学习笔记

第 05 课：机器是如何学习的？

1.字面理解：**让机器学会某种东西**

概念：让计算机程序（机器），不是通过人类直接指定的规则，而是通过自身运行，习得（学习）事物的规律，和事物间的关联。

2.计算机只能进行数值和运算，必须将这些事物数值化，将事物的变化和不同事物之间的关联转化为运算。通过在这些数值之上的一系列运算来确定它们之间的关系，再根据一个全集之中个体之间的相互关系，来确定某个个体在整体（全集）中的位置。

3.机器学习的原理：

举例：猫妈妈让小猫去捉老鼠

小猫比作**机器（Machine）**，它成为“老鼠分类器“的过程，就叫做**学习（Learning）。**

猫妈妈给的那些照片是用于学习的**数据（Data）。**

猫妈妈告知要注意的几点，是这个分类器的特征（Feature）。

学习的结果——老鼠分类器——是一个**模型（Model）**。这个模型的类型可能是逻辑回归，或者朴素贝叶斯，或者决策树……总之是一个分类模型。

小猫思考的过程就是**算法（Algorithm）**。

数据-算法-模型

4.有监督和无监督学习

老鼠分类器

小马种族聚类

第 06 课：机器学习三要素之数据、模型、算法

1. 数据：

需要构建一个向量空间模型（Vector Space Model，VSM）。VSM 负责将格式（文字、图片、音频、视频）转化为一个个向量。

```
x_1 = [1,0]
x_2 = [0,0]
x_3 = [0,0]
x_4 = [0,1]
x_5 = [0,1]
x_6 = [1,0]
```

无标注的特征向量

```
x_1 = [1,1,1]; y = 1
x_2 = [1,1,1]; y = 1
x_3 = [1,1,1]; y = 1
x_4 = [1,1,1]; y = 1
x_5 = [1,1,1]; y = 1
x_6 = [0,1,1]; y = 0
x_7 = [0,0,0]; y = 0
x_8 = [0,1,0]; y = 0
x_9 = [0,0,1]; y = 0
```

有标注的特征向量

特征工程：

确定用哪些特征来表示数据；

确定用什么方式表达这些特征。

2. 模型：

模型是机器学习的结果，这个学习过程，称为训练（Train）。可理解为一个函数：y=f(x)。

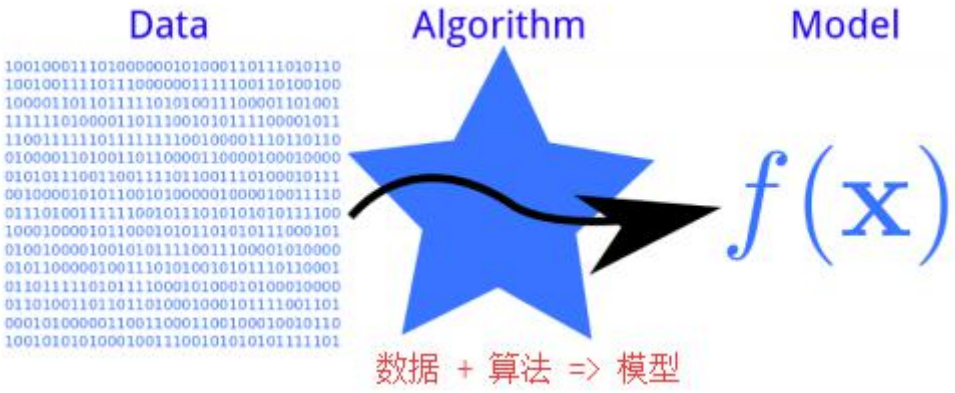
训练就是：根据已经被指定的 f(x) 的具体形式——模型类型，结合训练数据，计算出其中各个参数的具体取值的过程。

