# 机器学习一：基本概念

作者：damimi
原文链接：https://ld246.com/article/1589900749050
来源网站：链滴

背景：如何购买芒果?

1.列出每个芒果的特征（feature）；包括颜色、大小、形状、产地、品牌等。

2. 我们要预测的标签（label）；可以是连续值，（如芒果的甜度、水分、成熟度的综合打分。）也以是离散值（如好、坏等。）；标签的获取可以通过直接品尝获取，也可通过经验丰富的专家进行标。

3. 标记好的特征及标签的芒果可以看作一个样本（sample），也经常称为示例（Instance）。

4. 一组样本构成的集合称为数据集（data set）。一般数据集分为训练集和测试集，训练集中的样本来训练模型，测试集中的样本用来检验模型的好坏。

5. 通常一个D维向量 $\boldsymbol{x} = [x_1, x_2, \cdots, x_D]^{\mathrm{T}}$ 表示一个芒果的所有特征构成的向量，称为特征向量（feature Vector）,其中每维表示一个特征。芒果的标签通常用y表示。

6. 假设训练集D由N个样本组成，其中每个样本都是独立同分布，即独立地从相同的数据分布中抽取，记为：

$$\mathcal{D} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \cdots, (\boldsymbol{x}^{(N)}, y^{(N)})\}.$$

给定训练集D，我们让计算机从一个函数集合 $\mathcal{F} = \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots\}$ 中自动寻找一个“最优”的函数f*(x)来近似每个样本的征向量x和标签y之间的真实映射关系。对于一个样本x,我们可以通过函数f*(x)来预测其标签值：$\hat{y} = f^*(\boldsymbol{x}),$

或标签的条件概率：$\hat{p}(y|\boldsymbol{x}) = f_y^*(\boldsymbol{x}).$

这样，下次买芒果时，可以根据芒果的特征，使用学习到的函数f*(x)来预测芒果的好坏。为了评价的正性，我们独立同分布的抽取一组芒果作为测试集D′，并在测试集中所有芒果上进行测试，计算预测果的准确率

$$Acc(f^*(\boldsymbol{x})) = \frac{1}{|\mathcal{D}'|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}'} I(f^*(\boldsymbol{x}) = y),$$

其中I(·)为指示函数，|D′|为测试集大小。

机器学习的基本流程如下图：