1 绪论: 根据已存在的实例进行学习的例子

在绪论讲义中我们将介绍基本概念以及在今后整个课程中会用到的符号。我们将会遇到一些根据已有实例进行学习的材料以及一些补充材料。

1.1 基本概念及定义

已知一个集合 X，一个容许答案集合 Y，以及存在一个目标函数 $y^*: X \to Y$，并且我们仅对 X 的有限子集 $\{x_1, x_2, \cdots, x_l\} \subset X$ 存在已知目标函数值 $y_i = y^*(x_i)$。“对象-答案”有序对 (x_i, y_i) 称为**先例**。有序对 $X^l = (x_i, y_i)_{i=1}^l$ 的全体称为**训练样本**。

根据先例训练的目的，是从样本 X^l 重建关系 y^*，因此我们可以构造接近目标函数 $y^*(x)$ 的**判定函数** $a: X \to Y$，它不仅仅适用于训练样本，同时也适用于整个对象集合 X。

判定函数 a 应该可以有效地被计算机实现；因此它也可以被称为**算法**。

1.1.1 对象和特征

对象 x 的**特征**——是对象的某种特点的测量结果。特征的形式化表征是一个映射 $f: X \to D_f$，此处的 D_f 是一个特征容许值的集合。特别地，对任意的算法 $a: X \to Y$ 都可以看做某种特征。

根据集合 D_f 的性质，我们可以将特征分为以下几种类型。

如果 $D_f = \{0, 1\}$，那么 f 称为二**元**特征；

如果 D_f——有限集合，那么 f 称为**标定**特征；

如果 D_f——有限有序集合，那么 f 称为**顺序**特征；

如果 $D_f = \mathbb{R}$，那么 f 称为**定量**特征。

如果所有的特征都有相同的类型，即 $D_{f_1} = D_{f_2} = \cdots = D_{f_n}$，那么原始数据称为**同类的**，否则称为**异类的**。

假定我们已有一组特征 $f_1, f_2, \cdots, f_n$。向量 $(f_1(x), f_2(x), \cdots, f_n(x))$ 称为对象 $x \in X$ 的**特征描述**。在今后，我们将不把 X 中的对象同它的特征描述区分开来，即认为 $X = D_{f_1} \times D_{f_2} \times \cdots \times D_{f_n}$。训练样本 X^l 的特征描述

的全体，我们将它写成 $l \times n$ 的表，称为**对象-特征矩阵**：

$$F = \|f_j(x_i)\|_{l\times n} = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \cdots & \cdots & \cdots \\ f_1 x_l & \cdots & f_1 x_l \end{pmatrix} \tag{1.1}$$

对象——特征矩阵是用于表征上述数据以及附加问题数据的最普遍的标准表示。

1.1.2 答案与问题类型

根据容许答案集合 Y 的性质，我们可以将它分为以下几个类型。

如果 $Y = \{1, \cdots, M\}$，那么这个问题可以被分成 M 个不相交的类。在这种情况下，对象集合 X 被分成不同的类 $K_y = \{x \in X : y^*(x) = y\}$，并且算法 $a(x)$ 应当给出问题 "x 应该属于那个类?" 的答案。某个包含的对象的类称为**模式**，而问题就称为**模式识别**。

如果 $Y = \{0, 1\}^M$，那么这个问题可以被分成 M 个相交类。这种情况下的最简单的情形可以归结为，解决 M 个带有两个不相交的类的不相关的问题的分类问题。

如果 $Y = \mathbb{R}$，那么这就是一个**回归估计**问题。

预测问题其实就是分类问题或回归估计的一部分，即当 $x \in X$——对象 x 的之前的状态描述，$y \in Y$——对对象未来状态的描述。

1.1.3 算法模型与学习方法

定义 1.1. 参数映射群族 $A = \{g(x, \theta)|\theta \in \Theta\}$ 称为**算法模型**；此处的映射 $g : X \times \Theta \to Y$——这是某个确定的函数，$\Theta$——是参数 θ 的容许值的集合，它被称为**参数空间**或者**搜索空间**。

