Understanding Machine Learning

Yuyang Zhang

2017年7月15日

目录

1 why can machine learn

1.1 基本符号

1.1.1 输入

Domain Set: 一个任意集合$\mathcal{X}$，也作领域集。可以理解为所有样本的集合，其中每个样本通常以一个能够表征其特征的向量表示。

Label Set: 标签集$\mathcal{Y}$。样本所属于的类别，通常二分类问题，标签集为$\{0,1\}$或者是$\{-1,+1\}$。

Training Data: 训练数据$\mathcal{S}=\{(x_1,y_1),...,(x_m,y_m)\}$，也叫训练集，样本集。

1.1.2 输出

Predicting Rule: 预测规则，$h:\mathcal{X}\to\mathcal{Y}$。该规则是一个由样本集合到标签集的映射，可以理解为预测器(predictor)，**假设(hypothesis)**，分类器(classifier)，等等。

1.1.3 数据生成模型

我们假定样本$\mathcal{S}$是由概率分布$\mathcal{D}$生成，并且根据一个标记函数$f:\mathcal{X}\to\mathcal{Y}$，来标记样本的类别，对于任意$i=1,...,m$都有 $y_i=f(x_i)$，然而我们并不知道概率分布$\mathcal{D}$与标记函数f，我们的目标就是找一个合适的假设h，令其与标记函数f可以对样本做出相同的标记。

1.1.4 衡量标准

我们定义分类误差为：未能成功预测随机数据点正确标签的概率，即对于随机的一个$x\in\mathcal{X}$， $h(x)\neq f(x)$的概率。

定义$\mathcal{A}\subseteq\mathcal{X}$为一个领域子集， $\mathcal{A}$中的任意实例$x\in\mathcal{A}$的出现概率由$\mathcal{D}(\mathcal{A})$所决定。通常，我们称$\mathcal{A}$为一个事件， $\mathcal{A}=\{x\in\mathcal{X}:\pi(x)=1\}$，其中$\pi:\mathcal{X}\to\{0,1\}$，表示样本是否被观测到。我们也将$\mathcal{D}(\mathcal{A})$写作 $\mathcal{P}_{x\sim\mathcal{D}}[\pi(x)]$。此时我们可以定义假设$h$的错误率为：

$$\mathcal{L}_{\mathcal{D},f}(h)\overset{def}{=\!=}\mathcal{P}_{x\sim\mathcal{D}}[h(x)\neq f(x)]\overset{def}{=\!=}\mathcal{D}(\{x|h(x)\neq f(x)\})$$

其中误差的测量是基于概率分布$\mathcal{D}$和标记函数f的， $\mathcal{L}_{\mathcal{D},f}(h)$也称为泛化误差， **损失** 或者$h$的真实误差。

1.2 经验风险最小化

机器学习的过程都是基于训练集$\mathcal{S}$的，训练集$\mathcal{S}$由未知分布$\mathcal{D}$从领域集$\mathcal{X}$中采样得出，并由标记函数f标记，机器学习的输出是一个基于训练集$\mathcal{S}$的假设， $h_{\mathcal{S}} : \mathcal{X} \to \mathcal{Y}$。

由于我们并不知道分布$\mathcal{D}$与标记函数f，所以我们只能根据训练集$\mathcal{S}$来判断我们所选择假设的表现，定义训练误差为：

$$\mathcal{L}_{\mathcal{S}} \xlongequal{def} \frac{|\{x_i|h(x_i) \neq y_i, i = 1, .., m\}|}{m}$$

训练误差也称作经验误差和经验风险。

因为训练集$\mathcal{S}$是领域$\mathcal{X}$的一个子集，所以训练样本是真实世界的一个缩影，正如概率统计的一个核心思想，通过样本反映总体。所以我们认为利用样本集寻找一个较好的假设是可行的，即最小化训练误差$\mathcal{L}_{\mathcal{S}}(h)$，这称之为经验风险最小化， ERM(Experience Risk Minimize)。

