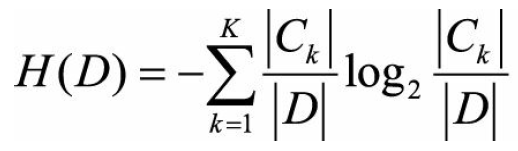


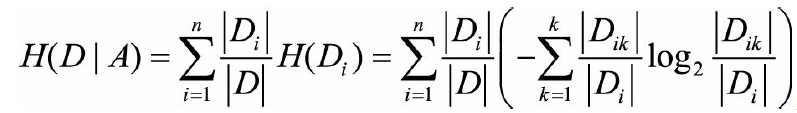
决策树作为最基础、最常见的有监督学习模型，常被用于**分类问题**和**回归问题**，在市场营销和生物医药等领域尤其受欢迎，主要因为树形结构与销售、诊断等场景下的决策过程十分相似。将决策树应用集成学习的思想可以得到随机森林，梯度提升决策树等模型。

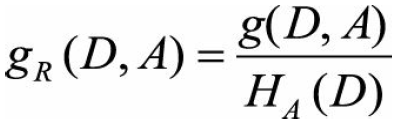
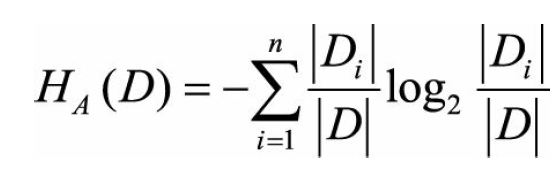
**决策树生成过程**：（1）特征选择；（2）树的构造；（3）树的剪枝

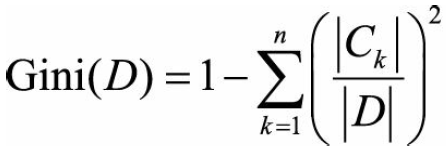
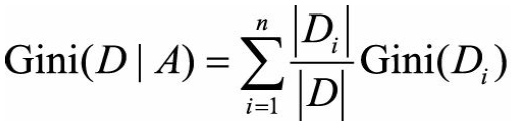
常用决策树算法：ID3, C4.5, CART为三种特征选择方法