例 1.1. 在具有 n 个数值特征 $f_j : X \to \mathbb{R}, j = 1, \cdots, n$ 的问题中，我们广泛地使用带有向量参数 $\theta = (\theta_1, \cdots, \theta_n) \in \Theta = \mathbb{R}^n$ 的**线性模型**：

$$g(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)——对回归估计问题，\ Y = \mathbb{R};$$

$$g(x,\theta)=\operatorname{sign}\sum_{j=1}^{n}\theta_j f_j(x)\text{——对分类问题，}Y=\{-1,+1\}\text{。}$$

特征不仅仅是原始维度，也可以是与之相关的函数。特别地，多维线性模型甚至也可用于一维特征。

例 1.2. 一种经典的在给定点 $(x_i,y_i)\in\mathbb{R}^2, i=1,\cdots,l$ 进行一元函数逼近的方法，包含在构建**多项式模型**中。如果我们给出 n 个特征 $f_j(x)=x^{j-1}$，那么如同例 *1.1* 中的函数 $g(x,\theta)$ 将会定义为 $n-1$ 阶的在给定特征 x 下的多项式。

基于训练样本 X^l 的最佳模型参数 θ 的搜寻过程称为**拟合**或者算法的**学习**过程。

定义 1.2. 学习方法 (*learning algorithm*) ——是映射 $\mu:(X\times Y)^l\to A$，它把任意的有限的选择 $X^l=(x_i,y_i)_{i=1}^l$ 与某个算法 $a\in A$ 联系起来；也就是说，基于选择 X^l 方法 μ 建立了算法 a。学习方法必须能够有效地用计算机实现。

因此，根据先例学习的问题可以清楚地分为两个阶段。

在**学习**阶段，方法 μ 基于选择集 X^l 构建出算法 $a=\mu(X^l)$。

在**应用**阶段，算法 a 对与新的对象 x 给出答案 $y=a(x)$。

学习阶段是最复杂的过程。通常情况下，它可以被归结为模型参数的搜寻，这样的参数能最优化给定的质量泛函。

1.1.4 质量泛函

定义 1.3. 损失函数 (*loss function*) ——这是一个用于描述算法 a 在作用于对象 x 时的错误值的非负函数 $\mathcal{L}(a,x)$。如果 $\mathcal{L}(a,x)=0$，那么答案 $a(x)$ 称为**正确的**。

定义 1.4. 算法 a 在选择集 X^l 下的**质量泛函**:

$$Q(a,X^l)=\frac{1}{l}\sum_{i=1}^{l}\mathcal{L}(a,x_i) \tag{1.2}$$

泛函 Q 也被叫做**平均损失**或者**经验风险**,因为它是根据**经验数据** $(x_i, y_i)_{i=1}^l$ 计算出来的。

取值只有 0 和 1 的损失函数叫做**二元的**。在这种情况下，$\mathcal{L}(a, x) = 1$ 表示算法 a 在对象 x 处允许错误，而泛函 Q 称为算法 a 在选择集 X^l 下的**错误率**。

在 $Y = \mathbb{R}$ 时，我们最常使用如下的损失函数:

$\mathcal{L}(a, x) = [a(x) \neq y^*(x)]$——错误指示器 [1]，它经常用于分类问题中；

$\mathcal{L}(a, x) = |a(x) - y^*(x)|$——与正确答案的偏差；泛函 Q 称为算法 a 在选择集 X^l 下的**平均错误**；

$\mathcal{L}(a, x) = (a(x) - y^*(x))^2$——平方损失函数；泛函 Q 称为算法 a 在选择集 X^l 下的**平方平均错误**；它通常用于回归问题。