1.2.1 过拟合

一个假设在训练集上效果优异，但在真实世界中表现却糟糕，这种现象称之为过拟合。当我们过度追求经验风险最小化原则时，我们就有可能面临过拟合的风险。

1.3 归纳偏好

虽然经验风险最小化会有过拟合的风险，相比于抛弃这个原则，我们更愿意去修正这个原则，考虑我们对于假设的归纳偏好。我们通常的解决方案是根据我们定好的归纳偏好，在有限的假设空间中去搜索所要用的假设，这些假设的集合成为假设类，记为$\mathcal{H}$，则我们的学习过程可以记为：

$$ERM_{\mathcal{H}}(\mathcal{S}) \in \underset{h \in \mathcal{H}}{\arg\max}\, \mathcal{L}_{\mathcal{S}}(h)$$

我们常见的正则化，就是一种归纳偏好， $L1$正则化表明我们的归纳偏好是更喜欢参数稀疏的假设。

1.4 为什么可以学习到东西

本节旨在说明：当拥有足够多样本时，在有限假设空间$\mathcal{H}$中，经验风险最小化$ERM_{\mathcal{H}}$原则不会出现过拟合，即我们通过训练样本，可以找到一个足够好的假设h_s，在真实世界中表现也足够好。

定义1 可实现性假设 存在$h^* \in \mathcal{H}$，使得$\mathcal{L}_{\mathcal{D},f}(h^*) = 0$。：

该假设意味着，对于随机样本集$\mathcal{S}$，由概率分布$\mathcal{D}$采样，由标记函数f标记，以概率1使得$L_{\mathcal{S}}(h^*) = 0$，其中样本集$\mathcal{S}$中的样本是独立同分布的。我们定义$h_{\mathcal{S}}$为对$\mathcal{S}$ 利用$ERM_{\mathcal{H}}$得到的结果:

$$h_{\mathcal{S}} \in \arg\max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{S}}(h)$$

因为样本集$\mathcal{S}$仍是领域集$\mathcal{X}$的子集，是根据分布随机得到的实例集合，会有一定概率使得采样得到的样本不具有代表性，并不能反映真实的总体情况，所以根据样本集$\mathcal{S}$得到的假设$h_{\mathcal{S}}$ 并不一定准确，在真实世界中的表现有可能很差。因此，我们选择一定程度的容忍，容忍会有一定几率采样到不具有代表性的样本，一般来说，我们定义采样得到不具有代表性样本的概率之多为δ，则$1-\delta$为置信参数。同时对于假设的预测效果，我们能容忍的损失上限为ϵ，称为精度参数，如果 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$，那么这是一个差的假设，反之，则是一个好的假设。

对于我们的样本集$\mathcal{S}$，最好的假设$h_{\mathcal{S}}$仍有 $\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$，则这组样本采样失败，因为我们的$ERM_{\mathcal{H}}$无法从$\mathcal{S}$中学到有用的东西。当样本集中所有样本$(x_1, ..., x_m)$，都不具有代表性，我们记为:

$$\{x_i | x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}$$

则采样失败的概率上界为:

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\})$$

因为每个样本都是不具有代表性的样本，所以是采样失败的概率的上界。设$\mathcal{H}_B$为差的假设的集合：

$$\mathcal{H}_B = \{h | \mathcal{L}_{\mathcal{D},f}(h) > \epsilon, h \in \mathcal{H}\}$$

同时设：

$$M = \{x | x \in \mathcal{S}, \exists h \in \mathcal{H}_B, \mathcal{L}_{\mathcal{S}}(h) = 0\}$$

为误导集，误导集使差的假设在训练样本$\mathcal{S}$ 上表现良好，但在真实世界中表现较差。因为有可实现假设

$$\exists h^*, \mathcal{L}_{\mathcal{D},f}(h^*) = 0$$

且有

$$h_{\mathcal{S}} \in \underset{h \in \mathcal{H}}{\arg\max}\, \mathcal{L}_{\mathcal{S}}(h)$$