3. 算法：

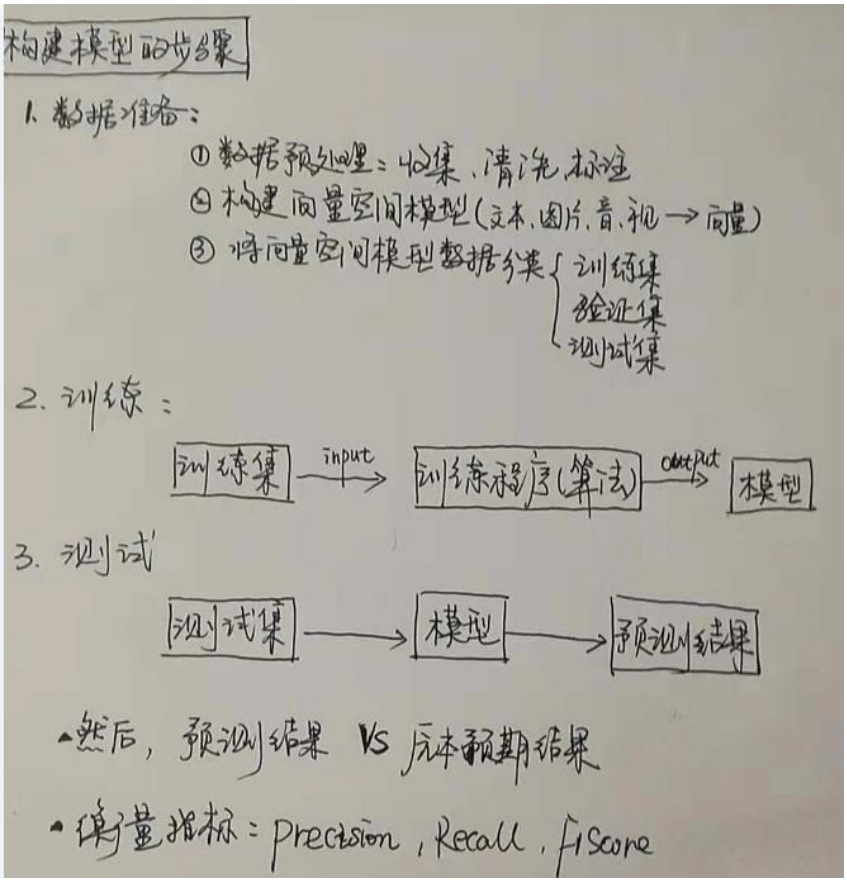
有监督学习的目标就是： 让训练数据的所有 x 经过 f(x) 计算后，获得的 y’ 与它们原本对应的 y 的差别尽量小。

