绪论

机器学习가 무엇이냐: 根据样本数据学习（训练）得到模型，用模型对新数据进行预测与决策。

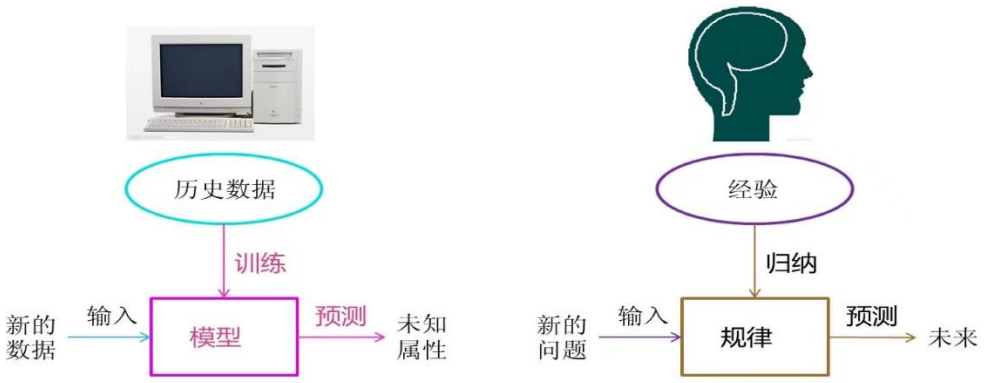
三要素：任务（T，要解决的问题），经验（E，训练数据），性能（P，模型在任务上的表现）

目标：通过利用经验来改善性能

研究内容：如何从“数据”产生“模型”的算法。

基本过程：基于数据，确定一个映射函数f，以及函数参数시타，建立映射关系：y = f (x;시타)

그럼 学习（训练）이 무엇이냐: 从数据中学得模型的过程。



机器学习的主要内容

数据预处理：数据清洗，集成（데이터들을 한곳으로 모은것），变换(형식 통일 시키는것)，数据规约（사용할 데이터를 간소하게 만드는것），规范化

特征工程：원시데이터를 알고리즘을 학습하는데 사용할 수 있는 特征으로 만드는 과정. 如特征选择，特征提取（特征构造）。

有监督学习:

分类与预测：분류 시스템안에서，数据对象의 내용에 맞게 数据对象을 所属类别로 확정 하는것. 回归分析，k近邻算法，决策树，贝叶斯分类，支持向量机，随机森林，集成学习，神经网络等

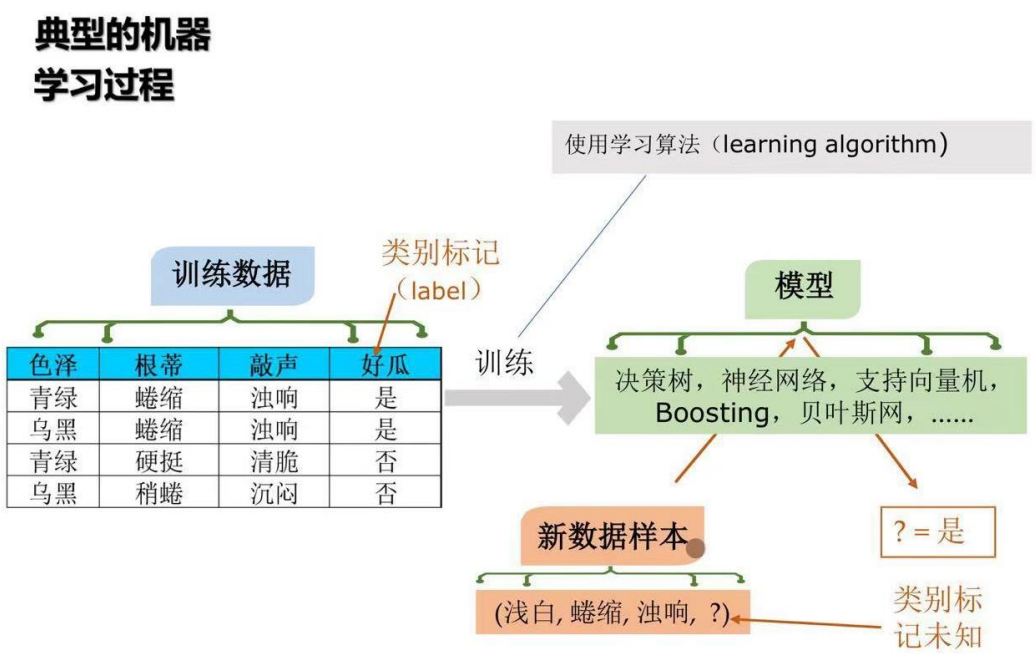
无监督学习:

聚类分析：对数据进行划分的过程， 分类랑 다른점은 划分的类是未知的。그래서 无监督学习라 하하는 것.

