《机器学习》读书笔记

黄奕诚

目录

1 绪论

1.1 引言

- 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。
- 机器学习研究的主要内容：在计算机上从数据中产生“模型”的算法，即“学习算法”。

1.2 基本术语

- 数据集（data set）：一组记录的集合
- 示例（instance）/样本（sample）：每条记录，即关于一个事件或对象的描述
- 属性（attribute）/特征（feature）：反映事件或对象在某方面的表现或性质的事项
- 属性值（attribute value）：属性上的取值
- 属性空间（attribute space）/样本空间（sample space）/输入空间：属性张成的空间，记为 $\mathcal{X}$
- 特征向量（feature vector）：一个示例（在样本空间对应的坐标向量）
- 学习（learning）/训练（training）：从数据中学得模型的过程
- 训练数据（training data）：训练过程中使用的数据
- 训练样本（training sample）：训练数据中的每个样本
- 训练集（training set）：训练样本组成的集合
- 假设（hypothesis）：对应了关于数据的某种潜在规律的学得模型
- 真实（ground-truth）：潜在规律自身
- 学习器（learner）：学习算法在给定数据和参数空间上的实例化
- 标记（label）：关于示例结果的信息

- 样例（example）：拥有标记信息的示例
- 标记空间（label space）/输出空间：所有标记的集合，记为 $\mathcal{Y}$
- 分类（classification）：预测的是离散值的学习任务（二分类 $\mathcal{Y} = \{-1, +1\}$或$\{0, 1\}$；三分类 $|\mathcal{Y}| > 2$）
- 回归（regression）：预测的是连续值的学习任务（$\mathcal{Y} = \mathbb{R}$）
- 测试（testing）：使用学得模型进行预测的过程
- 测试样本（testing sample）：被预测的样本
- 无监督学习（unsupervised learning）：训练数据中没有标记信息的学习任务，代表是聚类（clustering）
- 监督学习（supervised learning）：训练数据中具有标记信息的学习任务，代表是分类和回归
- 泛化（generalization）能力：学得模型适用于新样本的能力

1.3 假设空间

- “从样例中学习”是一个归纳的过程。
- 可以把学习过程看作一个在所有假设组成的空间中进行搜索的过程，搜索目标是找到与训练集“匹配”（fit）的假设。
- 假设空间可以表示为一课属性值中通配符逐渐被具体数值取代的树。
- 可以用许多策略对假设空间进行搜索，如自顶向下（从一般到特殊）、自底向上（从特殊到一般）。
- 可能有多个假设与训练集一致，即存在着一个与训练集一致的“假设集合”，称之为“版本空间”（version space）。

1.4 归纳偏好

- 多个与训练集一致的假设所对应的模型在面临新样本时，可能产生不同的输出。而对于一个具体的学习算法而言，必须要产生一个模型。此时学习算法本身的偏好会起到关键的作用。

- 归纳偏好（inductive bias）：机器学习算法在学习过程中对某种类型假设的偏好。
- 奥卡姆剃刀（Occam's razor）：若有多个假设与观察一致，则选最简单的那个。【常用的、自然科学研究中最基本的原则】
- 设 f 为希望学习的真实目标函数，则基于训练数据 X 的算法 $\mathfrak{L}_a$ 在训练集之外的所有样本上的误差与学习算法无关，即

$$\sum_f E_{ote}(\mathfrak{L}_a|X,f) = \sum_f E_{ote}(\mathfrak{L}_b|X,f)$$

"没有免费的午餐"定理（NFL 定理）：所有学习算法的期望性相同。

1.5 发展历程

1. 二十世纪五十年代到七十年代初："推理期"——赋予机器逻辑推理能力
2. 二十世纪七十年代中期开始："知识期"
   a. 机械学习（信息存储与检索）
   b. 示教学习（从指令中学习）
   c. 类比学习（通过观察和发现学习）
   d. 归纳学习（从样例中学习）
      - 符号主义学习（决策树、基于逻辑的学习）
      - 连接主义学习（神经网络）
      - 统计学习（支持向量机、核方法）
      - 深度学习

1.6 应用现状

- 计算机科学诸多分支学科领域（如计算机视觉、自然语言处理）
- 交叉学科（如生物信息学）
- 数据挖掘（机器学习领域和数据库领域是数据挖掘的两大支撑）
- 人类日常生活（天气预报、搜索引擎、自动驾驶、政治选举等）
- 促进人们理解"人类如何学习"

