Unsupervised learning.

Optimization objectives of K-means.

$c^{(i)}$ = index of cluster $(1, 2, \cdots, K)$ to which example $x^{(i)}$ is currently assigned
μ_k = cluster centroid k ($\mu_k \in \mathbb{R}^n$)
$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

Optimization objectives:

$$J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K) = \frac{1}{m} \sum \| x^{(i)} - \mu_{c^{(i)}} \|^2$$ ⇐ Distortion Function (畸变函数).

$$\min_{c^{(i)}, \mu_k} J(c^{(1)}, \cdots, c^{(m)}, \mu_1, \cdots, \mu_K)$$

最小化所有数据点与其所关联的数据中心点之间的距离之和.

Random initialization:

① $K < m$: the # of centroid k should be less them # of training sample m.

② Randomly pick k training examples.

③ Set $\mu_1, \cdots, \mu_K$ equal to these k examples.

Local optimal:

Sometimes we might encounter the local optima, what we can do is try K-means multiple times with different random initialization.

Choose number of clusters:

Elbow method

① run k means in a range of numbers.

② plot the cost corresponding to each k.

Dimensionality Reduction:

1. Data Compression:

Reduce data from 2D to 1D : use a new feature to represent those original features in a coherent relationship.

3D - 2D : project data to new directions / features with lower dimensionality.

2. Data Visualization.

Visualize the data to help us analyse them.

Principal Component Analysis problem Formulation.

Find a vector / surface (k-dimension) onto which to project the data so as to minimize the projection error.

找到一个向量，把所有的数据都投到该向量上，希望平均均方误差(distance)小.

PCA is not linear regression

PCA	Linear Regression.
reduce dimensionality of data	predict y according to feature x.
minimize projected error	minimize prediction error.
projection corresponding to direct vector.	projection corresponding to x-axis

principal component analysis algorithem

In PCA, we need to do:

① Compute direction vector u
② Compute the new features $z^{(i)}$

Algorithem = reduce from n to k.

Step 1: 均值归一化，计算所有特征的均值，ensure every feature has zero mean

Step 2: Compute "Covariance matrix"

$$\text{sigma} = \Sigma = \frac{1}{m}\sum_{i=1}^{n}\underbrace{(x^{(i)})}_{n\times 1}\underbrace{(x^{(i)})^T}_{1\times n} \rightarrow n\times n$$

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \cdots & x^{(n)} \\ | & | & & | \end{bmatrix}$$

Step 3: Compute "eigenvectors of matrix Σ" ⇒ 协方差矩阵的特征值.

```
[U S V] = svd(sigma);
```

sigma: $n\times n$ matrix.

$$U = \underbrace{\begin{bmatrix} | & | & | & & | \\ u^{(1)} & u^{(2)} & u^{(3)} & \cdots & u^{(n)} \\ | & | & | & & | \end{bmatrix}}_{k}$$

$(n\times n)$

$u \in \mathbb{R}^n$

$u^{(1)} \sim u^{(k)}$

$x \in \mathbb{R}^n \Rightarrow x \in \mathbb{R}^k$.

$z^{(i)} = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} \\ | & | & & | \end{bmatrix}^T x^{(i)} = \begin{bmatrix} — (u^{(1)})^T — \\ — (u^{(2)})^T — \\ \vdots \\ — (u^{(k)})^T — \end{bmatrix} \cdot x^{(i)}$

z: k-dimension

$n \times k$

U_{reduce}.

$k \times n$

$n \times 1$

$k \times 1$

$z_j = (u^{(j)})^T \cdot x$ => 第 j 个 direction vector 的对应的 new feature z_j.

Reconstruction from compressed representation

PCA is reducing the data from n-D to k-D ($n > k$), so that is should be reversible which means that $x_{approx} \approx x$.

$$x_{approx} = U_{reduce} \cdot z$$

So $x_{approx} \approx x$.

Choose the number of principal components (k):

PCA is minimizing the averaged projection error:

$$\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - x_{approx}^{(i)} \|^2$$

Total variation of data, how much the data varies:

$$\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} \|^2$$

Choose k to be smallest value so that

$$\frac{\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - x_{approx}^{(i)} \|^2}{\frac{1}{m} \sum \| x^{(i)} \|^2} \leq 0.01$$

← make sure the ratio is less than 0.01 (1%).

99% of variance is retained. many data are highly corelated.

call [U, S, V] = svd(sigma). once.

$$V = \begin{bmatrix} S_{11} & & & 0 \\ & S_{22} & & \\ & & S_{33} & \\ 0 & & & \ddots \\ & & & & S_{nn} \end{bmatrix}$$

for given k:

$$\frac{\frac{1}{m} \sum \| x^{(i)} - x_{approx}^{(i)} \|^2}{\frac{1}{m} \sum \| x^{(i)} \|^2} = 1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \leq 0.01$$

chose k for which

$$\frac{\sum_{i=1}^{k} s_{ri}}{\sum_{i=1}^{n} s_{ri}} \geq 0.99$$

Advice:

Only run PCA on traing set to learn Ureduce, then apply Ureduce in cross-validation or test sets in the same way.

① $x^{(1)}, x^{(2)}, \cdots, x^{(m)} \in \mathbb{R}^{10000} \xrightarrow{PCA} z^{(1)}, z^{(2)}, \cdots, z^{(m)} \in \mathbb{R}^{1000}$

② traing use new features.

③ In test/cross-validation, apply Ureduce to transfer x to z.

Bad use of PCA:

① To prevent overfitting: PCA just throws some information of data/features, but some times this may cause the loss of valuable information!

② To be a part of training:

Before running PCA, first try running whatever you want to do with the original/raw data, only if that doesn't do what you want, the implement PCA.