Multivariate methods

Classification

> *假设某影院的历史上片有不同类型，像是爱情片，惊悚片或是侦探片，我们手头上有一个历史数据的dataframe，包含variables和predictors，现在新引进一部电影*
> ***Q:** 请问应把他归类于哪种？？*

Discriminant analysis

CVA/LDA/QDA

- Principle(LDA/QDA): to find linear functions

$$z = \vec{a}'\vec{x}$$

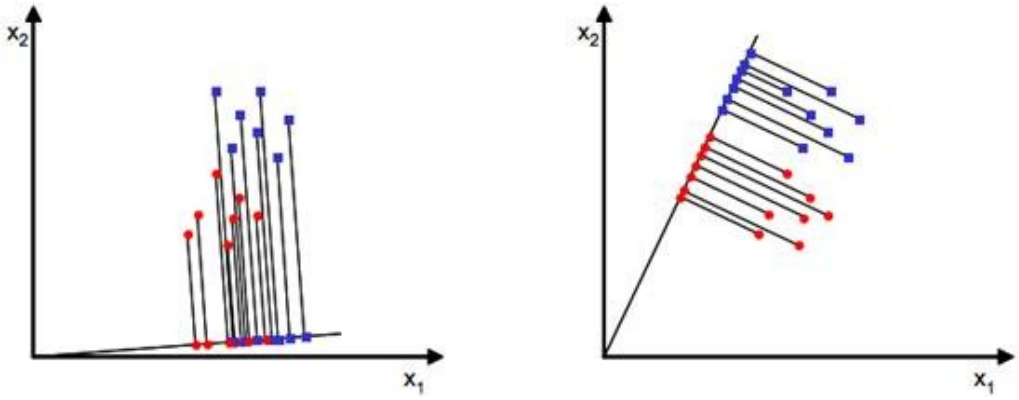
  - $\vec{x}$: column-vector
  - $\vec{a}'$: caonical vector
  - equal covariance matrices in LDA; unequal covariance matrices in QDA
- Baysian discriminant rule: *an object should be allocated to the class which has the largest posterior probability*

$$Class(\vec{x}) = argmax(P(y = C_g|\vec{x}))$$

- Principle (CVA): minimize the pooled **within**-class (W) covariance matrix and maximize **between**-class (B) covariance matrix

$$F = \frac{\vec{\mathbf{a}}' B\vec{\mathbf{a}}}{\vec{\mathbf{a}}' W\vec{\mathbf{a}}}$$

  - two-class problem: $W = \sum_C, B = \vec{\mu}_1 - \vec{\mu}_2$
  - B → infinity, W → 0
  - 想像在数据群旁有一条直线，这些样本点分别在这条线上进行投影形成新维度的样本，为了尽可能的区分不同的类别，我们只需使同类数据间的方差最小，而异类间的总均值之间距离最大



- maximize F: derivation of solution of maximization of F with respect to x
  - $F' = \frac{dF}{da} = 0$, so $\vec{a}$ is the eigonvetors of the matrix $W^{-1}B$

$$\vec{a} = \sum^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$$

Classification tree

- Gini index for all class K for leaf i: $1 - \sum_k P_{ik}$ (选最小)
  - P_{ik} = number of objects in leaf i belonging to class k / number of objects in leaf i
  - 想像在建树过程中，我们需要逐步选最优变量进行划分(characteristic of the object and being generalizable to the population of interest)，而最优变量的选择根据Gini值来决定，每次选变量都要对未划分的变量进行计算
  - Gini值的求解：假设当前一轮中的其中一个变量为gender, 进而根据变量的value形成2片leaves, 也就是说要分别算男性和女性各自不同的值，对于每种性别还要计算在不同类别的Gini

$$P_G = P_M + P_F = P_M 0 + P_M 1 + P_F 0 + P_F 1$$

- Gain for variables: the numbers of cases in a leaf with no misclassification (选最大)
  - *Note:* "cases" here is similar to "objects" in Gini
  - no misclassification means all nodes in one leaf belong to the same class

Factor analysis

- Goals:
  1. clusters variables into homogeneous sets
  2. allows us to select one single variable (factor) to represent many
  3. clustering of objects (people)
  4. [final goal] **to find a small number, m, of common factor leading to large community and small specific variance**
- Model:

$$X_j = \lambda_j F + e_j$$

where $F\ N(0, I)$ and $e\ N(0,)$

  - λ_j (the "loading" for X_j)
  - some important definitions:
    - **factor loading**: a measure of the importance of a common factor in explaining a variable (the degree to which each of variables correlates with each of the factors)
    - **community**: a measure of the variance in that variable htat is accounted for by the

common factor via the FA model (how much weight to put on X_j by the common factor)
        - **specific variance**: the variance of the results on that varaible that is not shared with the other variables via the common factor
        - **factor score**: the estimated value of a factor for given values of the variables on partivular individual
- computing community:
    1. (sum of square loading)^2
    2. 1- Uniqueness
- How many factors?
    - evaluating the d is important to determine the **upper bound** of the factor
    - using **PCA** to decide the number of factors
    - to solve the identifiability issues: we need to compute the degree of freedom

$$d = d.f.(m,n) = \frac{1}{2}(n-m)^2 - \frac{1}{2}(n+m)$$

where m factors and n variables
        - If $d < 0$ there are infinitely many solutions
        - If $d = 0$ then there is a unique solution to the problem (except for rotation)
        - In practice we usually have that d > 0 and an exact solution does not exist
- The **difference** between PCA and factor analysis:
    1. F postulates a model, P does not.

