主成分分析（PCA）

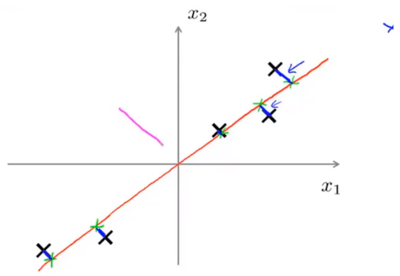
吴恩达机器学习公开课
(*http://study.163.com/course/courseLearn.htm?courseId=1004570029#/learn/video?lessonId=1052316982&courseId=1004570029*)

1. 降维

a) 节省内存
b) 可视化

2. PCA 目标：

找到一个超平面，最小化点到超平面的距离



3. PCA 算法

功能：将 n 维数据转化为 k 维后的数据；

输入：n 维数据：$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

输出：转化后的 k 维数据：$\{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$

步骤：

（1）归一化数据：$x_j^{(i)} = \dfrac{x_j^{(i)} - \mu_j}{s_j}$

（2）计算协方差：$\Sigma = \dfrac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$

（3）计算特征向量：$[U,S,V] = svd(\Sigma)$，其中 svd 表示 SVD 分解；

$U = \left[U^{(1)}, U^{(2)}, \ldots, U^{(m)}\right]$，取前 K 个 $U_{reduce} = U^{(1)}, U^{(2)}, \ldots, U^{(k)}$

（4）$z^{(i)} = {U_{reduce}}^T x^{(i)}$

5. 如何选择 K

a. 令：$x_{approx}^{(i)} = U_{reduce} z^{(i)}$

b. 对于 K=1;

计算 $U_{reduce}, z^{(1)}, z^{(2)}, \ldots, z^{(m)}, x_{approx}^{(1)}, x_{approx}^{(2)}, \ldots, x_{approx}^{m}$

Check if: $$\frac{\frac{1}{m}\left\| x^{(i)} - x_{approx}^{(i)} \right\|^2}{\frac{1}{m}\sum \left\| x^{(i)} \right\|^2} \le 0.01$$

c. 这里计算上式有一个简化过程：

$[U, S, V] = svd(\Sigma)$，后

$S = \begin{pmatrix} S_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & S_{nn} \end{pmatrix}$， 对于给定的 k 计算：$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \le 0.01$；

6. PCA 使用

X --> z --> $h_\theta(z) = \frac{1}{1 + e^{-\theta^T z}}$