Machine Learning Week-5

ramay7

March 15, 2017

Contents

1 K-means Algorithm

Randomly initialize K cluster centroid $\mu_1, \mu_2 \cdots, \mu_k \in R^n$.

Repeate {
. for i = 1 to m
. $C^{(i)}$:= index(from 1 to K) of cluster centroid closest to $X^{(i)}$
. for k = 1 to K
. μ_k := average (mean) of points assiagned to cluster k
}

2 Principal Component Analysis Algorithm

PCA is not linear regression.

2.1 Data Preprocessing

Training set: $X^{(1)}\ X^{(2)}, \cdots, X^{(m)}$
Prerocessing (feature scaling/mean normalization):
. $\mu_j = \frac{1}{m}\sum_{i=1}^{m} X_j^{(i)}$
. Replace each $X_j^{(i)}$ with $X_j - \mu_j$
. If different features on different scales (e.g. X_1 = size of house, X_2 = number of bedrooms), scale features to have comparable range of values.

2.2 Reduce data from n-dimensions to K-dimensions

Compute "covariance matrix":
. Sigma $= \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$
Compute "eigenvectors" of matrix Sigma: . [U, S, V] = svd(Sigma)
where svd is the abbreviation of singular value decomposition.
. $U_{reduce} = U(:, 1 : K)$;

2.3 Reconstruction from compressed representation

. $X_{approx}^{(i)} = U_{reduce}Z^{(i)}$
where U_{reduce} is a n by k matrix and $Z^{(i)}$ is a k by 1 matrix.

3 Choosing K(number of principal conponents)

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:

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

which indicates 99% of variance is retained".

For $[U, S, V] = svd(Sigma)$:

$$S = \begin{bmatrix} S_{11} & 0 & 0 & ... & 0 \\ 0 & S_{22} & 0 & ... & 0 \\ 0 & 0 & S_{33} & ... & 0 \\ . & . & . & ... & 0 \\ 0 & 0 & 0 & ... & S_{nn} \end{bmatrix}$$

For given k:

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

4 Application of PCA

- To compress the data so it takes up less computer memory/disk space
- To reduce the dimension of the input data so as to speed up a learning algorithm
- To visualize high-dimensional data (by choosing k = 2 or k = 3)