# 3.1基本形式

假定有d个属性描述示例，**线性模型的目标是获得预测函数**：

其中w和b即为需要学习的内容。

因为向量w直观表达了各个属性在预测中的重要性，因此线性模型具有较为良好的可解释性。

# 3.2线性回归

根据给定数据集学习获得一个线性模型**尽可能准确地预测实值的输出标记**。以下以仅包含一个属性的示例为例。

Q：如何确定w和b呢？

A：**基于均方误差最小化解方程**。即求解w和b使得最小化。将上式分别对w和b求偏导，可以分别得到w和b的闭式解。

线性模型存在丰富的变化，比如对数线性回归，示例对应的输出在指数尺上变化。

更一般的，考虑单调可微函数g()作为联系函数，得到的模型被称为广义线性模型，令：

# 3.3对数几率回归

针对分类问题，需要在上述广义线性模型的基础上找一个**单调可微的函数**将**分类任务的真实标记y**和**线性回归模型的预测值**联系起来。

**对数几率函数**：是一种Sigmoid函数，输出值在0附近变化很陡，值域在[0, 1]区间。对应的模型称为“**对数几率回归**”（是一种分类学习方法）。优点：直接对分类可能性进行建模，无需事先假设数据分布。

# 3.4线性判别分析（LDA）

核心思想：给定训练样本集，**设法将样本投影到一条直线上**，使得**同类样本**的投影点尽量**近**，**异类样本**点的投影尽量**远**。

**最大化目标：**

同类样本点投影尽可能近->**同类样本的投影点协方差尽可能小**

异类样本点投影尽可能远->**类中心间的距离尽可能大**

可以将LDA推广到多分类任务。

# 3.5多分类学习

常见思路：将多分类任务拆解为多个二分类任务

常见拆解策略：  
1）**OvO**：将所有n个类别两两配对，从而产生n\*(n-1)/2个二分类任务。**训练时每个分类器仅被送入两个类别的数据**，测试数据被同时送入所有分类器，最终预测结果通过投票产生。

2）**OvR**：只需要训练n个分类器，每个分类器确定某个数据属于某个类别还是不适于该类别。**训练时每个分类器被送入全部训练数据**，训练代价高于OvO。

3）**MvM**：多个类别作为正例，若干其他类别作为反例。

**纠错输出码（ECOC）**：常用的MvM技术，将编码的思想引入类别拆分，并尽可能在解码过程中具有**容错性**。（每个分类器的预测结果作为编码的一位，最终通过编码确定预测结果，**码距**的存在保证了一定程度上的容错性）

# 3.6类别不平衡问题

类别不平衡：分类任务中不同类别的训练样本数差距很大。

基本策略：“**再缩放**”

常用方法：

1）**欠采样**：去除一些样本使得正反例数量接近，代表算法：EasyEnsemble

2）**过采样**：增加一些样例使得正反例数量接近，代表算法：SMOTE

3）**阈值移动**：在联系函数g()中不再假设0.5作为分类阈值，而是将正反例的数量比值作为分类阈值。