线性模型

线性回归：试图学得一个通过属性的线性组合来进行预测的函数。简单，可解释性

核心思想：让预测值和真实值的均方差（距离）最小化

- 最小二乘法就是基于预测值和真实值的均方差最小化的方法来估计参数 w 和 b：

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

回归와 分类의 다른점：回归가 예측해야하는 목표 함수는 连续值이다.

回归分析法：통계학 원리를 사용해서, 대량 통계데이터에 대해서 수학적으로 처리를 해서, 변수과 출력값의 상관관계를 얻어낸다. 그리고 **建立相关性较好的回归方程**， 并加以外推，用于**预测因变量的变化**的分析方法

回归分析의 종류를 나누는 방법이 있는데,

自变量개수로 나누면 → 一元回归分析 & 多元回归分析

函数表达式로 나누면 → 线性回归分析 & 非线性回归分析

广义线性回归：将线性回归的预测值做一个非下线性的函数变化去逼近真实值

二分类问题

逻辑斯蒂回归야 말로 二分类问题에 사용되는 통계 모델이다, 기본 사상은 利用输入特征对事件发生的概率进行建模。그니까 sigmoid함수를 사용해서 선형회귀의 아웃풋을 확률로 만들어서 0-1 사이의 값을 갖게 하는거임.

核心思想

逻辑斯蒂回归将线性回归的输出通过 **Sigmoid 函数**（或逻辑斯蒂函数）转换成概率，以确保预测的结果在 0 到 1 之间。其主要步骤包括：

1. **线性组合**：将输入特征进行线性组合，得到一个分值 z：

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

其中 w 表示特征的权重，b 是偏置项。

2. **Sigmoid 函数**：将线性组合的结果 z 通过 Sigmoid 函数转换为概率 p：

$$p = \frac{1}{1 + e^{-z}}$$

其中 p 是事件发生的概率（例如，预测某样本属于类别 1 的概率）。

3. **分类决策**：若 p 大于某个阈值（通常是 0.5），则将样本分类为正类；否则分类为负类。

중요한건 얘는 连续属性值 문제만 해결할수 있고, 离散属性值 문제는 해결하지 못하기 때문에 离散值에 대해서는 따로 처리 해줘야 한다**对于离散属性值的处理**(시험에 나왔음):

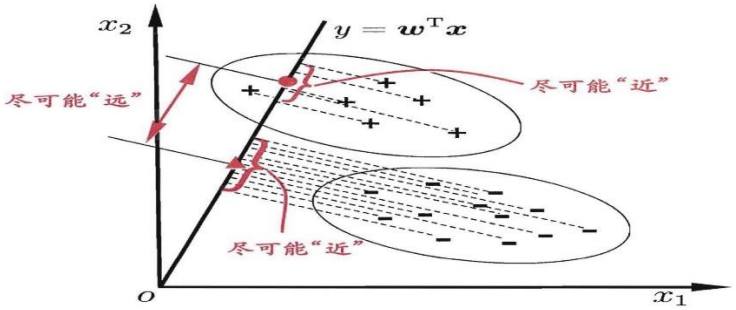
若属性间存在"序"关系 → 通过**连续化**将其转化为**连续值**

若属性间不存在"序"关系 → 通常可将**k个属性值**转化为k维向量

线性判别分析(LDA, Linear Discriminant Analysis)도 二分类问题에 쓰이는 선형학습방법이다.

1.훈련데이터셋에서 设法将样例投影到一条直线上， 使得同类样例的投影点尽可能接近， 异类样例的投影点尽可能远离。

2.새로운 样本에 대해서 분류를 할때는, 投影到同一条直线해서 投影点位置 에 따라서 분류 한다



多项式回归是研究一个因变量与一个或多个自变量间多项式的回归分析方法。그니까 1대1 함수 혹은 1대多 회귀분석인거임. 변수가 한 개면 一元多项式回归; 여러 개면 多元多项式回归。

다항식회귀문제는 变量转换해서 多元线性回归问题로 바꿔 해결할수 있음.

