姓名：化佳博　　学号：160600229　　2019.5.29 考试

大数据分析基础考试用纸

机器学习：人工智能的分支，研究机器模拟人类的学习行为，以获取新知识技能，改善自身性能（并识别已有知识的学科。

面临的挑战：1.高维特征空间与样本树 2.寻找最优解困难 3.可解释性差 4.大数据量难计算

三个基本概念：1.局部感受野 2.混合OR池化 3.共享权重。

传统神经网络：采用反向传播进行：采用迭代来训练整个网络，随机设定初值，计算输出，根据与Label差距修改参数直至收敛

深度神经网络：分层计算。原因为反向传播，对于深层网络，残差传至前面已太小，出现梯度扩散。若对所有层同时训练，则复杂度太高(时间)，若只训练一层，偏差会逐层传递，会有过拟合问题。

数据：加工，计算，分析，应用（数据源是每个人）

大数据：无法在一定时间内用常规软件工具对其内容进行抓取管理与处理的数据集合：海量+复杂类型。

核心特征：海量（Volume）+类型多（Variety），价值密度低（Value），速度快时效高（Velocity），真实性（Veracity）。

大数据&云计算：SaaS：分布式数据挖掘。Paas：分布式处理&分布式数据库。IaaS：云存储，虚拟化

大数据的意义：不在于掌握量大，在于专业化处理数据（产业：盈利在于提高加工能力实现data增值）。

挑战与核心内容：挑战：存储、分析，管理。核心：大数据案例，大数据技术。

统计学习的方法：监督学习、非~、半~、强化学习，三要素：模型(非概率模型)+策略+算法=方法。

Data(训练集+测试集)⇒假设空间(Model) M1, M2, M3… ⇒(评价案例)最优模型。

实现统计学习的步骤：1.得到一个有限的训练集数据集合 2.确定包含所有可能模型的假设空间（学习模型集合）3.确定选择准则(策略) 4.实现求解最优模型的算法(算法) 5.通过学习方法选择最优模型 6.利用学习的最优模型对数据分析OR预测。

监督学习任务：学习一个模型，使其能够对任意给定的输入，对其输出作出较好的预测

回归问题：输入空间：$X=\{x\}$可能的集合。实例由特征向量表示：$x=\{x^{(1)}\cdots x^{(n)}\}^T$，输出空间：$Y=\{y\}$可能的集合。样本输入输出对：$(x_i,y_i)$，训练集表示 $T=\{(x_1,y_1)\cdots(x_n,y_n)\}$。输入输出均为连续变量的预测问题。分类问题：输出为有限个离散变量的问题。

监督学习基本假设：1.假设X,Y有联合概率分布P(X,Y)，但联合概率分布定义未知 2.假设训练集是依联合分布(同分布)独立

损失函数/代价函数：度量模型一次预测的好坏 $L(Y,f(X))\geq 0$ 风险函数/期望损失：平均意义下预测的好坏

常用损失函数：0-1：$\begin{cases}0, Y=f(x)\\1, Y\neq f(x)\end{cases}$，平方：$(Y-f(x))^2$，绝对：$|Y-f(x)|$，对数：$L(Y,P(Y|X))=-\log P(Y|X)$

⇒风险函数 $R_{exp}(f)=E_p[L(Y,f(X))]=\int_{X\times Y}L(y,f(x))P(x,y)dxdy$　　最小值 $\min_{f\in F}R_{exp}(f)$ △

平均损失/经验风险/经验损失：$R_{emp}=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$；$\lim_{N\to\infty}R_{emp}=R_{exp}$，故 $R_{emp}\approx R_{exp}$

经验风险最小化(ERM)：样本足够大时效果较好(样本容量小⇒(正则化)⇔结构风险最小化(SRM)： △

监督学习：给定的训练集T，选假设空间，损失函数 $\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$ △

模型评估&模型选择：1.给定训练集T 2.↙ 3.确定模型(训练误差确定) 4.评估模型：测试集→测试误差

测试误差反映了(学习方法对未知的测试数据集的预测能力)泛化能力

过拟合：学习时选择模型包含参数过多，以至于对这一模型对已知数据预测很好，但对未知数据预测反而很差

一味追求提高对训练集预测能力，所选模型复杂度往往比真模型高（选择适当复杂度的模型，达到测试误差最小的目的）

交叉验证：重复使用数据，把其切分，再组合为训练集与测试集，在此基础上反复训练、测试试与预测。

可以将和《大数据分析基础》课程有关的内容写在上面。考试期间仅允许携带这一张纸。

if-Then规则的性质：互斥&完备 模型：条件概率模型 损失函数：正则化的极大似然函数 学习策略：损失函数是目标函数最小化

决策树表示：给定条件下类的条件概率分布（这一分布定义在特征空间的一个划分上）.

学习目标：根据训练集构建决策树模型，使其可以对实例正确分类. 本质：从训练集中归纳出分类规则

算法：递归地选择最小化特征值，并根据该特征对训练数据进行分割（生成时局部取优，剪枝时全局取优）

熵：随机变量不确定性的度量 $H(X)=-\sum_{i=1}^{n} p_i \log p_i$ 信息增量：熵不确定性减少程度 $g(D,A)=H(D)-H(D|A)$.

$H(D|A)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log\frac{|D_{ik}|}{|D_i|}$ 增益比 $g_R(D,A)=\frac{g(D,A)}{H_A D}$，$H_A D=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$

ID3：核心：在决策树各结点上应用信息增益准则选择特征递归来构造决策树 C4.5：用信息增益比选择特征.