产生$\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$ 是因为样本集不好，样本集采样失败，所以，当且仅当$\mathcal{S} \subseteq M$时，才会出现$\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon$，我们将其表示为：

$$\{x_i | x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\} \subseteq M$$

所以我们有

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \leq \mathcal{D}^m(M)$$

而M又可以写作：

$$M = \bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}$$

则：

$$\mathcal{D}^m(\{x_i | x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \leq \mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\})$$

由$P(A \cup B) \leq P(A) + P(B)$得：

$$\mathcal{D}^m(\bigcup_{h \in \mathcal{H}_B} \{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\})$$

其中我们将$\mathcal{S}$拆开，

$$\mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\} = \mathcal{D}^m(\{x_i | h(x_i) = f(x_i), x_i \in \mathcal{S}, i = 1, ...m\})$$

$$\mathcal{D}^m(\{x | x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\} = \prod_{i=1}^{m} \mathcal{D}(\{x_i | h(x_i) = f(x_i), x_i \in \mathcal{S}\})$$

根据$1 - \epsilon \leq e^{-\epsilon}$，对于等号右边的连乘的每一项都有：

$$\mathcal{D}(\{x_i | h(x_i) = y_i, x_i \in \mathcal{S}\}) = 1 - \mathcal{L}_{\mathcal{D},f}(h) \leq 1 - \epsilon$$

对于所有的$h \in \mathcal{H}_B$：

$$\mathcal{D}^m(\{x|x \in \mathcal{S}, \mathcal{L}_{\mathcal{S}}(h) = 0\} \le (1-\epsilon)^m \le e^{-\epsilon m}$$

可得：

$$\mathcal{D}^m(\{x_i|x_i \in \mathcal{S}, i = 1, ...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) \le |\mathcal{H}_B|e^{-\epsilon m} \le |\mathcal{H}|e^{-\epsilon m}$$

所以，若我们对采样失败概率的容忍大于采样失败的概率上限，我们认为是最后得到的$h_{\mathcal{S}}$学到了有用的的东西的，记为：

$$|\mathcal{H}|e^{-\epsilon m} \le \delta$$

借此我们可以推断出我们所需要拥有的样本数量为：

$$m \ge \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$$

推论1 设$\mathcal{H}$为一个有限假设集合， $\delta \in (0,1)$，$\epsilon > 0$，**当**

$$m \ge \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$$

成立时，从而对于任何标记函数f、任何分布$\mathcal{D}$，可实现性假设最少以$1-\delta$的概率，对于每个ERM假设$h_{\mathcal{S}}$，有以下不等式成立：

$$\mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) \le \epsilon$$

上述推论表明，对于足够大的m，由$ERM_{\mathcal{H}}$规则生成的有限假设会以$1-\delta$概率得到小于误差ϵ的近似正确解，概率在于容忍采样失败概率δ，近似在于容忍误差ϵ，即我们的算法，可能（容忍了失败概率）会学到令我们基本满意（容忍了一定误差）的东西。而概率近似正确，即PAC理论，将在第二章详述。

1.5 思路总结

1. 为什么不能有效学习？为什么会$\mathcal{L}_{\mathcal{D},f}(h) > \epsilon$?

2. 因为样本不具有代表性，学到的东西有问题，采样失败。

3. 怎么解决？

4. 降低采样失败概率至我们可以容忍的范围，小于δ。
5. 采样失败的概率上限小于δ。
6. $\sup \mathcal{D}^m(\{x_i|x_i \in \mathcal{S}, i=1,...m, \mathcal{L}_{\mathcal{D},f}(h_{\mathcal{S}}) > \epsilon\}) = |\mathcal{H}|e^{-\epsilon m} \leq \delta$
7. $m \geq \frac{\ln(|\mathcal{H}|/\delta)}{\epsilon}$
8. 当样本足够充足时，学习到的假设概率近似正确。