Machine Learning of Andrew Ng
pdf by CarolusRex

# Clustering

- Unsupervised learning introduction
- K-means algorithm
  - optimization objective
  - local optima
  - right value of K
  - drawbacks
- PCA(Principal Component Analysis)
  - Data processing
  - PCA algorithm
    - choosing k
  - reconstruction
  - notice

# Unsupervised learning introduction

data without label

# K-means algorithm

```
Randomly initialize $K$ cluster centroids $\mu_1, \mu_2, ..., \mu_k \in \mathbb{R}^n$
Repeat{
  for i = 1 to m
    $c^{(i)}$ = index (from $1$ to $K$) of cluster centroids closet to $x^{(i)}$
  for k = 1 to K
    $\mu_k$ = average(mean) of points assigned to cluster $k$
}
```

## optimization objective

$J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K) = \frac{1}{m} \sum ||x^{(i)} - \mu_{c^{(i)}}||^2$
$min_{(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K)} J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_K)$

## local optima

one method is fit more than one time, and choose the parameter whose cosf function $J$ is the smallest.

## right value of K

curves of cost function $J$ - number of clusters $K$, choose the elbow K.

## drawbacks

The discussion of drawbacks of K-means can be got here.

# PCA(Principal Component Analysis)

Reduce form $n$-dimension to $k$-dimension: Find $k$ vectors, $u^{(1)}, ..., u^{(k)}$ onto which to project the data, so as to minimize the projection error.

PCA is not linear regression. Its distance is between the point to the subpoint, not is the difference of values.

## Data processing

Train Set: $x^{(1)}, x^{(2)}, ..., x^{(m)}$
Preprocessing(feature scaling/mean normalization)
$\mu_j = \frac{1}{m} \sum x_j^{(i)}$
Replace each $x_j^{(i)}$ with $x_j - \mu_j$
If different features on different scales, scale features to have comparable range of values. $x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{s_j}$

## PCA algorithm

Compute "convariance matrix": $\Sigma = \frac{1}{m} \sum_{i=1}^{n} (x^{(i)})(x^{(i)})^T$
Compute "eigenvectors" of matrix $\Sigma$: $[U, S, V] = svd(sigma);$

```
Ureduce = U(:, 1: k);
z = Ureduce' * x;
```

### choosing k

Average squared projection error: $\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}-x_{approx}^{(i)}\|^2$
Total variation in the data: $\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2$
Typically, choose $k$ to be smallest value so that $\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}-x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2}\leqslant\epsilon\ (0.01)$ which means $1-\epsilon\ (0.99)$ of variance is retained.
It can be proved that the $k$ satisfy $\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}-x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2}\leqslant\epsilon\ (0.01)$ is equivalent to $\frac{\sum_{i=1}^{k}S_{ii}}{\sum_{i=1}^{n}S_{ii}}\geqslant 1-\epsilon\ (0.99)$

## reconstruction

$x_{approx}^{(1)}=U_{reduce}\cdot z^{(1)}$
It turns out that the $U$ matrix has the special property that it is a Unitary matrix. It has the property that $U^{-1}=U^{*}$, further, we have $U^{-1}=U^{T}$ because of real number field.

## notice

To prevent overfitting is the bad use of PCA, which can be implemented by regularization.