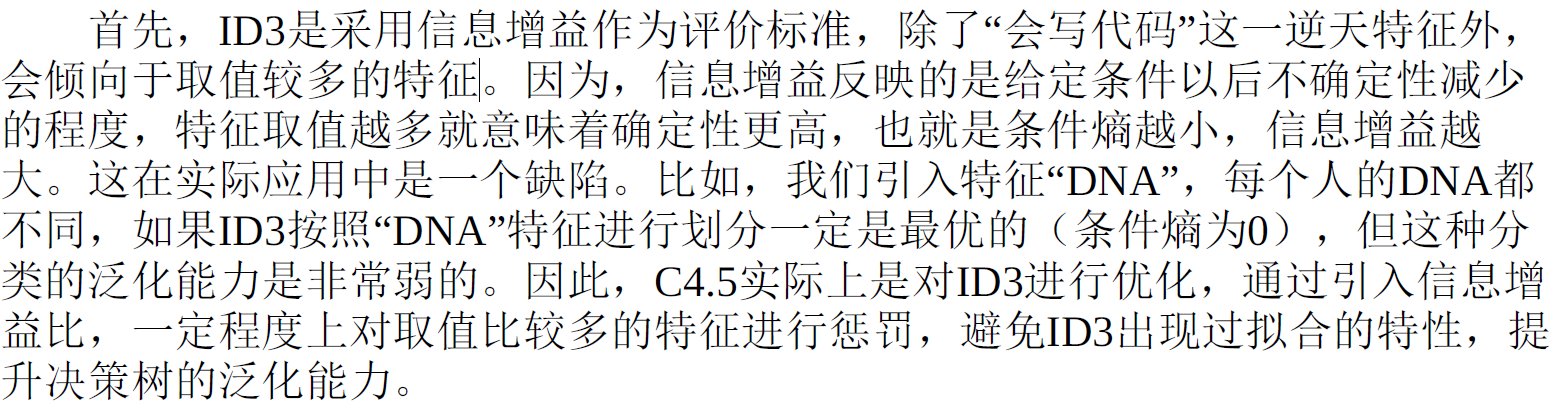
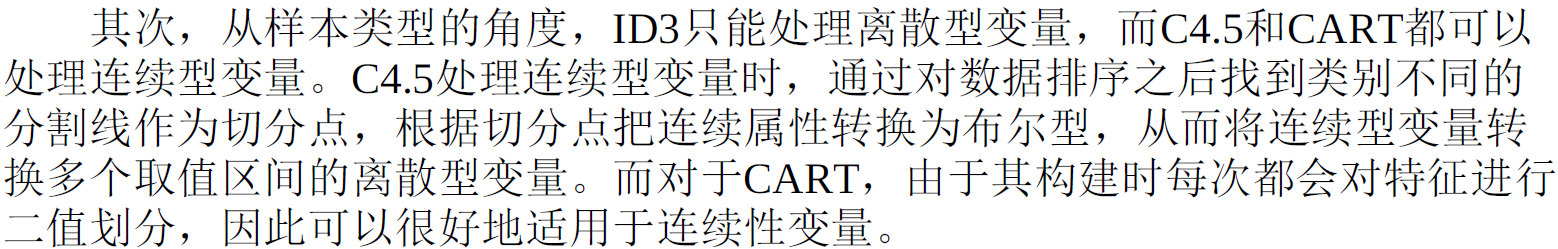
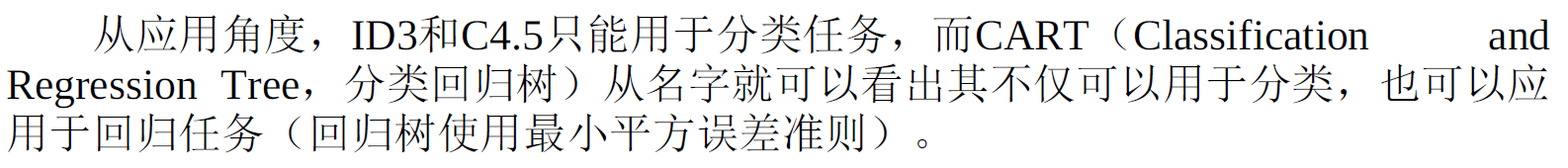
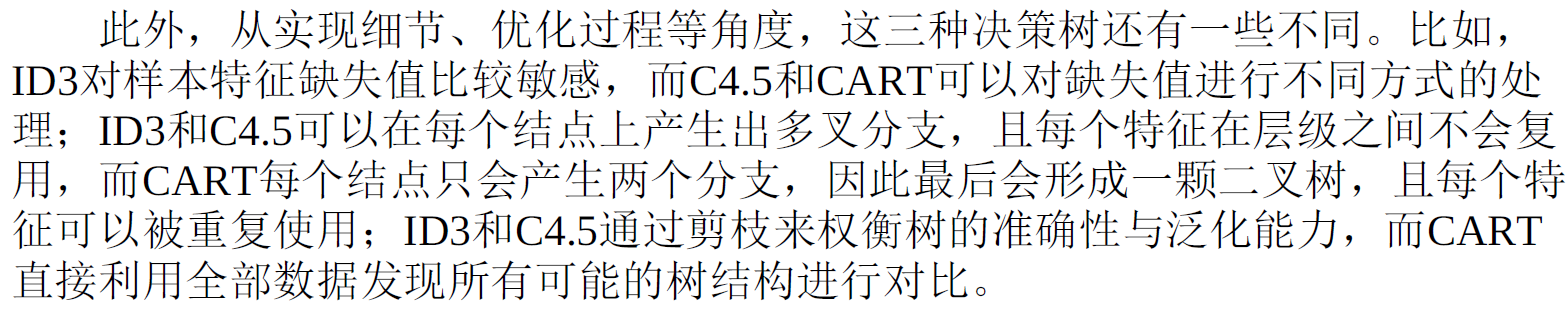
**ID3算法-最大信息增益**：对于样本集合D，class为K，D的经验熵表示为：

，其中Ck是D中属于第k类的样本子集，|Ck|表示子集个数，|D|表示样本集总个数。计算某个特征A对D的经验条件熵为：

，其中Di表示D中特征A取第i个值的样本子集，Dik表示Di中属于第k类的样本子集。所以信息增益g可以为两者之差。。注：书中的例子对理解公式很有帮助。

**C4.5-最大信息增益比**：特征A对于数据集D的信息增益比定义为：，其中，成为D关于A的取值熵。

CART(classficiation and regression tree)-最大基尼指数（Gini）：Gini描述的是数据的纯度，与信息熵含义类似。，CART在每一次迭代中选择基尼指数最小的特征及其对应的切分点进行分类，但与ID3,C4.5不同的是，CART是一棵二叉树，采用二元切割法，每一步将数据按特征A的取值切分两份，分别进入左右子树。特征A的Gini指数定义为：