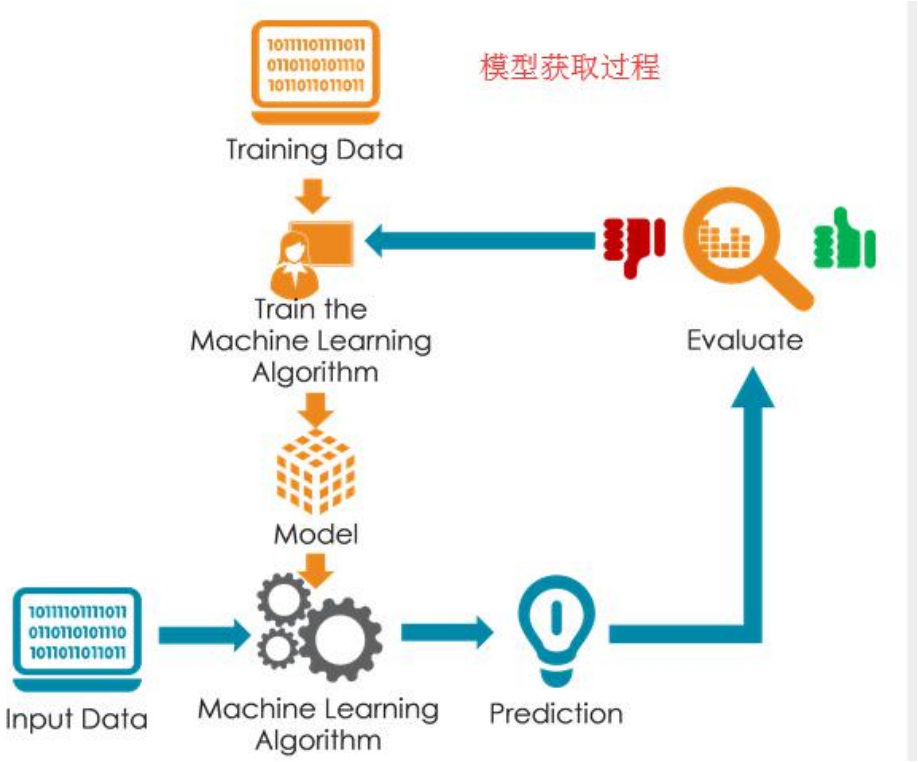
第 07 课：模型的获取和改进

1.获得模型的过程——训练——是将算法应用到数据上进行运算的过程。



步骤：

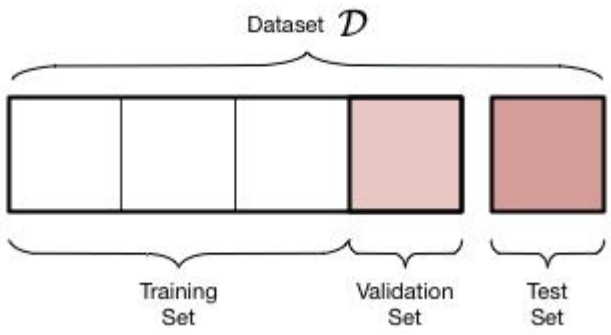




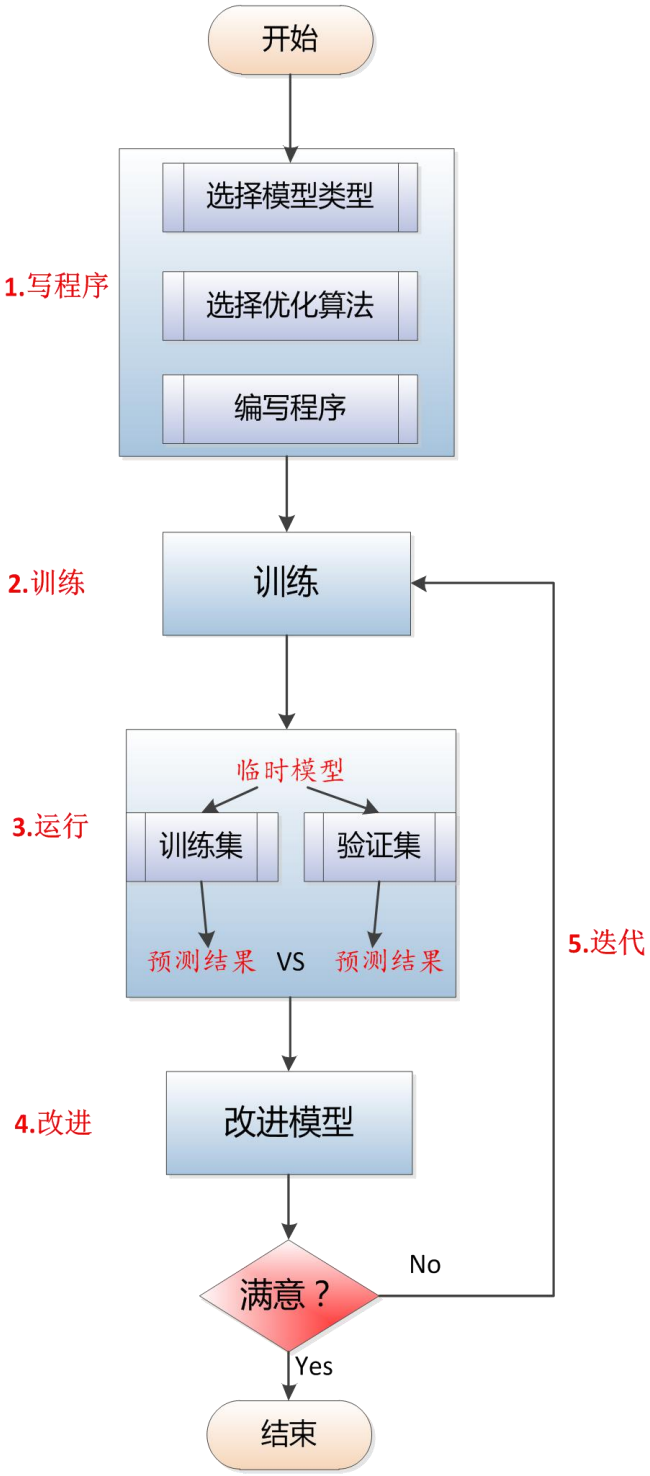
训练集（Train Set）：用来做训练的数据的集合。

验证集（Validation Set）：在训练的过程中，每个训练轮次结束后用来验证当前模型性能，为进一步优化模型提供参考的数据的集合。

测试集（Test Set）：用来测试的数据的集合，用于检验最终得出的模型的性能。每个集合都应当是独立的，和另外两个没有重叠。



2.训练的过程



3.改进模型

1）提高训练数据质量

数据的归一化（Normalization）、正则化（Regularization）标准化操作；采用 Bootstrap 等采样方法处理有限的训练/测试数据； 根据业务进行特征选取。

2.调参（算法）

例如用梯度下降方法学习 LR 模型时的步长（Alpha），用 BFGS 方法学习 Linear-chain CRF 时的权重（w）等。

制定目标->制定策略->执行->验证->调整策略

组合调参

3.模型类型选择

换个模型试试。比如，对于某个分类问题，Logistic Regression 不行，可以换 Decision Tree 或者 SVM 试试。

一般情况下，DL 模型（CNN、DNN、RNN、LSTM 等等）对于训练数据的需求比我们今天讲的统计学习模型要高至少一个量级。在训练数据不足的情况下，DL 模型很可能性能更差。