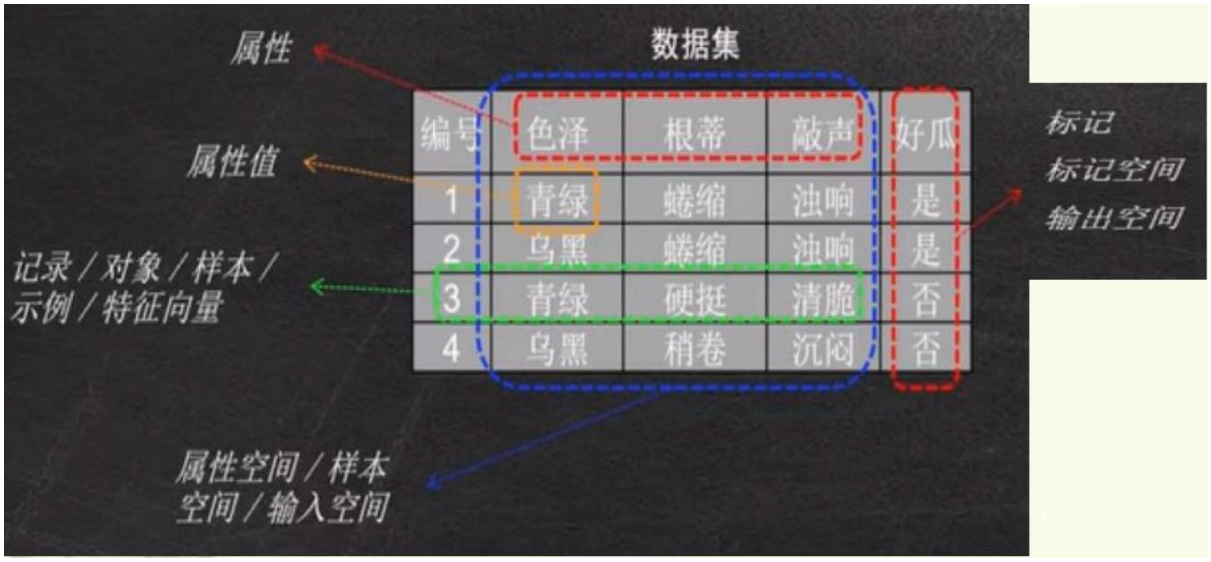
关联规则挖掘：发现大量数据集之间有趣的关联。

+孤立点分析：与数据的一般行为或模式不一致。多数为噪声，异常数据，常被剔除。

典型的机器学习过程：



数据集的基本术语：



基本术语：

分类：输出结果是离散值

回归：输出结果是连续值

监督学习：训练样本有标记

无监督学习：训练样本无标记

泛化能力：学得模型适用于新样本的能力

独立同分布：样本空间的全体样本都服从一个未知的分布且相互独立

数据挖掘（KDD）가 뭘까: 从数据集中提取或发现（挖掘）知识（有效，新颖的）的过程。보통 데이터에서 어떤 것을 뽑아낼수 있을지 모르는데, 이 기술을 사용해서 자주사용하는 모델을 뽑아내는 거임.(概念/类描述，关联分析，分类和预测，聚类分析，孤立点分析，趋势和演变分析)

그럼步骤는？：

1）数据准备——预处理

2）数据挖掘算法的选择——确定任务，选择有效挖掘算法（分类，聚类，关联规则挖掘……）

3）结果的解释评估——经过评价，排除无关模式

+细化：数据清洗——数据集成——数据选择——数据变换——**挖掘算法**——模式评估——知识表示

概念描述가 머임：대량데이터안에서 概述性总结을 진행하고, 간단하고 정확한 묘사를 얻는 것.

主要方法：概述性总结，数据泛化，두 개 데이터를进行对比，进行概化

归纳와 演绎의 차이가 뭘까

归纳：特殊到一般的“泛化”过程，即从具体的事实归结出一般性规律。

演绎：从一般到特殊的“特化”过程，即从基础原理推演出具体情况。

假设空间：由输入空间到输出空间的映射的集合。학습과정을 가설공간에서 탐색하는 과정이라고 볼 수 있는데, 그 목표는 훈련집과 알맞은 가설을 찾는 것.

예를 들어서 수박이 좋은 알인지는 색, 뿌리, 두드릴때 나는 소리, 총 3개의 요소에 의해 결정된다고 가정을 하면, 가설 공간은 이 3가지 요소의 属性值들의 조합 값만큼의 크기를 갖는다.

版本空间：학습과정을 가설공간에서 탐색하는 과정이라고 했는데, 가설 공간에서 적합하지 않는 것들을 다 배재한 후에, 훈련집과 알맞은 가설들의 집합이라고 할수 있다.

归纳偏好：机器学习算法在学习过程中对魔种类型假设的偏好。모든 알고리즘은 이 偏好가 있고, 알고리즘의 归纳偏好이 문제에 잘 맞아야지 좋은 성능을 낼 수 있다.

그렇기에 어떤 알고리즘을 쓸지에 대한 원칙이 있는데, 가장 자주 쓰이는 원칙이 바로,,,,,,

奥卡姆剃刀原理（시험에 나왔음）

= 解决问题的一种原则“简单有效原理”

核心思想：简洁性原则，即在多个可能的理论或模型中，选择假设最少，最为简单的那个。

쉽게 말해서 예측 성능이나 解释能力가 비슷한 모델들이 여러 개 있을때, 结构가 더 간단하고, 参数가 더 적은 모델을 선택해야한다.

特征选择할 때 자주 응용되기도 하는데, 가장 중요하고 解释能力가 강한 특징을 선택한다.

超参数调优：가장 간단한 超参数조합을 고르는데, 과도하게 超参数를 조절하는 것을 방지해서 模型이 过拟合하는 것을 막는다.

没有免费午餐定理（NLF：No Free Lunch Theorem）

어떠한 알고리즘도, 가능한 모든 문제에서 다른 모델들보다 더 좋은 퍼포먼스를 낼 수 없다.

즉 A알고리즘이 어떠한 문제에서 B알고리즘보다 좋은 성능을 보였다면, 다른 문제에서는 B가 더 좋을 수 있다. 즉 만능인 알고리즘이 없다는 뜻이다.

➔ 算法性能的相对性和条件性 ➔ 具体问题，具体分析!

명언 한 줄 남긴다,,,,

成功的机器学习引用不是拥有最好的算法，而是拥有最多的数据!

数据和特征决定了机器学习的上限，而模型和算法知识逼近这个上限而已。