被称为**最小化经验风险**（empirical risk minimization，ERM）的经典学习方法，也包含在此过程中；它要求在模型 A 中找到算法 a，使得满足质量泛函 Q 在给定的训练样本 X^l 中具有最小值:

$$\mu(X^l) = \arg\min_{a \in A} Q(a, X^l) \tag{1.3}$$

例 1.3. *在具有 n 个特征 $f_j : X \to \mathbb{R}, j = 1, \cdots, n$ 的回归估计 $(Y = \mathbb{R})$ 问题中，最小化风险的二次损失函数也可以看作是最小二乘法:*

$$\mu(X^l) = \arg\min_{\theta} \sum_{i=1}^{l} (g(x_i, \theta) - y_i)^2$$

1.1.5 学习问题的概率描述

在根据先例的学习问题中，集合 X 的元素——它们可能并不是真实的的对象，而仅仅是一些与之相关的简单的数据。这些数据或许并**不精确**，因为在特征 $f_j(x)$ 以及目标关系 $y^*(x)$ 的测量计算过程中常常存在误差。数据也可能是**不完整**的，因为并不能够测量一切可能的特征，而仅仅是测量那些实质上容易辨识的量。同一个 x 的描述结果也可能对应于不同的对象及答案。在那种情况下，$y^*(x)$ 严格地说来并不是函数。**问题的概率描述**能够消除这种不正确性。

为取代未知目标函数 $y^*(x)$ 的存在性，我们假设存在集合 $X \times Y$ 中

[1] 方括号通过规则 [真]=1，[假]=0，将逻辑值转换成数值

存在按概率密度 $p(x,y)$ 的未知概率分布，它随机地并且不依赖于 l 个观察结果 $X^l=(x_i,y_i)_{i=1}^l$ 的选择。这种选择叫做**简单的**或者**独立同一分布的**(independent identically distributed,i.i.d)。

问题的概率描述一般来说是更加广泛的存在，因为函数关系 $y^*(x)$ 也可以用概率分布 $p(x,y)=p(x)p(y|x)$ 来表示，只要令 $p(x|y)=\delta(y-y^*(x))$，此处的 $\delta(z)$——是狄拉克 δ 函数。

极大似然原则。 用问题的概率描述去代替用于近似未知关系 $y^*(x)$ 的算法 $g(x,\theta)$ 的模型，给出用于近似未知概率密度 $p(x,y)$ 的对象与答案的联合概率分布模型 $\varphi(x,y,\theta)$。然后寻找参数 θ，使得在这样的参数下选择的数据集 X^l 有最大的似然性，因此存在能最好逼近的概率密度模型。

如果选择 X^l 中的观察结果互相独立，那么所有观察结果的联合概率密度分布等于每个观察结果的概率密度 $p(x,y)$ 的积:$p(X^l)=p((x_1,y_1),\cdots,(x_l,y_l))=p(x_1,y_1)\cdots p(x_l,y_l)$。用密度模型 $\varphi(x,y,\theta)$ 代替 $p(x,y)$，我们就得到**似然函数** (likelihood):

$$L(\theta,X^l)=\prod_{i=1}^{l}\varphi(x_i,y_i,\theta)$$

似然值越大，选取与模型的拟合度越好。因此，我们需要找到，使得值 $L(\theta,X^l)$ 具有最大值的参数 θ。在数学概率论中，这叫做**极大似然原则**。它的形式化推理可以在参考文献 [13] 中找到。

此后，像寻找参数 θ，通过概率密度 $\varphi(x,y,\theta)$ 寻找构建算法 $a_\theta(x)$ 都没有困难。

最大似然与最小经验风险之间的关系。 可以用求 $-\ln L$ 的最小值来代替求 L 的最大值，因为它对于选择集的对象是可加的（具有和式):

$$-\ln L(\theta,X^l)=-\sum_{i=1}^{l}\ln\varphi(x_i,y_i,\theta)\to\min_\theta \tag{1.4}$$

如果我们定义**概率损失函数** $\mathcal{L}(a_\theta,x)=-l\ln\varphi(x,y,\theta)$，那么我们就会发现，这个泛函就和经验风险泛函是一致的。这样定义损失是十分自然的——如果序对 (x_i,y_i) 与模型 φ 的一致性越差，那么概率密度 $\varphi(x_i,y_i,\theta)$ 的