***如何对决策树进行剪枝***？

完全生长的决策树会导致过拟合。常用剪枝有pre-prunning(生成决策树的过程中提前停止树的增长)和post-prunning（在已生成的决策树熵剪枝，得到简化版决策树）。

预剪枝：设定一定条件让树停止增大。（1）设定树的深度；（2）设定叶子节点的样本量阈值；（3）计算每次分裂对测试机的准确度提升，如果小于阈值，则停止。

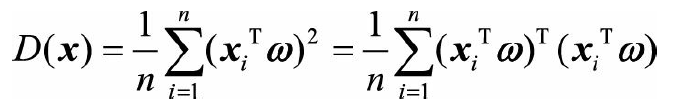
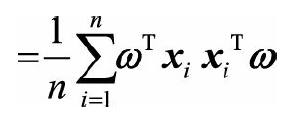
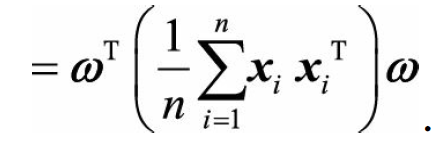
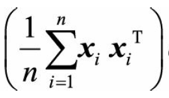
**注**：这些条件的设定比较依赖于经验。

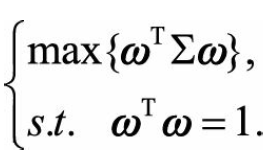
后剪枝：在决策树生成后，自底向上进行剪枝。剪枝过程将子树删除，用叶子节点替代，该节点的类别同样按照majority vote原则进行判断。同样也可在测试集熵的准确率进行判断，如果剪枝有利提升准确率，就实行。

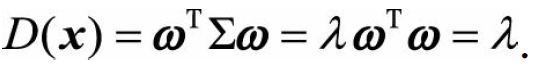
**第4章 降维**

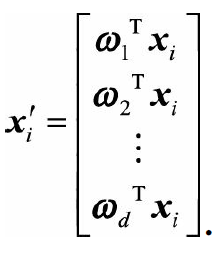
***如何定义主成分？从这个定义出发，如何设计目标函数使得降维达到提取主成分的目的？针对这个目标函数，如何对PCA问题进行求解***？

（1）假设有数据点分布在三维空间中，如果用xyz表示，则数据是三维的，但实际上这些数据点都分布在一个平面上，如果我们能够旋转xy平面使其与数据平面重合，那么数据就可以降维到x‘y’平面，那么这两个新轴就是我们要找的主成分。（2）从信噪比出发，信噪比越大表示数据质量越好，因此可以引出PCA的目标，就是最大化投影方差，也就是让**数据在主轴上投影的方差最大**。所以，目标就是找到一个投影方向，使得数据点在这个投影方向的投影方差尽可能大。向量的投影可以用两个向量的点积进行表示。故投影后的方差可以表示为：

  ，其中就是样本协方差矩阵，可以写为Σ。由于ω是单位方向向量，问题可以转化为求解一个最大化问题，即

，引入拉格朗日乘子，并对ω求导令其等于0，可以推出Σω=λω。则有，

可以看出，x投影后的方差就是协方差矩阵的特征值。我们要找到最大的方差也就是找到协方差矩阵的最大特征值，最佳投影方向就是最大特征值对应的特征向量。

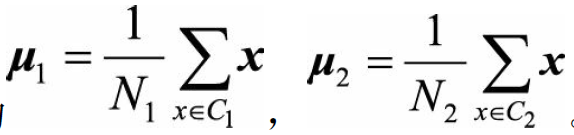
**PCA求解流程**：（1）数据去中心化处理；（2）求样本协方差矩阵；（3）对协方差矩阵进行特征值分解，将特征值从大到小排列；（4）取特征值前d大对应的特征向量，通过以下映射将n维样本映射到d维。

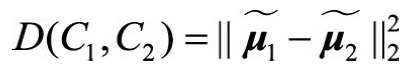
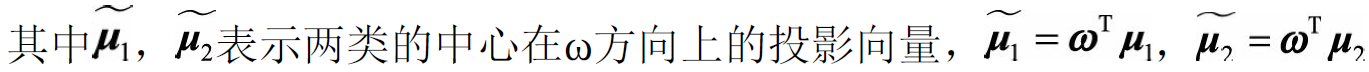
**PCA局限性**：无监督线性降维。可以通过核映射对PCA进行扩展获得核主成分分析，即KPCA