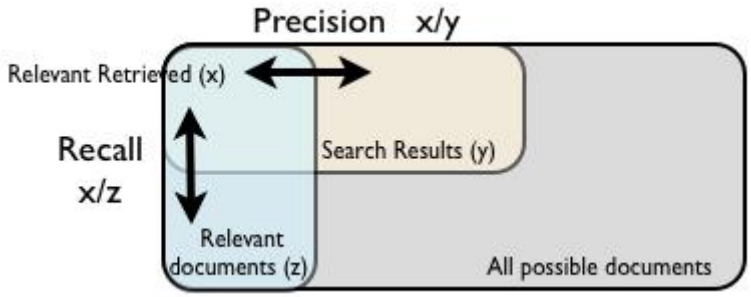
选对的，别选贵的。

第 08 课：模型的质量和评判指标

1.衡量模型质量

分类模型评判指标： **Precision、Recall 和 F1Score**

实际\预测	预测类为 Class_A	预测类为其他类
实际类为 Class_A	TP：实际为 Class_A，也被正确预测的测试数据条数	FN：实际为 Class_A，但被预测为其他类的测试数据条数
实际类为其他类	FP：实际不是 Class_A，*但被预测为 Class_A* 的数据条数	TN：实际不是 Class_A，*也没有被测试为 Class_A* 的数据条数



精准率：Precision=TP/（TP+FP），即在所有被预测为 Class_A 的测试数据中，预测正确的比率。

召回率：Recall=TP/（TP+FN），即在所有实际为 Class_A 的测试数据中，预测正确的比率。

F1Score = 2*(Precision * Recall)/(Precision + Recall)

