1 ML Week 8 : Machine Learning Clustering

Suppose you have an unlabelled set of observations$\{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, \ldots, x^{(n)}\}$. You run $k-$means with 50 random initializations and obtain 50 different clusterings solution of the data. What is the recommended way of choosing which solution to use.

1.1 The Distortion Function

For each of the clustering solutions, the distortion function is computed as

$$\frac{1}{m}\sum_{i=1}^{m} \parallel x^{(i)} - \mu_c^{(i)} \parallel^2$$

The optimal clustering solution is the solution that minizes the distortion function. Since a lower value for a distortion function, implies a better clustering, you should choose the clustering with smallest value of the distortion function.

Finding Closest Centroids

In the cluste assignment phase of the algorithm, the algorithm assigns every training example $x^{(n)}$ to the closest centroid, given the current position of the centroids.

$$C^{(i)} := J \text{ that minimizes } \|x - \mu_j\|^2$$

- $C^{(i)}$ index of the centroid clostest to $x^{(i)}$
- μ_j is the position of the $j-$th centroid.

1.2 Computing Centroid Means

$$\mu_k := \frac{1}{abs(C_k)} \sum_{i \in C_k} x^{(i)}$$

A good way to initialize k-mean is to select k distinct examples from the training set and set the cluster centroids equal to these selected examples.

On every iteration of $k-$means, the cost function $J(C^{(1)}, C^{(2)}, \ldots, C^{(n)}, \mu_1, \ldots \mu_k)$

The distortion function should either stay the same or decrease, in particular it should not increase.

2 Week 8 Unsupervised Learning

2.1 Question 1.

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

YES Given a database of information about your users, automatically group them into different market segments.

YES Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.

Given historical weather records, predict the amount of rainfall tomorrow (this would be a real-valued output)

Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

2.2 Question 2.

Suppose we have three cluster centroids μ_1=[1 2], μ_2=[-3 0] and μ_3=[4 2]. Furthermore, we have a training example x(i)=[-2 1]. After a cluster assignment step, what will c(i) be?

YES c(i)=1
c(i)=2
c(i)=3
c(i) is not assigned

2.3 Question 3.

K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

- Move each cluster centroid μ_k, by setting it to be equal to the closest training example x(i)
- YES Move the cluster centroids, where the centroids μ_k are updated.
- The cluster assignment step, where the parameters c(i) are updated.
- YES The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting c(i)) to the closest training example x(i).

2.4 Question 4.

Suppose you have an unlabeled dataset x(1),. . . ,x(m). You run K-means with 50 different random

initializations, and obtain 50 different clusterings of the
data. What is the recommended way for choosing which one of
these 50 clusterings to use?

SELECTED Compute the distortion function J(c(1),. . . ,c(m),μ_1,. . . ,μ_k), and pick the one that minimizes this.

Use the elbow method.

Manually examine the clusterings, and pick the best one.

Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.

2.5 Question 5.

Which of the following statements are true? Select all that apply.

WRONG Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid

CORRECT On every iteration of K-means, the cost function J(c(1),. . . ,c(m),μ_1,. . . ,μ_k) (the distortion function) should either stay the same or decrease; in particular, it should not increase.

WRONG K-Means will always give the same results regardless of the initialization of the centroids.

CORRECT A good way to initialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples. Submit Quiz

3 Principal Component Analysis

3.1 Recommended Applications of PCA

- Data Compression: reducing the dimensionof input data $x^{(i)}$, which will be used in a supervised learning algorithm (i.e. use PCA data so that your supervised learning algorithm runs faster.
- Data Visualization: Reduce data to 2D (or 3D) so that it can be plotted.

Close

3.2 Question 1.

Consider the following 2D dataset:

Which of the following figures correspond to possible values that PCA may return for u(1) (the first eigenvector / first principal component)? Check all that apply (you may have to check more than one figure).

(Select both parallel vectors - CORRECT)

3.3 Question 2.

Which of the following is a reasonable way to select the number of principal components k?

(Recall that n is the dimensionality of the input data and m is the number of input examples.)

- Choose the value of k that minimizes the approximation error

- Choose k to be the smallest value so that at least 1% of the variance is retained.
- Choose k to be 99% of n (i.e., k=0.99 × n, rounded to the nearest integer).
- Choose k to be the smallest value so that at least 99% of the variance is retained. [CORRECT - Selected]

3.4 Question 3.

Suppose someone tells you that they ran PCA in such a way that "95% of the variance was retained." What is an equivalent statement to this?

3.5 Question 4.

Which of the following statements are true? Check all that apply.

- Given an input x ∈ R^n, PCA compresses it to a lower-dimensional vector z∈ R^k.
- PCA can be used only to reduce the dimensionality of data by 1 (such as 3D to 2D, or 2D to 1D).
- If the input features are on very different scales, it is a good idea to perform feature scaling before applying PCA.
- Feature scaling is not useful for PCA, since the eigenvector calculation (such as using Octave's svd(Sigma) routine) takes care of this automatically.

3.6 Question 5.

Which of the following are recommended applications of PCA? Select all that apply.

CORRECT Data compression: Reduce the dimension of your data, so that it takes up less memory / disk space.

CORRECT Data compression: Reduce the dimension of your input data x(i), which will be used in a supervised learning algorithm (i.e., use PCA so that your supervised learning algorithm runs faster).

- As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.
- Data visualization: To take 2D data, and find a different way of plotting it in 2D (using k=2).