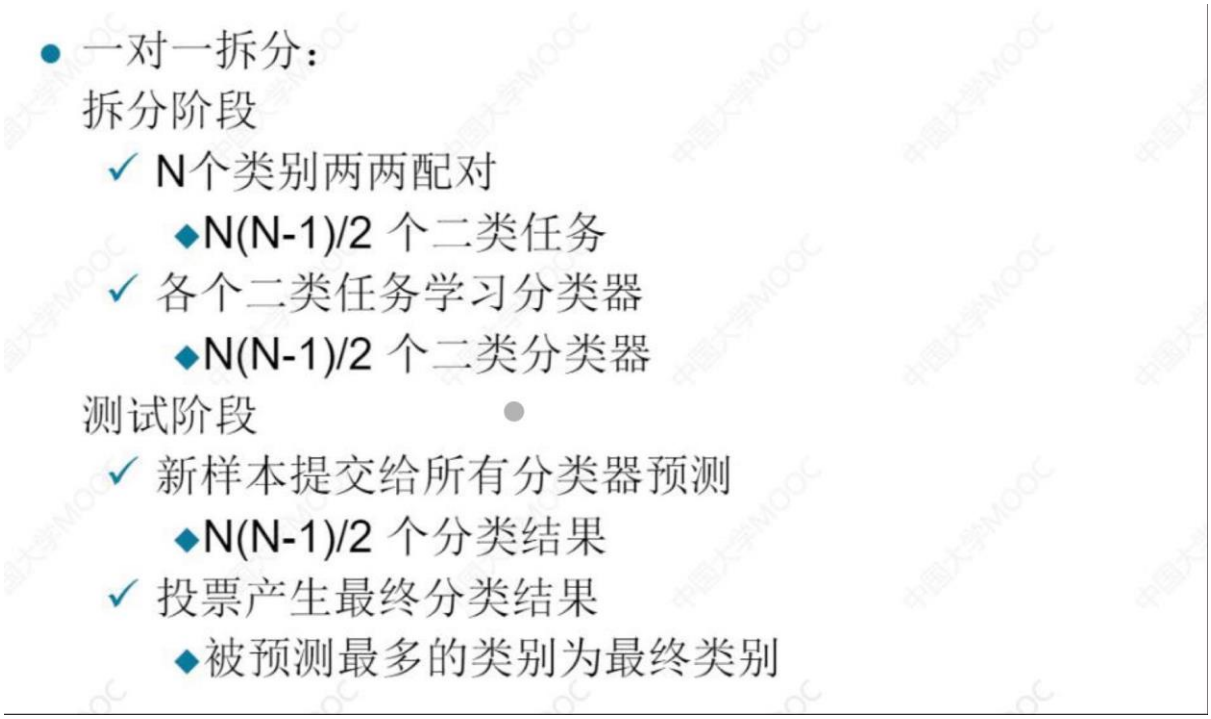
其他回归分析方法

1) KNN回归: 한样本의 k개 最近邻居를 찾아서, 평균값으로 예측하는거임
2) 决策树回归 : 주로 CART算法를 말하는 거임, 内部结点特征的取值은 "예", "아니요"임 **二叉树结构（基于基尼指标的划分度量）**

위에는 다 二分类问题였고, 지금부터는 자주 마주치게 되는 多分类问题이다. 주로 二分类学习器로 해결함.(对每个分类器的预测结果**进行集成**以获得最终的多分类结果)