显然上面三个值都是越大越好，但往往在实际当中 P 和 R 是矛盾的，很难保证双高。

衡量模型整体质量，要综合看所有 10 套指标，而不是只看一套。

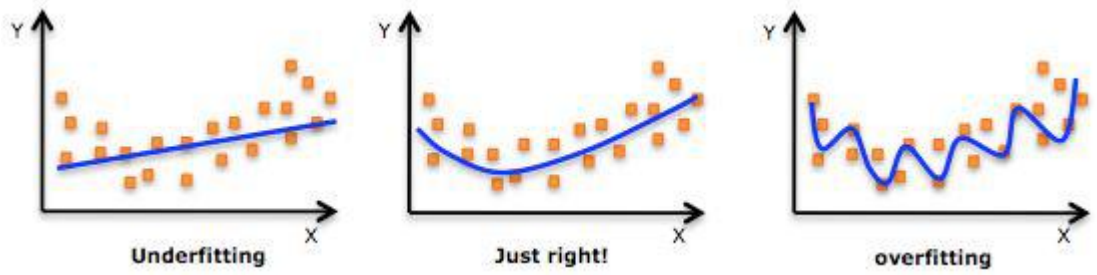
2.指标对应的是模型&数据集

任意的评价指标，都同时指向一个模型和一个数据集，两者缺一不可。

同样一套指标，用来衡量同一个模型在不同数据集上的预测成果，最后的分数值可能不同。

3.模型的偏差和过拟合

对训练集样本拟合程度的角度，可以分为两类：**欠拟合（Underfitting）**和**过拟合 （Overfitting）。**



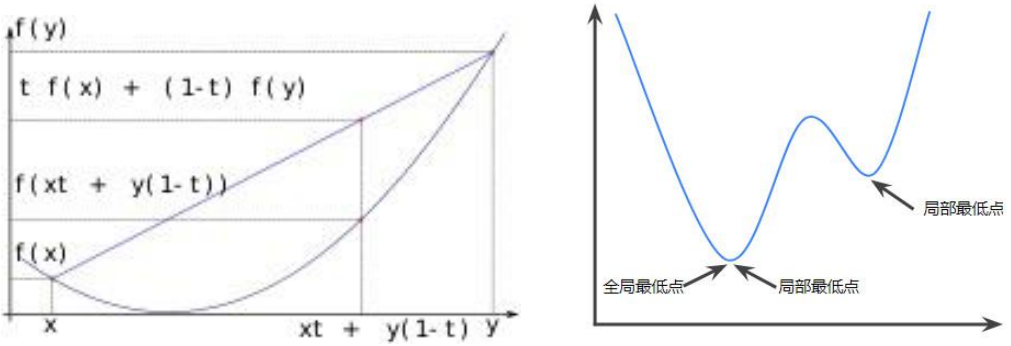
欠拟合：模型类型太过简单，特征选取不够。

过拟合：模型太过复杂，特征选择不当（过多或组合不当）。

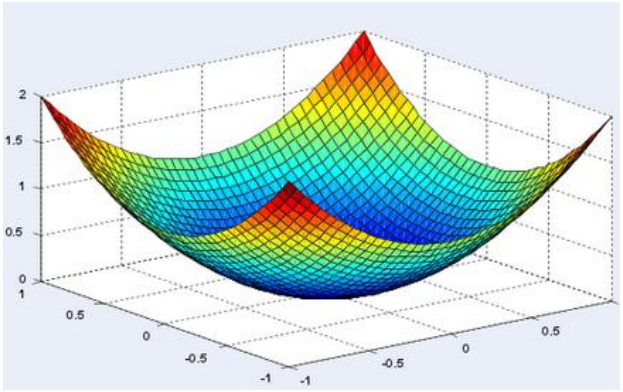
第 09 课：最常用的优化算法——梯度下降法

1.学习的目标（最小化目标函数**）**

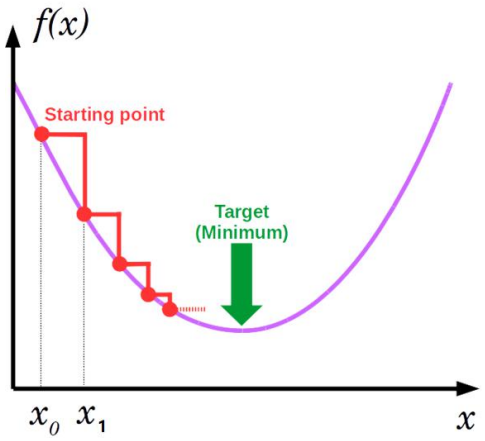
二维空间：



三维空间：

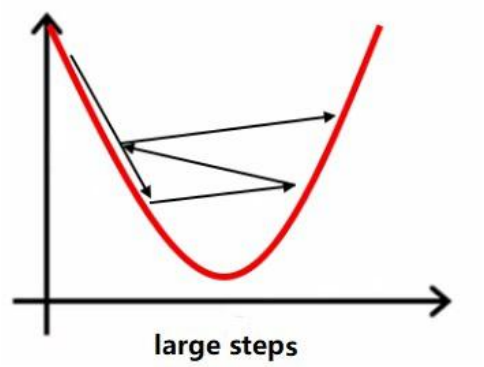


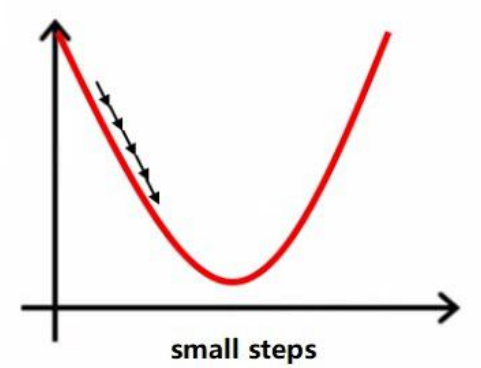
2.梯度下降法



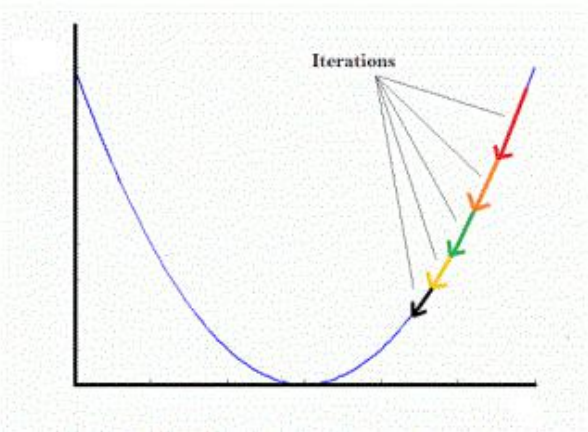
3.梯度下降的超参数

步长是算法自己学习不出来的，它必须由外界指定。这种算法不能学习，需要人为设定的参数，就叫做超参数。





为了平衡大小步伐的优缺点，也可以在一开始的时候先大步走，当所到达点斜率逐渐下降——函数梯度下降的趋势越来越缓和——以后，逐步调整，缩小步伐。比如下图这样：



4.梯度下降的难点

多个极小值，应该尝试不同的起点或者大步伐。