2 模型评估与选择

2.1 经验误差与过拟合

- 设在 m 个样本中有 a 个样本分类错误，则错误率（error rate）为 $E = a/m$，精度（accuracy）为 $1 - a/m$。
- 误差（error）：学习器的实际预测输出与样本的真实输出之间的差异。训练误差（training error）/经验误差（empirical error）：学习器在训练集上的误差。泛化误差（generalization error）：学习器在新样本上的误差。想要使泛化误差最小，而新样本未知，所以努力使经验误差最小化。
- 过拟合（overfitting）：学习器将训练样本自身的一些特点当作为所有潜在样本都会具有的一般性质。【关键障碍、无法彻底避免】欠拟合（underfitting）：学习器对训练样本的一般性质尚未学好。【较容易克服】若 “P≠NP”，过拟合就不可避免。

2.2 评估方法

为了对学习器对泛化误差进行评估，需要使用一个测试集（testing set）来测试学习器对新样本的判别能力，然后以测试集上的测试误差（testing error）作为泛化误差的近似。

若当前只有一个包含 m 个样例的数据集

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

，则对其进行适当的处理，从中产生训练集 S 和测试集 T。

2.2.1 留出法

直接将数据集 D 划分为两个互斥的集合，其中一个作为训练集 S，另一个作为测试集 T，即 $D = S \cup T, S \cap T = \emptyset$，在 S 上训练出模型后，用 T 来评估其测试误差，作为对泛化误差的估计。

- 划分尽可能保持数据分布的一致性，例如在分类任务中至少要保持样本的类别比例相似（分层采样）。

- 一般采用若干次随机划分、重复进行实验评估后取平均值作为评估结果。
- S 和 D 大小权衡没有完美的解决方案，常见做法是 2/3∼4/5 的训练样本比例。

2.2.2 交叉验证法

将数据集 D 划分为 k 个大小相似的互斥子集，即

$$D = D_1 \cup D_2 \cup \cdots \cup D_k, D_i \cup D_j = \emptyset (i \neq j)$$

每个子集 D_i 都尽可能保持数据分布的一致性（分层抽样）。然后从中选取 $k-1$ 个子集为训练集，剩下一个子集为测试集。可进行 k 次训练和测试，最终返回 k 个测试结果的均值。也称为“k 折交叉验证”（k-fold cross validation）。

- k 最常用的取值是 10，常用的还有 5、20 等。
- 留一法（Leave-One-Out）不受随机样本划分的影响，评估结果比较准确，但计算开销大。

2.2.3 自助法

以自助采样法（bootstrap sampling）为基础，给定包含 m 个样本的数据集 D，每次随机从 D 中挑选一个样本，将其拷贝放入 D'，再将该样本放回初始数据集 D 中。这个过程重复执行 m 次后，得到了包含 m 个样本的数据集 D'。此时将 D' 用作训练集，$D \backslash D'$ 用作测试集。

- D 有约 36.8% 的样本未出现在采样数据集 D' 中。
- 亦称为“包外估计”（out-of-bag estimate）。
- 自助法在数据集较小、难以有效划分训练/测试集时很有用，且能从初始数据集中产生多个不同的训练集。
- 因为自助法产生的数据集改变了初始数据集的分布，会引入估计偏差。

2.2.4 调参与最终模型

- 常用的调参做法：对每个参数选定一个范围和变化步长，进行计算开销和性能估计之间的折中。
- 在模型选择完成后，学习算法和参数配置已选定，此时用数据集 D 重新训练模型，使用所有 m 个样本，得到最终提交给用户的模型。

2.3 性能度量

给定样例集

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

其中 y_i 是示例 $\boldsymbol{x}_i$ 的真实标记。要评估学习器 f 的性能，即把学习器预测结果 $f(\boldsymbol{x})$ 与真实标记 y 进行比较。

回归任务中最常用的性能度量："均方误差"（mean squared error）

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

更一般地，对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$，均方误差可描述为

$$E(f; \mathcal{D}) = \int_{x\sim\mathcal{D}}(f(\boldsymbol{x}) - y)^2 p(\boldsymbol{x})d\boldsymbol{x}$$

对于分类任务——

2.3.1 错误率与精度

- 分类错误率

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

精度

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

- 对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$，错误率

$$E(f;\mathcal{D}) = \int_{x\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) \neq y)p(\boldsymbol{x})d\boldsymbol{x}$$

精度

$$\text{acc}(f;\mathcal{D}) = \int_{x\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) = y)p(\boldsymbol{x})d\boldsymbol{x} = 1 - E(f;\mathcal{D})$$