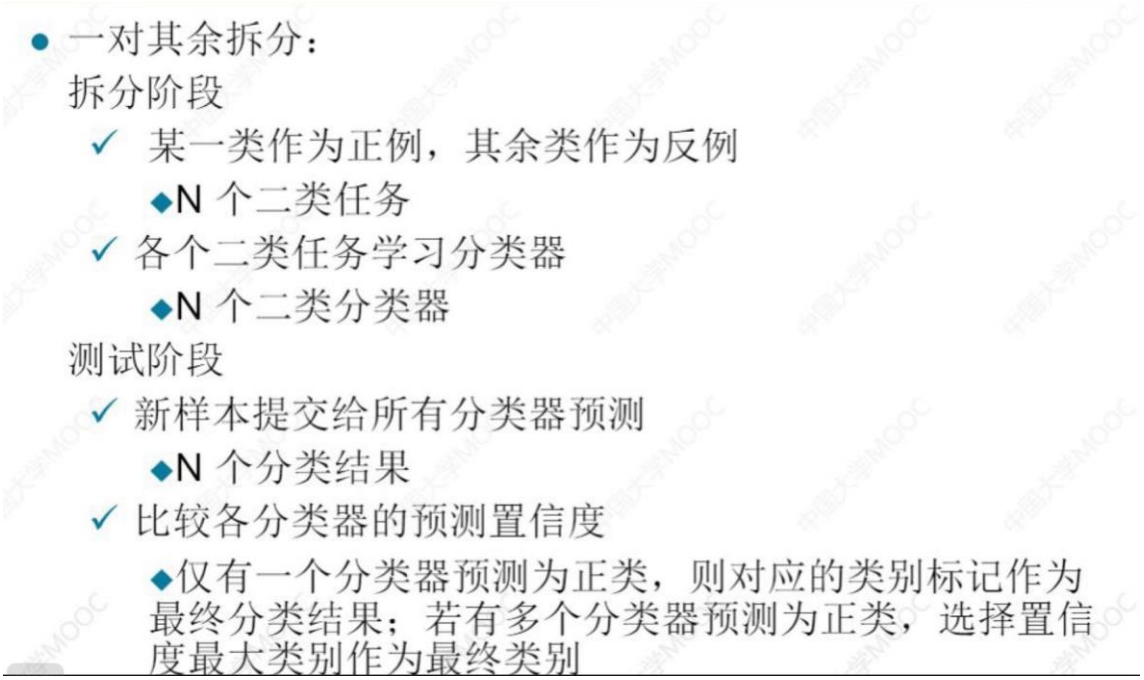
多分类学习은 一对一, 一对其余, 多对多로 나눠서 해결할수 있엉

一对一: 이 방법은 다중 클래스 분류를 여러 개의 이진 분류 문제로 변환하여 해결하는 방법

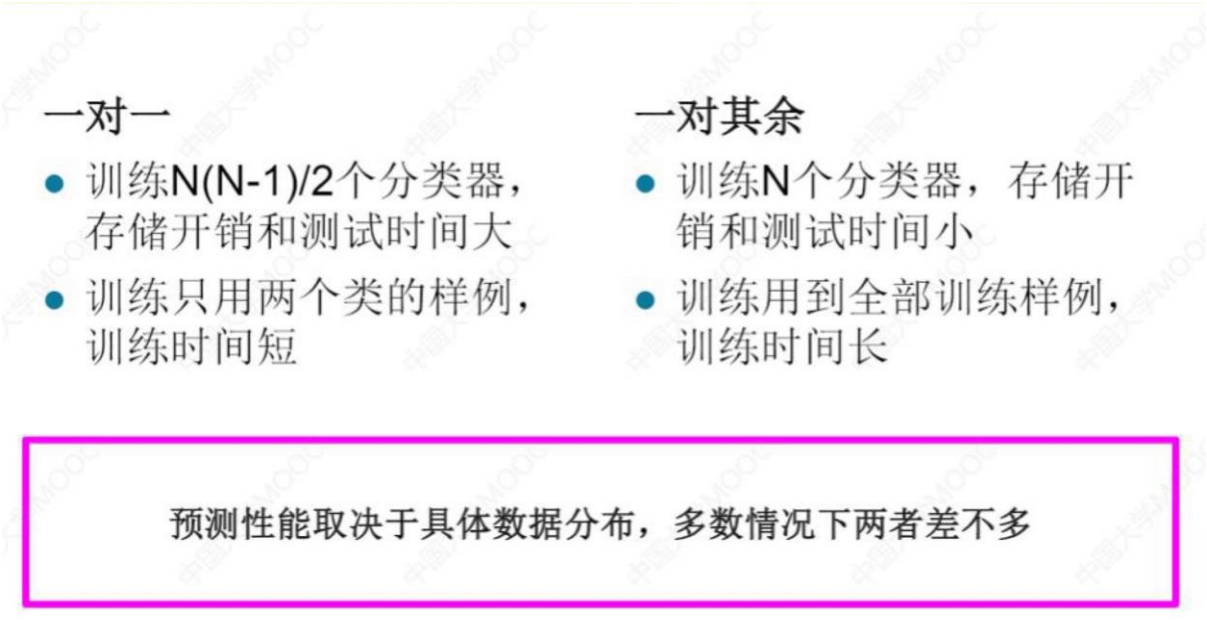


총 N개의 클래스가 있을 때, 두 개씩 조합해서 이진 분류기를 만들고, 각 두 클래스에 대해 별도의 이진 분류기를 학습시킴, 학습된 이진 분류기들이 각 새 샘플에 대해 예측을 수행하고, 투표를 진행해서 새 샘플이 어느 클래스에 가장 많이 할당되었는지를 확인해서 최종 클래스를 결정함.

一对其余:



특정 클래스르 긍정 사례로 취급하고, 나머지 모든 클래스를 부정 사례로 취급하여 이진 분류 문제로 변환함. 클래스가 N개 라고 했을 때 N개의 이진 분류 문제와 각각 별도의 분류기를 학습시킨다. 그 후, 새로운 샘플을 N개의 이진 분류기에 모두 넣어 예측하고, N개의 예측 결과가 나오니까 한 분류기만 해당 샘플을 긍정으로 예측한 경우 그 분류기의 클래스를 최종 결과로 선택함, 여러 분류기가 긍정으로 예측한 경우는 置信度가 가장 높은 분류기의 클래스를 최종 결과로 선택.



多对多 :

특정 클래스는 正类, 나머지 클래스는 反类로 간주하는 방식

纠错输出码（시험에 나왔음）는 다중 분류 문제를 다수의 이진 분류 문제로 변환하여 해결하는 기법, 에러 수정 코드의 개념을 적용하여 더 견고한 분류 결과를 얻음.

