第一章 绪论

机器学习定义

吃瓜群众希望根据经验从无数不同形态特点的瓜中选择出一个好瓜，于是把 **经验** 作为 **数据** 输入计算机，学习经验数据生成 **算法模型**，并作出判断。

重要会议

- 国际机器学习会议 **ICML**
- 国际神经信息处理系统会议 **NIPS**
- 国际学习理论会议 **COLT**
- 欧洲机器学习会议 **ECML**
- 亚洲机器学习会议 **ACML**
- 人工智能领域会议 **IJCAI**， **AAAI**
- 数据挖掘领域会议 **KDD**， **ICDM**
- 机器视觉模式识别领域 **CVPR**
- 中国机器学习大会 **CCML**
- 机器学习及应用研讨会 **MLA**

重要期刊

- 国际学术期刊 Journal of Machine Learning Research , Machine Learning
- 人工智能领域 Artifical Intelligence , Journal of Artificial Intelligence Research
- 数据挖掘领域 ACM Transaction on Knowledge Discovery from Data , Data Mining and Knowledge Discovery
- 计算机视觉与模式识别 IEEE Transaction on Pattern Analysis and Machine Intelligence
- 神经网络 Neural Computation, IEEE Transactions on Neural Networks and Learning Systems
- 统计学 Annals of Statustion

基本术语

- `sample`

  样本，也叫示例 `instance`

  如描述一个瓜的特性的一条记录：（色泽=青绿；根蒂=蜷缩；敲声=浊响）

- `attribute`

  属性，或 `feature`，如色泽、根蒂、敲声

- `attribute value`

  属性值，不多解释

- `data set`

  数据集，由样本组成

  $D = \{x_1, x_2, \ldots, x_m\}$

- `attribute space`

  属性空间，或样本空间 `sample space`，作为输入空间。比如西瓜的三个属性分别作为空间上的 x, y, z 轴形成的一个三维空间

- `feature vctor`

  特征向量，在样本空间中每个样本对应一个

  $x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$, xi 的维度 `dimensionality` 为 `d`

- 机器学习基本过程为学习或训练。使用由训练数据 `training data` 中的训练样本 `trainning sample` 组成的训练集 `training set` 进行训练。

- `learner`

  学习器，学习训练得到的模型

- 学习训练得到的模型对应某种潜在的规律 —假设`hypohesis`或真实、真相 `ground-tuth`。

- `lable`

  标记，关于训练结果好坏的信息。比如（（色泽=青绿；根蒂=蜷缩；敲声=浊响）好瓜）

- `example`

  拥有标记的样本 $(x_i, y_i) y_i \in Y$， Y 为标记空间/输出空间 `lable space`

- `classification`

  分类。预测 `prediction` 的是 **离散值**，如好瓜、坏瓜。

  可以分为二分类 `binary classification` 和多分类 `multi-class classification`，二分类中有正类 `positive class` 和反类 `negative class`。

- `regression`

  回归。预测的是 **连续值**，如瓜的成熟度0.95，0.37.

- `clustering`

  聚类。把训练集中的西瓜分为若干组/簇 `cluster`.

- 监督学习 `supervised learning`

  包括 **分类** 和 **回归**，有 **标记样本**

- 无监督学习 `unsupervised learning`

  如 **聚类**， 没有标记样本信息

- `generalization`

  泛化能力，在无监督学习中模型适用于新样本的能力

- 假设空间

  学习就是一个在假设空间中搜索，找到与训练集匹配 `fit` 的假设的过程

- 版本空间

  可能有多个**假设与训练集一致**，这些假设集合为版本空间

- 与训练集一致的假设

  表示能对训练集中所有的样本进行正确的判断

- 归纳偏好

  对于多个版本空间，需要选择一个更加合适的，最常见的是选择图线更加平滑的，算法有奥卡姆剃刀 `Occam's razor` 等

对于一个预测任务，对训练集$\{(x_1,y_1),(x_2,y_2),\dots,(x_i,y_i)\}$进行学习，建立输入空间 **X** 到输出空间 **Y** 的映射 $f:X\mapsto Y$,得到模型之后对被预测样本 `testing sample` **x** 进行测试得到其预测标记 **y = f(x)**。