值就会越小，从而损失函数 $\mathcal{L}(a_\theta, x)$ 的值就会越大。

反之亦然——大量的损失函数也可以通过某种方式组合成概率密度模型 $\varphi(x, y, \theta)$，因此最小经验风险与极大似然实质上是等价的。

例 1.4. 如果给定模型 $g(x,\theta)$。采取额外的概率假设，误差 $\varepsilon(x,\theta) = g(x,\theta) - y^*(x)$ 符合正态分布 $\mathcal{N}(\varepsilon; 0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\varepsilon^2}{2\sigma^2})$，它的平均值为零，方差为 σ^2。因此概率密度模型具有形式：

$$\varphi(x, y, \theta) = p(x)\varphi(y|x, \theta) = p(x)\mathcal{N}(g(x,\theta) - y^*(x); 0, \sigma^2)$$

据此可以推出，概率损失函数是精确到两个与参数 θ 无关的常量的二次函数：

$$-\ln\varphi(x, y, \theta) = -\ln p(x)\mathcal{N}(g(x,\theta) - y^*(x); 0, \sigma^2) = C_0 + C_1(g(x,\theta) - y^*(x))^2$$

因此存在两种途径形式化学习问题：第一种基于引入损失函数，第二种——基于引入数据生成的概率模型。两种方式相似地（甚至是完全相同地）优化问题。学习——就是进行优化。

1.1.6 过拟合问题及泛化能力的概念

最小化经验风险应该谨慎地进行运用。如果某个算法 a 到达了泛函 $Q(a, X^l)$ 的最小值，但这并不能保证 a 对任意的**测验选择集** $X^k = (x'_i, y'_i)_{i=1}^l$ 都能很好地接近目标关系。

如果算法在训练集中的工作质量明显高于未包含在训练集中的对象时，称为**过训练**（overtraining）或者**过拟合**（overfitting）现象。在实际问题中，会经常遇到这种情况。

很容易这样设想这样的方法，它能将经验风险最小化到零，但是它绝对无法进行学习。得到训练集 X^l 之后，方法会记住他们并建立起一个算法把每一个要求的对象 x 与训练集 X^l 中的对象 x_i 进行比较。如果两个对象相同 $x = x_i$，那么算法会给出正确的答案。对于此外的情况算法可能给出任意的答案。经验风险的最小可能值为零。不过这样的算法不能重建除学习对象之外的关系。结论就是：想要成功地学习，不仅仅要记忆，还需要泛化。

方法 μ 的**泛化能力**（generalization ability）由值 $Q(\mu(X^l), X^k)$ 描述，此处的选择集 X^l 与 X^k 要求是具有代表性的。为形式化概念 “代表性的选择”，经常采用标准假设，即选择 X^l 与 X^k——简单的，取自同一个未知概率分布的集合 X。

定义 1.5. *学习方法 μ 称为*__可靠的__，*如果对于足够小的值 ε 和 η 满足不等式*

$$P_{X^l,X^k}\{Q(\mu(X^l), X^k) > \varepsilon\} < \eta \tag{1.5}$$

参数 ε 称为**准确性**，参数 $(1-\eta)$——可靠性。

允许这样的等价形式：对任意的简单选择 X^l 和 X^k，评价值 $Q(\mu(X^l), X^k) \leqslant \varepsilon$ 满足不小于概率 $(1-\eta)$。

获取具有形式 (1.5) 的评价值是学习的统计学理论的基本问题。第一个评价值是在 60 年代末由弗拉基米尔 · 纳乌莫维奇 · 万普尼克（Vladimir Naumovich Vanpnik）与阿列克谢 · 雅克夫列维奇 · 泽范阑杰斯（Alexey Yakovlevich Chervonenkis）得到 [5,6,7]。目前统计学理论正在蓬勃发展 [34]，不过对与很多感兴趣的实际问题，泛化能力的评价值要么无法得知，要么非常之高。

