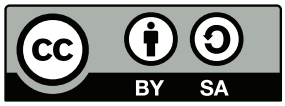
机器学习理论学习笔记(1):
基础知识

王龙腾
longtenwang@gmail.com

2017 年 3 月 9 日



在这一篇笔记中，我们将回顾机器学习的一些基本术语和它们的定义，以帮助读者建立起关于机器学习的理论架构。

1 何为机器学习

学习(learning)：假设用P来评估计算机程序在某任务类T上的性能，若一个程序通过利用经验E在T中任务上获得了性能改善，则我们说关于T和P，该程序对E进行了学习 [2]。周志华的书中这样说 [3]：

> 在计算机系统中，“经验”通常以“数据”的形式存在,因此机器学习所研究的主要内容，是关于在计算机上从数据中产生模型的算法，即所谓学习算法(learning algorithm)。有了学习算法，我们把经验数据提供给它，它就能给予这些数据产生模型；在面对新情况时，模型会给我们提供相应的判断。

如无特别指出，模型泛指从数据中学得的结果。我们有时也称从模型中学得数据的过程为训练(training)。此外，关于该定义的进一步说明请看 [1]

2 数据的抽象化定义：样本空间与标记空间

样本(sample)：亦称为示例(instance)。数据集里一条对事件或对象的描述。

特征(feature)：亦称为属性(attribute)。反映事件或对象在某方面的表现或性质的事项。

把上述文字抽象，假设对于某一学习任务T，对于样本$\boldsymbol{x}$，我们可以用d个特征对其进行描述，即$\boldsymbol{x}=\{x_1,x_2,\ldots,x_d\}$,特征$x_i$的取值集合为$X_i$，我们发现，这些特征“张成”了一个空间，我们称之为样本空间，记为$\mathscr{X}$。样本$\boldsymbol{x}$为这个空间中的一个向量，整数d称为样本空间的维度。

标记(label)：也译为标签。指关于示例结果的信息。有了标记的示例称为样例(example)。我们用 $(\boldsymbol{x}_i,y_i)$来表示第i个样例,其中$y_i\in\mathscr{Y}$是示例$\boldsymbol{x}_i$的标记，$\mathscr{Y}$是所有标记的集合，称为标记空间或者输出空间。

注意到实际的学习任务T中，样本空间的维度可能会很大，而且特征的取值集集合可以取到一些无穷集。所以，我们用样本空间中的三种采样来实现我们的学习任务，它们分别是训练集，校正集，测试集。

训练集(training set)：用来训练学习算法的示例的集合。

校正集(validation set)：用来调整学习算法的参数的样例的集合。

测试集(testing set)：用来评估学习算法的样例的集合。

3 学习算法的结果与性能评估

概念类(concept class)：从样本空间$\mathscr{X}$到标记空间$\mathscr{Y}$的映射称为概念，记为c。

若c满足$\forall(\boldsymbol{x},y),c(\boldsymbol{x})=y$则我们称其称为目标概念，我们希望从数据中学得的目标概念构成的集合，我们称之为概念类，记为$\mathscr{C}$。

假设空间(hypothesis space)：记为$\mathscr{H}$，有两个含义：

1）**对于具体的学习任务T，我们考虑所有可能概念的集合**。举周志华老师讲过的西瓜的例子，学习任务T是从一堆瓜中分辨出好瓜，我们可以观察到瓜的色泽，根蒂，敲声，于是我们可以从这一堆瓜的学习中得到，好瓜是具备某种色泽，某种根蒂，某种敲声的瓜这么一个概念。如果瓜的色泽，根蒂，敲声的分别有3，21种可能的取值。考虑到有可能一些特征与我们形成的概念无关，或者根本就没有所谓好瓜这个概念，所以所有可能概念的数目应该为$4\times3\times3+1=37$这37条概念就构成了关于判别好瓜这个分类的学习任务的假设空间。注意到，由于关于好瓜的概念以及我们通过学习算法得到的关于好瓜的概念可能不止一条，所以概念类$\mathscr{C}$的元素个数$|\mathscr{C}|$会大于1。此外，也许目标概念就不存在，所以会有$|\mathscr{C}|=0$

2)**对于学习算法$\mathcal{L}$，我们考虑的所有可能概念的集合**。举一个例子，在二维直角坐标系中有若干个点，每个点用蓝色或者白色标记，如果我们设计的学习算法是找到一条直线将蓝点和白点分开，那么我们的假设空间就是二维实平面中所有线性划分的集合。。

假设空间中的每一个元素称为假设，记为h。**如无特别说明，在理论探讨中，我们讨论的是关于学习算法的假设空间。**

损失函数(loss funciton)：用来衡量用学习算法预测得到的标记与真实标记之间的差别的函数。对于样例$(\boldsymbol{x}, y)$若学习算法得到的假设为h，则预测得到的标记为$h(\boldsymbol{x})$,真实标记为y,则损失函数可以记为$\ell(h(\boldsymbol{x}), y)$。

常用的损失函数有：

$$\ell_{\mathbb{I}}(h(y), y) = \begin{cases} 1, h(y) \neq y \\ 0, h(y) = y \end{cases} \tag{1}$$

若记$z = h(y) - y$，则我们有在支持向量机中常用的损失函数

$$\ell_{0/1}(z) = \begin{cases} 1, z < 0 \\ 0, otherwise \end{cases} \tag{2}$$

$$\ell_{hinge}(z) = \max(0, 1 - z) \tag{3}$$

$$\ell_{exp}(z) = \exp(-z) \tag{4}$$

$$\ell_{log}(z) = \log(1 + \exp(-z)) \tag{5}$$

经验风险(empirical risk)：也称为经验误差(empirical error)。用来衡量假设在样本空间的采样上与真实标记差别的一个量。对于$h \in \mathscr{H}$,如果采样中共有n个样例，则该假设的经验风险定义为：

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\boldsymbol{x}_i), y) \tag{6}$$

显然，在学习算法$\mathcal{L}$的设计中，我们的目标之一是设计出能找到那条经验风险最小的假设的算法。

泛化风险(generaliztion risk)：也称为泛化误差(generalization error)。用来衡量假设在样本空间上与真实标记的差别的一个量对于$h \in \mathscr{H}$，如果我们假设，样本空间$\mathscr{X}$中的所有样本服从一个隐含未知的分布$\mathcal{D}$，则h的泛化风险可以定义为：

$$R(h) = \mathbb{E}\left[\ell(h(\boldsymbol{x}), y)\right] \tag{7}$$

在实际问题中，我们很难将风险控制为0，退而求其次，令ϵ为$R(h)$的上限，来表示预先设定的学得模型所应满足的误差要求，我们称其为**误差参数**。

在机器学习算法的设计中，我们最终希望设计出一个能找到使得泛化风险满足误差参数的假设的学习算法。对于假设h,在实际问题中，我们通常不知道样本空间的未知分布$\mathcal{D}$，我们策略是用经验风险$\hat{R}(h)$来估计泛化风险$R(h)$。用经验风险估计的泛化风险理论探讨请看下一篇笔记。

参考文献

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[2] Tom M Mitchell et al. *Machine learning*. McGraw-Hill Boston, MA:, 1997.

[3] 周志华. 机器学习. 清华大学出版社, 2016.