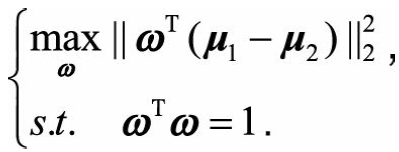
**LDA(Linear Discriminant Analysis)**：有监督学习算法，常被用来数据降维。中心思想是最大化类间距离和最小化类内距离。

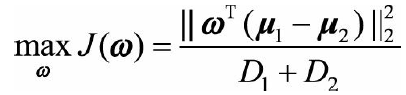
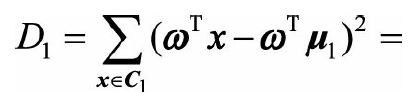
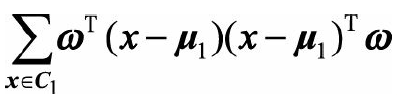
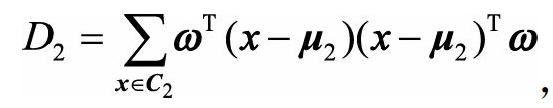
***对于具有class label的数据，如何设计目标函数使得降维过程中不损失class info？如何求解***？

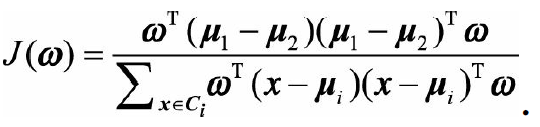
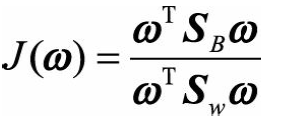
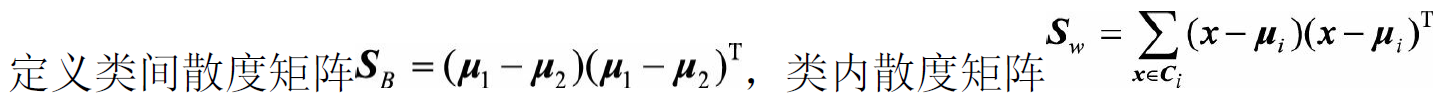
LDA首先是为了分类服务的，因此之哟啊找到一个投影方向，使得投影后的样本尽可能按照原始class分开。

，为两个类别样本均值，目标是投影之后两类之间距离尽可能大，距离为

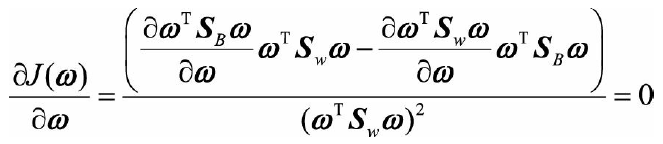
 

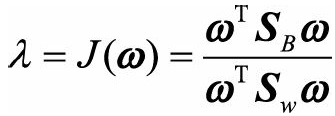
则需要优化的问题为：，当ω与u1-u2方向一致时，距离最大。但是这样还不够，我们还需要最小化类内距离。可以通过优化类内方差，使其尽可能小。我们将整个数据集的类内方差定义为各个分类的方差之和，将目标函数定义为类间距离和类内距离的比值，可得，

，  

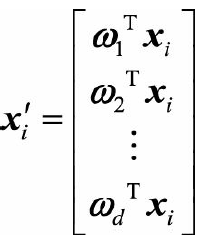
🡪 , 

最大化J(ω)，对ω求偏导并令其为0：

🡪 🡪 

，令，可得

，因此我们最大化的目标对应了一个矩阵的特征值，所以LDA降维变成了一个求矩阵特征向量的问题。J(ω)对应了矩阵最大的特征值，而投影方向就是这个特征值对应的特征向量。

**LDA求解过程**:（1）计算数据集中每个class样本的均值向量uj，以及总体均值向量u；（2）计算类内三度矩阵Sw，全局散度矩阵St，得到类间散度矩阵Sh=St-Sw；（3）对矩阵进行特征值分解，将特征值从大到小排列；（4）取特征值前d大的对应的特征向量，通过以下映射将n维样本映射到d维

***PCA和LDA的相似与差别***？

（1）目标函数：PCA选择的是投影数据方差最大的方向。由于它是无监督的，因此PCA将设方差越大，信息量越多，从而降维；LDA选择的是投影后类内方差小、类间方差大的方向。它用到了class label信息，为了找到数据中具有discriminant的维度，使得原始数据在这些方向上投影后区分class；（2）举例：语音降噪用PCA，因为噪声方差小，用PCA可以去掉方差小的维度数据；如果要区分人声，则 应该使用LDA进行降维，使每个人的语音信号具有区分性。（3）总之：对无监督的任务使用PCA降维；有监督的应该用LDA降维。