机器学习内容总结

说明：课本中所有算法涉及的推导与证明部分不考。在全面复习的基础上重点掌握基本概念，算法的基本思想与简单的计算。

第一章 绪论：

1.2 基本术语

- 机器学习：致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，“经验”通常是以“数据”的形式存在。机器学习研究的主要内容是在计算机上从数据中“产生模型”的算法，即“学习算法”，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，并使用该模型对新的情况给出判断（例如好瓜）
- 假设：通过学习所获得的模型，对应了数据的某种潜在规律，因此所获得的**模型**，也就是假设
- 根据**预测目标**的不同，学习任务分为 3 类：分类（预测的是离散值，例如好瓜，坏瓜）、回归（预测的是连续值，例如西瓜成熟度 0.95,0.37）和聚类（无预测值，即没有标记信息）
- 预测样本：学得模型后，被预测的样本称为预测样本。
- 根据**训练数据是否拥有标记信息**，学习任务又可以分为 3 类：监督学习（分类、回归）、无监督学习（聚类）和半监督学习（两者结合）两个空就填监督学习，无监督学习，分类和回归是前者代表，聚类是后者代表。
- **泛化能力：学得的模型适用于新样本的能力（课本：即使对聚类这样的无监督学习任务，我们也希望学得的簇划分能适用于没在训练集中出现的样本，学得模型适用于新样本的能力，称为“泛化能力”）**，一般而言，训练的样本越多越有可能通过学习获得强泛化能力的模型，具有强泛化能力的模型能很好地适用于整个样本空间

1.3 假设空间

- **归纳**和**演绎**是科学推理的两大基本手段
  - **归纳：从特殊到一般的泛化过程**，即从具体的事实归结出一般性规律
  - **演绎：从一般的特殊的特化过程**，即从基础原理推演出具体情况
- 归纳学习分广义与狭义，广义是从样例中学习，狭义是从训练数据中学得概念，狭义学习也称概念学习，概念学习中最基本的是布尔概念学习。
- 假设空间，版本空间（可能有多个假设与训练集一致，即存在一个和训练集一致的假设集合，就是版本空间）
- 假设空间的搜索过程，即不断删除与正例不一致的假设，与反例一致的假设，最终获得与训练集一致的假设

1.4 归纳偏好

- 机器学习算法在学习过程中对某种类型假设的偏好称作归纳偏好
  - **奥卡姆剃刀**是一种**常用的**、自然科学研究中最基本的原则，即“若有多个假设与观察一致，选最简单的那个”
  - **奥卡姆剃刀并非唯一可行的原则**
  - **奥卡姆剃刀本身存在不同的演绎**
- 一个算法 ζ_a 如果在某些问题上比另一个算法 ζ_b 好，必然存在另一些问题，ζ_b 比 ζ_a 好，即没有免费的午餐定理。

第二章 模型评估与选择

2.1 经验误差与过拟合

- **误差：**样本真实输出与预测输出之间的差异
  - **训练误差（经验误差）：**学习器（学习器就是模型）在训练集上的误差
  - **泛化误差：**学习器在新样本上的误差，泛化误差等于偏差、误差、噪声之和
  - （课本原文：学习器的实际预测输出与样本的真实输出之间的差异”称为误差，学习器在训练集上的误差称为“训练误差”，在新样本上的误差称为“泛化误差”）
  - **机器学习的目标**：得到**泛化误差小**的学习器
- 过拟合：若学习器把训练样本学习的太好，将训练样本本身的特点当做所有样本的一般性质，以至于把训练样本所包含的不太一般的特性都学到了，导致泛化性能下降，过拟合只能缓解不能避免，例如加正则项，优化目标函数，决策树中的剪枝
- 欠拟合：学习能力低下而造成的，对训练样本的一般性质尚未学好，训练数据较少时更容易发生欠拟合，克服欠拟合例如决策树中扩展分支，神经网络学习中增加训练轮数

2.2 评估方法

- 常见的几种模型评估方法（评估结果的稳定性和保真性）
  - **留出法：**直接将数据集划分为两个互斥集合，作为训练集和测试集
  - **交叉验证法：**首先将数据集分层采样，划分为 k 个大小相似的互斥子集（k 常取 10）；然后，每次用 k-1 个子集的并集作为训练集，余下的那个子集作为测试集，进行 k 组训练和测试，最终返回这 k 个测试结果的均值，最后将上个步骤随机使用不同的划分方式重复 p 次，最终求得 p 次均值。 特例：留一法
  - **自助法：**以**自助采样法**作为基础，对数据集 D，有放回采样 m 次，得到训练集 D’，D/D’用做测试集。没在测试集中出现的样本用于测试，这样的测试结果叫做**包外估计**

