PCA

主成分分析, 应用于降维, 可以提高模型训练效率.

步骤

1. 对特征进行均值归一化
2. 计算特征的协方差矩阵(特征有n维, 有m个样本)(其中$x^{(i)}$为第i个样本, n * 1), 得到协方差矩阵(n * n).

$$\sum = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)})(x^{(i)})^T$$

3. 对协方差矩阵使用奇异值分解

$$[U, S, V] = svd(\sum)$$

4. 获取U(n * n)的前K列, 转置后(k * n) 乘以$x^{(i)}$得到原始样本降维后的特征向量

选择K的方法

评估指标

- 投影误差的总和/训练样本的模平方的和

$$Average\ squared\ projection\ error\ divide\ by\ total\ variation\ in\ the\ data:$$
$$\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)} - x_{approx}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} <= 0.01$$

上述式子小于等于0.01(1%), 即可说明99%的方差被保留了, 个人理解是99%的信息被保留了.

将K从小到大依次变化, 分别计算该值, 选取小于等于0.01的值. 0.01是超参数.

但是这种方法的效率低, 计算量大.

- 计算S对角矩阵(n * n)

$$1 - \frac{\sum_{i=1}^{k}S_{ii}}{\sum_{i=1}^{n}S_{ii}} <= 0.01$$

不同的计算方式, 相同的作用, 相当于只要计算上式分式的结果即可.

使用方法

1. 在训练集上使用无标签的数据进行PCA压缩降维, x->z, 并且得到U矩阵, 将z, y数据进行训练
2. 在交叉验证集或测试集上使用U_{reduce} (即U的前K列)的转置乘以样本x, 使得x->z, 然后进行预测推断

说明

一般而言, 使用PCA降维可以将样本的维度降低1/5到1/10之间.

PCA降维是不能用于处理过拟合问题的.