没有免费午餐定理NFL

考虑到两种策学习算法 $\mathcal{E}_a,\mathcal{E}_b$，其中前者较为优越，后者为"随机胡猜"。

- 样本空间 χ 和假设空间 $\mathcal{H}$ 都是离散的；
- $P(h|X,\mathcal{E}_a)$: 算法 **a** 给予训练数据 **X** 产生假设 **h** 的概率；
- f : 希望学习的真实目标函数；
- $\mathbb{I}(\cdot)$: 只是函数，真为1假为0；

则 $\mathcal{E}_a$ 的 **训练集外误差** ——在训练机之外所有样本上的误差为

$$E_{ote}(\mathfrak{L}a|X,f)=\sum_h\sum_{x\in\mathcal{X}-X}P(x)\mathbb{I}(h(x)\neq f(x))P(h|X,\mathfrak{L}a)$$

对于二分问题，f 按照均匀分布对误差求和：

$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}a|X,f)&=\sum_f\sum_h\sum_{x\in\mathcal{X}-X}P(x)\mathbb{I}(h(x)\neq f(x))P(h|X,\mathfrak{L}a)\\&=\sum x\in\mathcal{X}-XP(x)\sum_h P(h|X,\mathfrak{L}a)\sum_f\mathbb{I}(h(x)\neq f(x))\\&=\sum x\in\mathcal{X}-XP(x)\sum_h P(h|X,\mathfrak{L}a)\frac{1}{2}2^{|\mathcal{X}|}\\&=\frac{1}{2}2^{|\mathcal{X}|}\sum x\in\mathcal{X}-XP(x)\sum_h P(h|X,\mathfrak{L}a)\\&=2^{|\mathcal{X}|-1}\sum x\in\mathcal{X}-XP(x)\cdot 1\end{aligned}$$

所以 **总误差与算法无关！**，它们的期望相同，但是是在 **均匀分布** 的前提下。所以 **NFL** 告诉我们 **学习算法自身与归纳偏好相匹配起到决定性作用**。

习题

1.1

若只包（青绿，蜷缩，浊响，是）、（乌黑、稍蜷，沉闷，否）的两个样例，试给出相应的版本空间

首先 3 * 3 * 3 + 1 = 28 个，然后枚举一下，把假设空间中能对所有的样本进行正确的判断的保留，剩下

色泽=青绿 根蒂=蜷缩 敲声=浊响

色泽=青绿 根蒂=蜷缩 敲声=*

色泽=青绿 根蒂=* 敲声=浊响

色泽=* 根蒂=蜷缩 敲声=浊响

色泽=青绿 根蒂=* 敲声=*

色泽=* 根蒂=蜷缩 敲声=*

1.3

若数据包含噪声，则假设空间中可能不存在与所有训练样本都一致的假设。在此情形下，试设计一种归纳偏好用于假设选择

去掉一些有欧相同属性却不同分类的数据

1.4

本章1.4节在论述“没有免费的午餐”定理时，默认使用了“分类错误率”作为性能度量来对分类器进行评估。若换用其他性能度量l,试证明没有免费的午餐”定理仍成立

证明：NFL首先要保证目标函数均匀分布...

第二章 模型评估与选择

误差与过拟合

- `error rate` 错误率 = 分类错误的样本数占总样本数的比例：$E = a/m$
- `accuracy` 精度 = 1 - 错误率：$1 - a/m$
- 误差：

误差名	使用范围
误差	实际输出与样本真实输出之间
训练误差/经验误差	在训练集上的误差
泛化误差	在新样本上的误差

- 过拟合 `overfitting`
- 欠拟合 `underfitting`

  比如对一张带有锯齿的绿色树叶进行学习，过拟合的结果是认为所有树叶都必须带有锯齿，欠拟合的结果是绿的的都为树叶。

评估方法

对机器学习的泛化误差进行评估，需要一个 **测试集** ，用测试集上的 **测试误差** 多位泛化误差的近似。**测试样本尽可能不出现在训练集中**

留出法 hold-out

将数据集 D 划分为两个互斥集合训练集 S 和测试集 T, $D = S \cup T, S \cap T = \varnothing$

- 常规划分比例：2/3 ～ 4/5

- 两个集合尽可能保持数据分布一致
- 由于划分的随机性，单次的留出法结果往往不够稳定，需多次随机划分，重复实验取平均值

交叉验证 cross validation

将数据集划分为 k 个大小相似的互斥子集 $D = D_1 \cup D_2 \cup \ldots \cup D_k, D_i \cap D_j = \varnothing (i \neq j)$