2.3 性能度量

- **回归任务**最常用的性能度量是**“均方误差”**——也就是数学上的标准差

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{x}_i)-y_i)^2$$

- **分类任务**中常用的性能度量：**错误率和精度、查准率和查全率、ROC 和 AUC、代价敏感错误率和代价曲线**

第三章

3.1 基本形式

- 线性模型：试图学得一个**通过属性的线性组合**来进行预测的**函数**，线性模型有很好的**可解释性/可理解性**

3.2 线性回归

- 目的：学得一个线性模型以尽可能准确地预测实值输出标记，**算法性能与回归函数复杂度无关**
- **最小二乘法（参数/模型估计）**：试图找一条直线，使得所有样本到直线上地欧式距离之和最小（课本：基于均方误差最小化来进行模型求解的方法称为“最小二乘法”，在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小）
- **单一线性回归目标函数**

  $f(x)=wx_i+b$ 使得 $f(x)\cong y_i$

3.3 对数几率回归（基本思想）

- **用线性回归模型预测结果去逼近真实标记的对数几率，对应的模型（实际上是一种分类学习方法）**

3.4 线性判别分析 LDA（基本思想）

- **是一种经典的线性学习方法，给定训练样例集，设法将样例投影到两条直线上，使得同类样例的投影的尽可能接近，异类样例的投影点尽可能远离。在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别**

3.6 类别不平衡问题（要会举例子）

- **不同类别训练样例数相差很大的情况（正类为小类）**

例如有 998 个反例，但是正例只有 2 个，那么学习方法只需要返回一个永远将新样本预测为反例的学习器，就能达到 99.8%的精度；然而这样的学习器往往没有价值，因为它不能预测出任何正例

第四章 决策树

4.1 基本流程

- 决策树的结构：**一般一颗决策树包含一个根节点（包含样本全集）、若干内部节点（每个内部节点对应一个属性测试）以及若干叶节点（每个叶节点对应一个决策结果），每个节点包含的样本集合根据属**

性测试的结果被划分到子结点中，从根节点到每个叶节点的路径对应了一个判定测试序列

- 目的：为了产生一颗泛化能力强，即处理未见示例能力强的决策树，其流程遵循简单且直观的**分而治之策略**

4.2 划分选择（基本运算）——计算题

- 决策树学习的关键在于如何选择最优划分属性
- 一般而言，随着划分过程不断进行，我们希望决策树的分支节点所包含的样本尽可能属于同一类别，即结点的纯度越来越高
- 经典的**属性划分方法：信息增益（ID3），增益率，基尼指数**

4.3 剪枝处理

- 一定程度避免**因决策分支过多**，防止训练集把自身的一些特点当做所有数据都具有的一般性质而导致的过拟合，**降低过拟合风险**
- 剪枝的基本策略：**预剪枝和后剪枝**

注：决策树的**核心技术**为划分选择和剪枝处理

第五章 神经网络

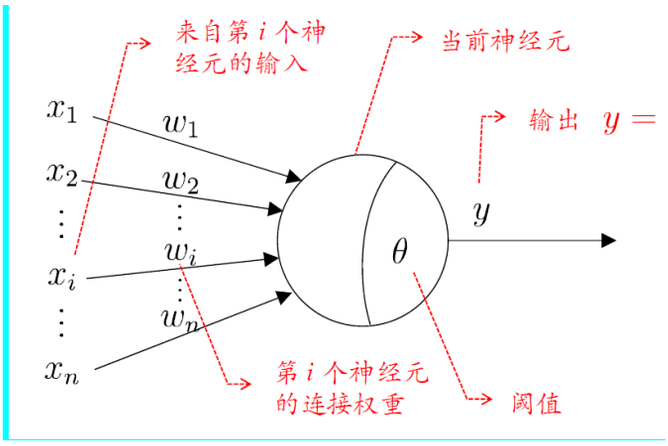
5.1 神经元模型

- 神经网络：是由具有适应性的**简单单元**组成的**广泛并行互联**的网络，它的组织能够**模拟生物神经系统**对真实世界物体所作出的反应。**神经网络中最基本的成分是神经元**
- M-P 神经元模型：**神经元接收到来自 N 个其他神经元传递过来的输入信号，这些输入信号通过带权重的连接进行传递，神经元接收到的总输入值将与神经元的阈值进行比较，然后通过激活函数的处理得到输出**

