机器学习复习10

2022年7月8日 星期五 11:56

Principal components analysis

- issues: some features are alike...
  or some have no relationships
  with the $h(x)$

- PCA counts
  ① get a form of

x	y
...	...

(features)

  ② get a form of $x_i - \bar{x}$ $y_i - \bar{y}$

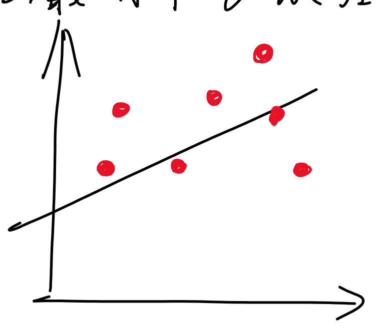
x	y
...	...

(new features)

  ③ get cov matrix
  ④ get matrix's feature number
  & feature vector
  ⑤ 样本点投影到对应的特征向量上...
  Final Data$(m \times k)$ = Data Adjust$(m \times n) \times$ EigenVectors$(n \times k)$

- Process before we got cov matrix
  1. let $\mu = \frac{1}{m}\sum_{i=1} x^{(i)}$
  2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$
  3. let $\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$
  4. Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$

- PCA basic theory
  1. 最大方差理论：
  最好的k维特征由n维样本点转化为
  k维后，每一维的样本方差都很大
  2. 最小平方误差理论：



$$\min \sum_{k=1}^{n} ||(x_k' - x_k)||^2$$

① 确定点：

$$J_0(x_0) = \sum_{k=1}^{n} ||x_0 - x_k||^2$$

$$= \sum_{k=1}^{n} ||x_0 - m||^2 + \sum_{k=1}^{n} ||x_k - m||^2$$

② 确定方向： $x_k' = m + a_k e$

最小平方误差： $J_1(a_1, \dots a_n, e) = \sum_{k=1}^{n} ||(x_k' - x_k)||^2$

$\Rightarrow a_k = e^t(x_k - m)$

固定 a_k，对 e 求偏导，代入 J 中，

$$J_1(e) = -e^t S e + \sum_{k=1}^{n} ||x_k - m||^2$$

其中 $S = \sum_{k=1}^{n} e^t (x_k - m)(x_k - m)^t e$

S 称为散列矩阵

求偏导： $\frac{\partial u}{\partial e} = 2Se - 2\lambda e$

$\Rightarrow Se = \lambda e$