纠错输出码과정(시험에 나왔음):

编码단계:

1) N개의 클래스를 랜덤하게 M次 나눔. 每次划分将一部分类华为正类, 나머지는反类로
2) M개의 二类分类器, 得到每个类标记长度为M的编码

解码단계:

1) 测试样本을 M개의 二类分类器에게 주고 예측을 시킴
2) 生成一个长度为M的预测编码
3) 距离最小的类别是最终类别

作用（시험에 나왔음）:

1. 通过构造冗余信息，具有纠错能力，从而提升分类效果。
2. 即使某些二分类器出现错误，也可以通过距离度量找到最优分类结果。

类别不平衡问题（시험에 나왔었음）

不平衡数据集是指**数据集各个类别的样本量极不均衡**

处理方法！！ 이거 시험에 나왔어→ 欠采样，过采样，改进方法

欠采样: 从**多数类样本**中**随机选择少量样本，**再**合并**原有少数类样本作为新的训练数据集（하지만 데이터 손해보는거고 모델도 전체중의 일부만 학습 하게 되는 거임）

过采样: 从少数类的样本中进行随机采样来增加新的样本（하지만 데이터셋에서 일부 데이터들이 반복적으로 사용되니까, 过拟合될수도 있어）

➔ 改进: SMOTE算法（少数类别样本을 분석하고 토대로 **人工合成新样本**해서 추가）

➔ 改进2: EasyEnsemble算法（多数样本을 **N개의 子集**로 만들고, 매 子集가 少数样本이랑 같게 만든다, 그리고 각 子集들과 少数样本이 조합함. 그리고 **AdaBoost**를 사용해서 훈련시킴. 결국 **集成基分类器**）