课本原文：神经元接收到来自 n 个其他神经元传递过来的输入信号，这些输入信号通过带权重的连接进行传递，神经元接受到的总输入值将与神经元的阈值进行比较，然后通过“激活函数”处理以产生神经元的输出。

- ✓ 输入：来自其他 *n* 个神经元传递过来的输入信号
- ✓ 输出：通过激活函数的处理以得到输出

输出 $y = f\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$



-

5.2 感知机与多层网络

- 感知机由**两层神经元**组成，输入层输入信号后传递给输出层，输出层是 M-P 神经元
- 感知机能够容易地实现逻辑与、或、非运算

5.3 误差逆传播算法/BP 算法（算法的主要思想，缓解过拟合）——训练多层网络

- **最成功的训练多层前馈神经网络的学习算法**
- 主要思想：首先将输入示例提供给输入层神经元，然后逐层将信号前传，知道产生输出层的结果，然后计算输出层的误差，再将误差逆向传播至隐层神经元，最后根据隐层神经元的误差来对连接权和阈值进行调整，该法带过程循环进行，知道达到某些停止条件为止

- BP 算法基于梯度下降策略，以目标的负梯度方向对参数进行调整

- BP 算法缓解过拟合两种策略：早停和正则化

- BP 网络及算法中的变量符号，理解

✓ 给定一个拥有 *d* 个输入神经元, *l* 个输出神经元, *q* 个隐层神经元的多层前向前馈网络结构.

✓ 第 j 个输出神经元的输入

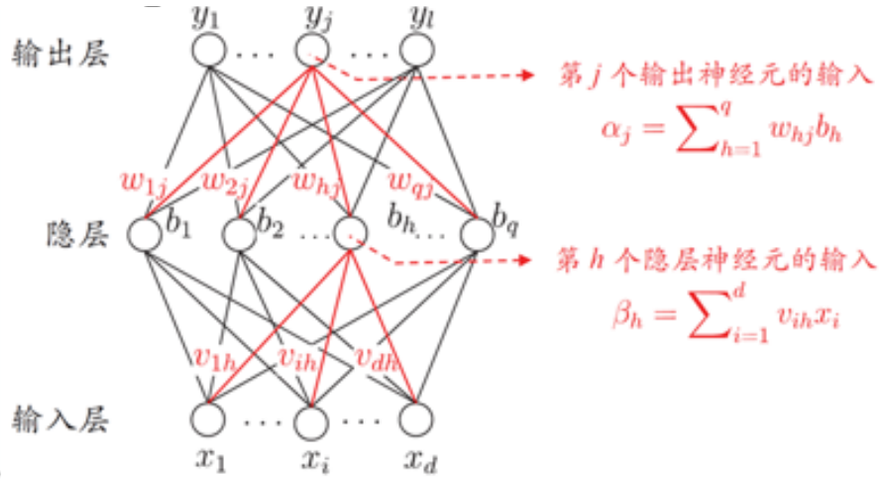
✓ 第 h 个隐层神经元的输入

① 首先，将输入示例提供给输入层神经元，
然后逐层将信号前传，直到产生输出层的结果;

