西瓜书复习笔记10

- K临近算法：
  - 懒惰学习：
    不对模型进行训练，直接对测试样本进行预测。（KNN）
  - 急切学习：
    训练阶段就对样本进行处理。
  - KNN：
    给定测试集，使用某种距离度量找出训练集中与其最靠近的k个训练样本，基于这k个邻居的信息进行预测。分类任务用投票法，回归任务用平均法。
  - 维数灾难：
    高维的情况下会出现样本稀疏、距离计算困难的问题。
  - 降维：
    通过数学变换将原始高维属性空间转为一个低维的子空间。
  - 直接降维：
    删属性、L1正则、特征选择
  - 线性降维：
    PCA
- PCA：
  - 线性代数复习：
    向量：箭头
    基向量：坐标系
    加法和乘法：移动和缩放
    内积、点积：等于B投影后的乘积（实数）

$$
\begin{aligned}
&(a_1, a_2, a_3)^T, (b_1, b_2, b_3)^T \\
&= a_1 b_1 + a_2 b_2 + a_3 b_3
\end{aligned}
$$

    基变换：换坐标系
    行列式：基向量变换后，所形成的面积缩放比例
    行列式为0的话，相当于市降维了。

$$
\det\left(\left[\begin{array}{ll} a & b \\ c & d \end{array}\right]\right) = ad - bc = 0
$$

    令A为变换的策略，向量X为变化前，向量y为变化后：

$$A^{-1}A\vec{x} = A^{-1}\vec{v}$$

AA^{-1}得到变化前的基坐标
矩阵的逆：
秩(Rank)：基变换后空间的维度
满秩：变换后维度不变
叉积：

$$\vec{v} \cdot \vec{w} = \vec{p}$$

大小是平行四边形大小，方向是与平行四边形垂直。
特征向量：基变化之后只有拉伸没有旋转的向量
特征值：特征向量上的拉伸倍数

$$A \cdot \vec{v} = \lambda \cdot \vec{v}$$

$\vec{v}$是特征向量
λ是特征值

$$\lambda\vec{v} = A\vec{v}$$
$$\lambda I\vec{v} = A\vec{v}$$

$A - \lambda I$是向量的变换，如果向量变换等于0，相当于是降维了，$det(A - \lambda I) = 0$，求出λ。

- PCA流程：
  1. 所有的点以原点为中心平移，使他们离原点的距离均值为0。
  2. 找到一条过原点的直线，使得所有的点在这条线上的投影尽可能分散（分散程度用方差表示），也就是投影后离原点距离最大。也就是转化为找到一个基向量使数据在新的坐标系中方差最大。

$$\mathrm{Var}(a) = \frac{1}{m}\sum_{i=1}^{N}(a_i)^2$$

  3. 构建协方差矩阵
     计算协方差矩阵的特征值和特征向量
     选择最大的特征值对应的特征向量
     将数据转到新的坐标系中
     （推导过程）

*4. 多维数据降维：
设有m条n维数据。
1. 将原始数据按列组成n行m列矩阵X
2. 将X的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值
3. 求出协方差矩阵
4. 求出协方差矩阵的特征值及对应的特征向量r
5. 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前k行组成矩阵P
6. 即为降维到k维后的数据

- KCPA：
  线性不可分的话，用PCA的核函数，先升维，再降维
- 流形学习：
  流形，比如三维空间中的二维曲线，圈圈。