- 子集数据分布保持一致
- 进行k次训练和测试，自身作为测试集，k-1 个座位训练集
- 得到 k 次结果取平均
- 需多次随机划分，重复 p 次实验取平均值，叫 **p次k折交叉验证**
- 当 **k = 样本数** 时为 **留一法LOO**，Leave-One-Out

自助法 bootstrapping

以上两种方法评估模型使用的训练集比 D 小，于是训练样本不同会导致估计偏差，自助法为了减小这一影响

- D 有 m 个样本，从中抽取一个拷贝到 D' 并放回 D (下次采样还有可能被采样到)，在 m 次采样中样本始终不被采样的概率极限

$$\lim_{m \mapsto \infty} (1 - \frac{1}{m})^m \mapsto \frac{1}{e} \approx 0.368$$

- 所以 D 中大约有36.8%的样本没有出现在 D' 中，将 D' 作为训练集，$D - D'$ 作为测试集
- **包外估计** 有 3/1 的没有在训练集中出现的样本用于测试

自助法在数据集较小，难以有效划分训练集/测试集时很有用，数据集足够多时前两者比较好。

调参

`parameter tuning`，学习算法中有许多参数 `parameter` 需要设定。

性能评估

有了评估方法，我们还需要一个衡量模型泛化能力的评估标准 `performance measure`。

均方误差 MSE

回归中常用的性能度量方法

- 样本集 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
- 预测结果 $f(x)$ 与真实标记 y

均方误差

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

错误率与精度

错误率定义

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

精度

$$\begin{aligned}acc(f;D)&=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i)\\&=1-E(f;D)\end{aligned}$$

查准率、查全率、F1

查准率 `precision`: 检索出的信息中有多少比例用户感兴趣；

查全率 `recall`: 用户感兴趣的信息中有多少被检索出来；

对于二分类问题：**样例总数 = TP + FP + TN + FN**

	实际值	预测值	全称
TP	Positive	Positive	True Positive
FP	Negative	Positive	False Positive
FN	Positive	Negative	False Negative
TN	Negative	Negative	True Negative

分类结果 **混淆矩阵** 如下表所示

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）

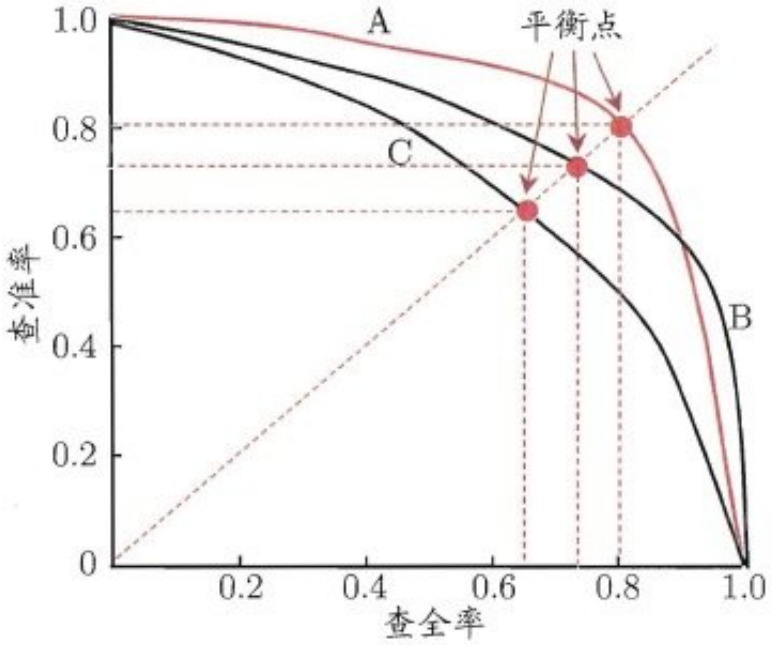
查准率

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

查全率

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

可以看出两者是矛盾量，一般来说P高往往R低。两者关系用 **P-R曲线** 描述，平衡点 **BEP**



F1度量

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

对于查准率与查全率的重视程度不同，可以设定偏好 β，$\beta=1$ 为标准F1；$\beta>1$ 查全率影响更大；$0<\beta<1$ 查准率影响更大。

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

多二维混淆矩阵

对于 n 个二分类混淆矩阵综合考察查全率和查准率的方法：

1. 计算个混淆矩阵的 P,R，再求平均，得到宏查准率、宏查全率、宏F1：

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

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

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

2. 对混淆矩阵各元素平均再基于这些计算出微查准率、微查全率、微F1：

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

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

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