决策树的剪枝：通过极小化决策树整体的损失函数或代价函数来实现，损失函数 $C_\alpha(T)=\sum_{t=1}^{|T|}H_t(D)+\alpha(|T|)=C(T)+\alpha|T|$ △

分类与回归树：(CART, Breinary) 预测误差 Σ△ 回归树：平方误差最小化准则；分类树：Gini，index 最小化预测. $+\frac{|D_2|}{|D|}Gini(D_2)$

回归树模型：$f(x)=\sum_{m=1}^{M}c_m I(x\in R_m)$ 最优值：$c_m=ave(y_i|x_i\in R_m)$ △ 联合后 $Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)$

Gini指数：分类问题中假设有K个类，样本点属于第k类的概率为 p_k，则 $Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$ 平方误差或Gini最小最优

两个要素：1.因变量Y必须是连续型变量或近似连续变量 2.研究X,Y线性相关关系的模型.

线性回规模型：$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$. 最小二乘法：寻找参数 β_0,β_1 估计，使__差平均和达最小，t检验，F检验.

标准化残差：$ZRE_i=\frac{e_i}{\hat{\sigma}}$ 学生化残差：$SRE_i=\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$△ $h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{L_{xx}}$. 变量选择RMSQ，Cp，AIC，BIC准则

模型预测：单侧预测，区间预测，0-1回归核心思想：0-1数据→可能性 $p(Y=1)\to\frac{p}{1-p}\to\ln\frac{p}{1-p}$.

定序回归思想：1.放弃直接建立定序数据Y与X之间相关关系 2.建立潜变量Z与X之间的普通线性回归关系 3.基于Z与X之间线性回归关系推演Y与X的回归关系（NB：1.无数值意义，不能代数运算 2.顺序重要）

计数回归注意：1.计数回归是非负的无上界整数 2.处理计数数据的第一选择，泊松分布

生存回归：1.生存数据普遍存在的问题：截断 2.最基本的描述方法：各种分位数.

聚类分析：对样品OR指标进行分类的一种多元统计分析方法. 目的：使类内对象的同质性最大化，和类△对象异质性最大化

基本思想：根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间相似程度的统计量. 利用统计量将样品或指标进行归类，样相似样品OR指标归一类，不相似归为其他，直到全部聚合.

数据处理方法：1.总合标准化 $x'_{ij}=\frac{x_{ij}}{\sum_{i=1}^{m}x_{ij}}$ 2.标准差标准化 $x'_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$ 3.极大值标准化 $x'_{ij}=\frac{x_{ij}}{\max_i\{x_{ij}\}}$ 4.极差标准化 $x'_{ij}=\frac{x_{ij}-\min_i\{x_{ij}\}}{\max_i\{x_{ij}\}-\min_i\{x_{ij}\}}$

距离的计算：绝对值：$\sum_{i=1}^{n}|x_{ik}-x_{jk}|$，欧氏：$\left(\sum_{k=1}^{n}(x_{ik}-x_{jk})^2\right)^{\frac{1}{2}}$，明科 $\left(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^p\right)^{\frac{1}{p}}$，切比 $p\to\infty$：$\max_k|x_{ik}-x_{jk}|$

直接聚类法：1.各分类对象单独视为一类，2.根据距离最小原则依次送出一对分类对象，贼新类，若其中一个分类对象已归于一类，则把另一个也归入，若一对分类对象正好属于已归两类，则两类归为一类，每次归并都划去该对象所在列与列序相同的行，m-1次后就可全部归为一类.

最短距离聚类法：在原 m×m 距离矩阵非对角元素中找 $d_{pq}=\min\{d_{ij}\}$，把分类对象 G_p, G_q 归并为新类 G_r. 然后按 $d_{rk}=\min\{d_{pk},d_{qk}\}(k\neq p,q)$ 计算新类间距离，可得一个新(m-1)阶矩阵，再反复进行，直到归为一类.

分析方法四种类型：基于距离、基于决策树、基于贝叶斯，规则归纳方法.

解决关键：构造合适分类△，从数据集到一组类△集的映射，一般这些类是被预先定义的非交叠的

两步骤：1.建立一个模型描述预定的数据类集或概念集 2.使用类型进行分类 $\forall C_s\in C, C_s\neq C_j$. Sim：相似性

基于距离分类算法思路：给定数据集 $D=\{t_1\cdots t_n\}$ 与一组类 $C=\{C_1\cdots C_n\}$，假设每组包括一些数值型的属性集：$t_i=\{t_{i1}\cdots t_{ik}\}$ 每个类也包含数值属性值 $C_j=\{C_{j1}\cdots C_{jk}\}$，则分配每个 t_i 到满足如下条件的类 C_j：$Sim(t_i,C_j)>=Sim(t_i,C_s)$

K-近邻分类算法：通过计算每个训练数据到待分元组的距离，取和待分类元组距离最近的K个训练数据，K个数据中哪个类别的训练数据占多数，则待分别类元组就属于哪个类别.

K-me 1.从n个数据对象任选K个作为初始聚类中心 2.循环 3.4.直到不变：3.根据每个聚类对象的均值（中心对象）计算每个对象与这些中心对象的距离，并根据最小距离重新对对象划分 4.重新计算有变化聚类的均值. 标准 $E=\sum_{i=1}^{k}\sum_{p\in C_i}|p-c_i|^2$

$(p-m_i)^2$