② 然后，计算输出层的误差，再将误差逆向传播至隐层神经元

③ 最后，根据隐层神经元的误差来对连接权和阈值进行调整
该法代过程循环进行，直到达到某些停止条件为止

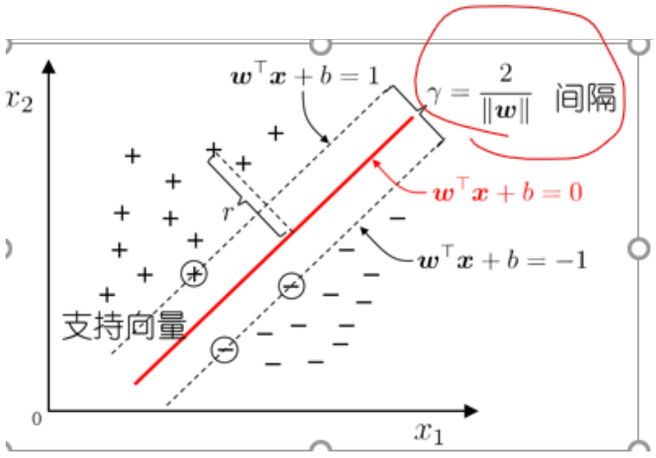
-



第六章 支持向量机

6.1 间隔与支持向量机

- **支持向量机最基本的思想**就是基于训练集 D 在样本空间中找到一个划分超平面，将不同类别的样本分开，使得两类样本中，距离超平面最近的样本之间的间隔最大。
- **线性模型：在样本空间中寻找一个超平面，将不同类别的样本分开**
- **支持向量机：**寻找具有最大间隔的划分超平面$\omega^T x+b=0$（与第三章的线性模型区别开），*ω（w1；w2；…;wd）是法向量，决定超平面的方向，b 是位移项，决定了超平面与原点之间的距离，寻找这两个参数使得间隔最大*
- **支持向量机一个特性：**训练完成后，大部分的训练样本都不需要保留，最终模型只与支持向量有关
- 距离超平面最近的几个训练样本，称作**支持向量**
- **两个异类支持向量到超平面的距离之和称作间隔**
- **下图中的红色圈出来的公式**



- **最小化最大间隔——公式 6.6 记住，Page 123（SVM 基本模型）**

6.2 对偶问题——计算 **Page 123 公式 6.**

- **对公式 6.6 使用拉格朗日乘子法得到公式 6.8 在化简就得到公式 6.11**

（老师画这里了，怎么考不确定）

- 拉格朗日函数
- 对偶问题公式

6.4 软间隔与正则化

- 软间隔基本想法：最大化间隔的同时，让不满足约束的样本应尽可能少

第七章 贝叶斯分类器

7.1 贝叶斯决策论

- 在所有相关概率都已知的理想情形下，**贝叶斯决策论**考虑如何基于这些概率和误判损失来选择最优的类别标记

7.2 极大似然估计

7.3 朴素贝叶斯分类器——计算题

- **朴素贝叶斯分类器原理：**采用了**“属性条件独立性假设”**，即对已知类别，假设所有属性相互独立，换言之，假设每个属性独立地对分类结果发生影响
- 属性条件独立性假设 $P(c|x)$ 可写为：

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

其中 d 为属性数目，x_i 为 x 在第 i 个属性上的取值。

由于对所有类别来说 $P(x)$ 相同，因此基于贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

7.4 半朴素贝叶斯分类器

- 基本想法：**对属性条件独立性假设进行一定程度的放松，适当考虑一部分属性间的相互依赖信息，**从而既不需要进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系

第八章 集成学习——多分类器系统

8.1 个体与集成

- **集成学习：**通过构建并结合多个学习器来完成学习任务
- 一般结构：先产生一组个体学习器，再用某种策略将它们结合起来
- （书上原话：集成学习的一般结构是先产生一组“个体学习器”，再用某种策略将他们结合起来）
- 同质集成：只包含同种类型的个体学习器，个体学习器称为基学习器，例如决策树集成中全是决策树，神经网络集成中全是神经网络（这样的集成是同质的）
- 异质集成：包含不同类型的个体学习器，个体学习器称为组件学习器。例如同时包含决策树和神经网络（这样的集成是异质的）
- 根据**个体学习器的生成方式**，目前的集成学习方法大致**可以分为两大类**

降低偏差，基于泛化性能曲线的“过拟合”和建构很强的集成

- 个体学习器间存在强依赖关系、必须**串行生成的序列化方法**，例如 Boosting（最著名的代表 AdaBoost）
- 个体学习器之间不存在强依赖关系、可**同时生成的并行化方法，**例如 Bagging 与随机森林
- +另外，与标准的 AdaBoost 只适用于二分类任务不同，Bagging 能不经修改的用于多分类、回归等任务。