$$F_1 = a_1 X1 + a_2 X_2 + \dots$$

$$X_1 = a_1 F1 + a_2 F_2 + \dots$$

    2. Factor analysis tries to explain the correlations or covariances of the observed variables by means of a few common factors, while principal component analysis is concerned with explaining the variance of the observed variables
    3. In principal components, the component loadings do not change regardless of choice of number of components, while in factor analysis, there can be substantial changes in factors if the number of determined factors is changed.
    4. Calculation of principal component scores is straightforward, calculation of factor scores is more complex with a variety of possible methods.
    5. There is usually no relationship between the principal components of the sample covariance matrix and those of the sample correlation matrix. For maximum likelihood factor analysis, the results of analysing either matrix are equivalent

“

主成分分析是研究如何通过少数几个主成分来解释多变量的方差-协方差结构的分析方法，也就是求出少数几个主成分(变量)，使它们尽可能多地保留原始变量的信息，且彼此不相关。它是一种变换方法，即把给定的一组变量通过线性变换，转换为一组不相关的变量，在这种变换中，保持变量的总方差不变，同时具有最大方差，称为第一主成分；具有次大方差，称为第二主成分。依次类推。若共有p 个变量，实际应用中一般不是找p 个主成分，而是找出m $(m < p)$ 个主成分就够了，只要这m 个主成分能反映原来所有变量的绝大部分的方差。主成分分析可以作为因子分析的一种方法出现。
因子分析是根据相关性大小把变量分组，使得同组内的变量之间相关性较高，但不同的组的变量相关性较低，每组变量代表一个基本结构，这个基本结构称为公共因子。对于所研究的问题就可试图用最少个数的不可测的所谓公共因子的线性函数与特殊因子之和来描述原来观测的每一分量。通过因子分析得来的新变量是对每个原始变量进行内部剖析。因子分析不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子和特殊因子两部分。具体地说，就是要找出

某个问题中可直接测量的具有一定相关性的诸指标，如何受少数几个在专业中有意义、又不可直接测量到、且相对独立的因子支配的规律，从而可用各指标的测定来间接确定各因子的状态。因子分析只能解释部分变异，主成分分析能解释所有变异。

PCA

> *1) 拿到一个数学系的本科生期末考试成绩单，里面有三列，一列是对数学的兴趣程度，一列是复习时间，还有一列是考试成绩。我们知道要学好数学，需要有浓厚的兴趣，所以第二项与第一项强相关，第三项和第二项也是强相关。那是不是可以合并第一项和第二项呢?*
> *2) 拿到一个样本，特征非常多，而样例特别少，这样用回归去直接拟合非常困难，容易过度拟合。比如北京的房价：假设房子的特征是（大小、位置、朝向、是否学区房、建造年代、是否二手、层数、所在层数），搞了这么多特征，结果只有不到十个房子的样例。要拟合房子特征->房价的这么多特征，就会造成过度拟合。*

- **Motivation**: to reduce the dimensionality of the highly correlated variables
- **Basic idea**: to find a small number of uncorrelated linear combinations of original variables which explain the **most variation** of the data
    - PCA的思想是找到几个**方向向量**(Vector direction)，将n维特征映射到k维上（k<n），这k维是全新的正交特征。这k维特征称为主元，是重新构造出来的k维特征，而不是简单地从n维特征中去除其余n-k维特征。当我们把所有的数据投射到该向量上，我们希望该向量尽可能多的覆盖所有数据使得投影上去的两个最终端点分得最远，也就是方差最大，而平方均方误差能尽可能小。方向向量是一个经过原点的向量，而投射误差是从特征向量向该方向向量作垂线的长度。
- An small **example**: http://www.cnblogs.com/jerrylead/archive/2011/04/18/2020209.html

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

1. 归一化数据预处理
2. 求特征协方差矩阵
    - 如果数据是3维，那么协方差矩阵是 这里只有x和y，求解得 对角线上分别是x和y的方差，非对角线上是协方差。协方差大于0表示x和y若有一个增，另一个也增；小于0表示一个增，一个减；协方差为0时，两者独立。协方差绝对值越大，两者对彼此的影响越大，反之越小。
3. 求协方差的特征值和特征向量
    - 特征值对应特征向量，每列特征向量对应一个主成分

4. 将特征值按照从大到小的顺序排序
    - 选择其中最大的k个，然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵。
5. 将样本点投影到选取的特征向量上。

Prediction

“

***data**: winning times for the men's olympic 100m print, 1896-2008 **objective**: to predict the winning times (seconds) in London 2012*

- **basic process**: Problem → Data → Model → Train Predict
- **assumption**: there exists the relationship between year (x) and times (t)
- **model**: linear regression
- **criteria**: max(loss function)

$$\sum (t_n - (W_0 + W_1 x_n))^2$$

where we cannot use absolute funtion or just abstraction funtion, cos the predicted value is positive and the parameters could be more likely to be found

- **mathematical method**: least square
- **computational components** (python): $\bar{X}, \bar{t}, \bar{X^2}, \bar{Xt}$

Nearest Neighbor Classifier