泛化能力的经验评价值 当不能从理论上进行计算时，我们会采用这种方法。

给定被选集 $X^L = (x_i, y_i)_{i=1}^L$。用 N 种方式将它分成两个不相交的子集——长为 l 的学习集 X_n^l 以及长为 $k = L - l$ 的验证集 X_n^k。对每种分解方式 $n = 1, \cdots, N$ 建立算法 $a_n = \mu(X_n^l)$ 并计算质量泛函 $Q_n = Q(a_n, X_n^k)$ 的值。所有分解方式的质量泛函的算术平均数称为**交叉验证**水平（cross-validation,CV）：

$$CV(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(Q_n^l), X_n^k) \tag{1.6}$$

对不同的 X^L 的分解方式可能导致不同交叉验证水平 [48]。最简单的情况就是将 N 值选定在 20 到 100 的范围内。“事实上” $t \times q$ **倍交叉验证**（$t \times q$-fold cross-validation）是标准方法，即选择样本被随机划分成长度为 q（或几乎为 q）的块，每个块依次成为验证样本，同时剩余的其他块就是学习样本。样本 X^L 以 t 次的不同方式分解为 q 个块。总共我们就能得到

$N = tq$ 种分解方式。由于所有的对象都能在验证中找到 t 次，所以这个方法给出了更加准确的水平估计。

交叉验证的缺点：计算上的非高效性；较高的方差；由于训练集的数量从 L 降到 l，因此导致训练不够充分。

1.2 实际问题实例

在人类事物的不同了领域中，我们时常能够遇到有关分类，回归以及预测的实际问题，并且它们的数量还在不断增长。

1.2.1 分类问题

例 1.5. 在**医学诊断**问题中，病人就相当于对象。特征就是疾病症状以及治疗手段的观察结果。比如二元特征——性别，是否头痛，是否虚弱，是否恶心等等。顺序特征——状态程度（非常好，一般，严重，非常严重）。定量特征——年龄，脉搏，血压，血液中的血红蛋白数，药物剂量等等。事实上，患者的特征描述就是病史的形式化表现。如果积累了足够多的病例，我们就可能能够解决一些不同的问题：对疾病进行分类（**鉴别诊断**）；制定最合理的治疗手段；预计病症的持续及结束时间；评估并发症风险；寻找综合征——疾病的最大病症的集合表现。这一系列的系统的价值在于，他们能够即时分析总结大量的案例——这是人类不能做到的。

例 1.6. 在银行发放贷款过程中的**债务评估**问题。在美国和其他发达国家 *60-70* 年代信用卡的繁荣时期，首次出现了对自动发放贷款程序的需要。在这种情况下，对象就是债务人——有资格承担债务的实际上的或是法律上的代表。在实际债务人的情况下，特征就由记录有债务人债务的债券组成，上面也可能记录由银行搜集的个人详细附加信息。二元特征的例子：性别，现存电话。标定特征——出生地，职业，所在单位。顺序特征——受教育状况，从事的职位。定量特征——年龄，工龄，家庭收入，在其他银行的欠债情况，贷款数量。学习集由已知的贷款历史组成。在最简单的情况之下，可以把债务人分为两类：“好”和“坏”。贷款应当只发放给第一类债务人。在更复杂一点的情况下，可以根据所得的全部特征信息来评估债务人的累积等级（*score*）。分数越高则表明债务人越可靠。我们将其称为——**信用评估**

(*credit scoring*)。在学习阶段，我们对特征信息进行综合和选择，并且对不同的特征分配不同的分数等级，以使得贷款风险最小化。以下问题是解决如何确定发放贷款的条件：确定贷款利率，还款期限以及合同中的其他条件。这些都是通过先例学习的问题。

例 1.7. 拥有大量实际客户的大中型公司需要进行**客户流失预警**（*churn predication*)。