2.3.2 查准率、查全率与 $F1$

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查准率（precision）

$$P = \frac{TP}{TP + FP}$$

查全率（recall）

$$R = \frac{TP}{TP + FN}$$

- 平衡点（Break-Even Point，BEP）：$R = P$ 时的取值，数值越高可以认为学习器越优。

- $F1$ 度量

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

实际上 $F1$ 是 R 和 P 的调和平均

$$\frac{1}{F1} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$$

- F_β 度量：考虑 R 与 P 的不同偏好，设 β 为查全率 R 对查准率 P 的相对重要性，则

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

实际上 F_β 是加权调和平均

$$\frac{1}{F_\beta} = \frac{1}{1 + \beta^2}(\frac{1}{P} + \frac{\beta^2}{R})$$

- 宏 $F1$：在各混淆矩阵上分别计算出各自的 (P_i, R_i)，再计算平均值：

$$\text{macro}-P = \frac{1}{n}\sum_{i=1}^{n} P_i$$

$$\text{macro}-R = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$\text{macro}-F1 = \frac{2 \times \text{macro}-P \times \text{macro}-R}{\text{macro}-P + \text{macro}-R}$$

- 微 $F1$：先将各混淆矩阵的对应元素进行平均得到四个指标，再基于这些平均值计算 $F1$：

$$\text{micro}-P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$

$$\text{micro}-R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}}$$

$$\text{micro}-F1 = \frac{2 \times \text{micro}-P \times \text{micro}-R}{\text{micro}-P + \text{micro}-R}$$

△ 混淆矩阵介绍：每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；每一行代表了数据的真实归属类别，每一行的数据总数表示该类别的数据实例的数目。例如共有 150 个样本数据，预测为 1、2、3 类各 50 个，分类结束后得到的混淆矩阵为

		预测		
		类 1	类 2	类 3
实际	类 1	43	2	0
	类 2	5	45	1
	类 3	2	3	49

2.3.3 ROC 与 AUC

ROC 全称是“受试者工作特征”（Receiver Operating Characteristic）曲线。横轴为“假正例率”（FPR），纵轴为“真正例率”（TPR）。

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

- 现实任务中 ROC 曲线的绘制方法：给定 m^+ 个正例和 m^- 个反例，根据学习器预测结果对样例进行排序，然后把分类阈值设为最大，此时 FPR 和 TPR 都为 0. 在坐标 $(0, 0)$ 处标记一个点，然后将分类阈值依次设为每个样例的预测值。设当前一个标记点坐标为 (x, y)，若当前为真正例，则对应标记点坐标为 $(x, y + \frac{1}{m^+})$；若当前为假正例，则对应标记点坐标为 $(x + \frac{1}{m^-}, y)$，然后用线段连接相邻点即得。

- AUC (Area Under ROC Curve) 即为 ROC 曲线下各部分的面积之和。设 ROC 曲线是由坐标为 $\{(x_i, y_i)|1 \leq i \leq m\}$ 的点按序连接而成，则 AUC 可估算为

$$\text{AUC} = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$

- 给定 m^+ 个正例和 m^- 个反例，令 D^+ 和 D^- 分别表示正、反例集合，则排序“损失”(loss) 定义为

$$\ell_{rank} = \frac{1}{m^+ m^-} \sum_{\boldsymbol{x}^+ \in D^+} \sum_{\boldsymbol{x}^- \in D^-} \left(\mathbb{I}\big(f(\boldsymbol{x}^+) < f(\boldsymbol{x}^-)\big) + \frac{1}{2} \mathbb{I}\big(f(\boldsymbol{x}^+) = f(\boldsymbol{x}^-)\big) \right)$$

它对应 ROC 曲线之上的面积，有

$$\text{AUC} = 1 - \ell_{rank}$$

2.3.4 代价敏感错误率与代价曲线

- 不同类型的错误可能造成不同损失，所以为错误赋予“非均等代价”(unequal cost)。

- 以二分类为例，可以设定一个“代价矩阵”，如下表所示。

真实类别	预测类别	
	第 0 类	第 1 类
第 0 类	0	$cost_{01}$
第 1 类	$cost_{10}$	0

- “代价敏感”（cost-sensitive）错误率

$$E(f;D;cost)=\frac{1}{m}\Bigg(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}\big(f(\boldsymbol{x}_i)\neq y_i\big)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}\big(f(\boldsymbol{x}_i)\neq y_i\big)\times cost_{10}\Bigg)$$

- 在非均等代价下，“代价曲线”（cost curve）可以刻画期望总体代价。设 p 是样例为正例的概率。横轴为正例概率代价

$$P(+)cost=\frac{p\times cost_{01}}{p\times cost_{01}+(1-p)\times cost_{10}}$$

  纵轴为取值为 $[0,1]$ 的归一化代价

$$cost_{norm}=\frac{\mathrm{FNR}\times p\times cost_{01}+\mathrm{FPR}\times(1-p)\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}$$

- 代价曲线的绘制方法：将 ROC 曲线上的每一点转化为代价平面上的一条线段，取所有线段的下界，围成的面积即为所有条件下学习器的期望总体代价。