结合策略：平均法、投票法、学习法

8.2 Boosting

8.3 Bagging 与随机森林

第九章 聚类

9.1 聚类任务

- 无监督学习：训练样本的标记信息是未知的，此类学习任务中**研究最多、应用最广**的是聚类（课本原话：在“无监督学习中”，训练样本的标记信息是未知的，目标是通过对无标记训练样本的学习来揭示数据的内在性质及规律，为进一步数据分析提供基础。此类学习任务中研究最多应用最广的是聚类）
- **聚类试图将数据集中的样本划分为若干个通常不相交的子集，每个子集被称为一个簇**
- 聚类过程，仅能自动形成簇结构，簇所对应的概念语义需由使用者来把握和命名

9.2 性能度量

- 物以类聚：同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同，也就是说，聚类结果的簇内相似度高且簇间相似度低（课本原话：直观上看，我们希望“物以类聚”，即同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。换言之，聚类结果的“簇内相似度”高且“簇间相似度”低）

9.3 距离计算

- 离散属性与连续属性的性质更接近一些，能直接在属性值上计算距离，这样的属性称为有序属性
- Page200VDM
- 以下均为课本原话
- 给定样本，最常用的是“闵可夫斯基距离”，可用于有序属性
- p=2 时，闵可夫斯基距离即“欧氏距离”
- p=1 时，闵可夫斯基距离即“曼哈顿距离”
- 对无序属性可采用 VDM，于是将闵可夫斯基距离和 VDM 结合即可处理混合属性
- 课本 P202　9.4.1 给定样本集.....一直到“其中.....是簇 Ci 的均值向量”结合公式 9.24

9.4 原型聚类（K 均值算法）——page203 计算看一遍，上面的过程用自己话可以复述出来

原型聚类有三种方法，k-means 算法，学习向向量量化算法（代表算法 LVQ），高斯混合聚类（高斯混合聚类采用概率模型来表达聚类原型）

第十章 降维

10.1~10.3：降维的原因与目的，低维嵌入的主要思想和作用，最大可分性与最近重构性

K 近邻学习 k-NN **是一种常用的监督学习方法，首先确定训练样本，以及某种距离度量，然后对于某个给定的测试样本，找到训练集中距离最近的 k 个样本，然后根据这 k 个邻居的信息进行预测。(书上原话：k 近邻学习是一种常用的监督学习方法，其工作机制非常简单：给定测试样本，基于某种距离度量，找出训练集中与其靠近的 k 个训练样本，然后基于这 k 个“邻居”的信息拉进行预测，通常在分类任务中可使用“投票法”，在回归任务中可使用“平均法”)**

通常在分类任务中使用投票法，回归类任务中采用平均法

缓解维数灾难的一个重要途径就是**降维**

原因：数据样本虽然是高纬的，但与学习任务密切相关的也许仅是某个低维分布，即高维空间中一个低维嵌入，因而可以对数据进行有效的降维

目的/作用：通过某种数学变换，将原始高纬属性空间转变为一个低维子空间，在这个子空间中样本密度大幅度提高，距离计算也变得更为容易

低维嵌入的主要思想：一般来说，欲获取低维子空间，最简单的是对原始高维空间进行线性变换

基于线性变换来进行降维的方法称为线性降维方法

最大可分性：样本点在这个超平面上的投影能尽可能分开

最近重构性：样本点到这个超平面的距离都足够近。

名词解释:

机器学习:

机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，从而在计算机上从数据中产生“ 模型”用于对新的情况给出判断

假设通过学习所获得的模型，对应了数据的某种潜在的规律，因此所获得的模型亦称"假设"

归纳从特殊到一般的"泛化"过程，即从具体的事实归结出一般性规律

演绎从一般到特殊的“ 特化”过程，即从基础原理推演出具体状况

过拟合

若学习器把训练样本学习的“ 太好”，将训练样本本身的特点当做所有样本的一般性质，以至于把训练样本所包含的不太一般的特性都学到了，导致泛化性能下降

欠拟合学习能力低下而造成的，对训练样本的一般性质尚未学好

错误率分错样本占样本总数的比例

精度分对样本占样本总数的比率

查准率(准确率):正例被预测出来的比率

查全率(召回率):预测出来的正例中正确的比率

偏差度量了学习算法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

表达了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。

误差样本真实输出与预测输出之间的差异。

泛化:

机器学习的目标是使学得的模型能很好地适用于“ 新样本”，该模型适用于新样本的能力，称为泛化。

神经网络

:是由具有适应性的简单单元组成的广泛并行互联的网络，它的组织能够模拟生物神经系统对真实世界物体所作出的反应。

局部极小解:是参数空间中的某个点,其邻域点的误差函数值均不小于该点的函数值。

全局最小解:是指参数空间中所有点的误差函数值均不小于该点的误差函数值。

线性模型:在样本空间中寻找一个超平面，将不同类别的样本分开

支持向量

:机解的稀疏性:支持向量机训练完成后，大部分的训练样本都不需保留，最终模型仅与支持向量有关

贝叶斯网:

“ 信念网”，它借助有向无环图来刻画属性间的依赖关系，并使用条件概率表来表述属性的联合概率分布。

贝叶斯判定准则

:为最小化总体风险，只需在每个样本上选择那个能使条件风险最小的类别标记，即

什么是过拟合及其一般的缓解方式?

过拟合，即若学习器把训练样本学习的“ 太好”，将训练样本本身的特点当做所有样本的一般性质，以至于把训练样本所包含的不太一般的特性都学到了，导致泛化性能下降。过拟合的缓解方式为:通过加正则项，优化目标函数

线性判别分析LDA的主要思想是什么?

给定训练样例集，设法将样例投影到二条直线上，使得同类样例的投影点尽可能接近;异类样例的投影点尽可能远离。

线性回归目的是什么?

给出单一属性的线性回归目标函数;若采用最小二乘法求解线性回归模型，给出最小二乘法的闭式解及其推导过程。

决策树学习的目的是什么?决策树学习的目的是为了产生一棵泛化能力

强，即处理未见示例能力强的决策树

决策树学习算法中剪枝的目的是什麼?剪枝的基本策略的分为哪两种?

决策树学习算法中剪枝是决策树学习算法对付“ 过拟合”的主要手段，可通过主动去掉一些分支来降低过拟合的风险。剪枝的基本策略的分为预剪枝、后剪枝。

支持向量机中，软间隔基本想法是什麼?

最大化间隔的同时，让不满足约束的样本应尽可能少。

支持向量回归的特点是什么?

支持向量回归允许模型输出和实际输出间存在2的偏差，落入中间2间隔带的样本不计算损失，从而使得模型获得稀疏性。

贝叶斯概率模型的训练过程的本质什么?统计学界的两个学派是什么?

贝叶斯概率模型的训练过程的本质是参数估计过程。统计学界的两个学派为频率主义学派和贝叶斯学派。

贝叶斯概率模型中的频率主义学派的主导思想是什么?

认为参数虽然未知，但却存在客观值，因此可通过优化似然函数等准则来确定参数值

贝叶斯概率模型中的贝叶斯学派的主导思想是什么?

认为参数是未观察到的随机变量、其本身也可由分布，因此可假定参数服从一个先验分布，然后基于观测到的数据计算参数的后验分布。

拉普拉斯修正的优点是什麼?

拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题，并且在训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值渐趋向于实际概率值。

简述半朴素贝叶斯分类器的基本想法。

半朴素贝叶斯分类器适当考虑一部分属性间的相互依赖信息，从而既不需进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系。

简述贝叶斯网的概念。

贝叶斯网，亦称“ 信念网”,它借助有向无环图来刻画属性间的依赖关系，并使用条件概率表来表述属性的联合概率分布。

简述朴素贝叶斯分类器原理，并给出具体公式

朴素贝叶斯分类器采用“ 属性条件独立性假设”，对已知类别，假设所有属性相互独立，换言之，假设每个属性独立地对分类结果发生影响。

基于属性条件独立性假设， 写为: $P(c \mid \mathbf{x})$

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

由于对所有类别来说 相同$P(x)$,贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

D 的信息熵Ent(D)的公式表达。 $$\mathrm{Ent}(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

例 D^1(根蒂=蜷缩)有8个，其中正例5个，反例3个

$Ent(D^1)= -（(5/8）*\log_2(5/8)+ (3/8）*\log_2(3/8))$

试使用贝叶斯决策对该学生x进行男女性别分类

男生： $P(w_2)$ 0.1 $P(w_1) = 0.9$

现有一待识别的学生，其观察值为x，从类条件概率密度分布曲线上查得

根据贝叶斯判定准则得 $p(x \mid w_1) = 0.2,\ p(x \mid w_2) = 0.4$

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

h1=p（w1）p（x|w1） θ.9*0.2=0.18

h2=p（w2）p（x|w2） θ.1*0.4=0.04

h1>h2 因此是女生