1 Week 7 Unsupervised Learning

1.1 K-means Clustering

- In this this exercise, you will implement the K-means algorithm and use it for image compression.
- You will first start on an example 2D dataset that will help you gain an intuition of how the K-means algorithm works. After that, you wil use the K-means algorithm for image compression by reducing the number of colors that occur in an image to only those that are most common in that image.
- You will be using ex7.m for this part of the exercise.

1.2 Implementing K-means

- The K-means algorithm is a method to automatically cluster similar data examples together.
- The intuition behind K-means is an iterative procedure that starts by guessing the initial centroids, and then refines this guess by repeatedly assigning examples to their closest centroids and then recomputing the centroids based on the assignments.
- The inner-loop of the algorithm repeatedly carries out two steps:
  (i) Assigning each training example to its closest centroid
  (ii) Recomputing the mean of each centroid using the points assigned to it.
- The K-means algorithm will always converge to some final set of means for the centroids. Note that the converged solution may not always be ideal and depends on the initial setting of the centroids.
- Therefore, in practice the K-means algorithm is usually run a few times with different random initializations.
- One way to choose between these different solutions from different random initializations is to choose the one with the lowest cost function value (distortion).
- Random initialization The initial assignments of centroids for the example dataset in `ex7.m` were designed so that you will see the same gure as in Figure 1.
- In practice, a good strategy for initializing the centroids is to select random examples from the training set.

Question 1.

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

Question 2.

Suppose we have three cluster centroids μ_1=$[12]^T$, μ_2=$[-30]^T$ and μ_3=$[42]^T$. Furthermore, we have a training example x(i)=$[31]^T$. After a cluster assignment step, what will $c^{(i)}$ be?

Question 3.

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

Question 4.

Suppose you have an unlabeled dataset $\{x(1), \ldots, x(m)\}$. 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?

Question 5.

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

2 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.

2.1 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.

2.2 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.

3 Week 8 Unsupervised Learning

3.1 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.

3.2 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

3.3 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).

3.4 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.

3.5 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

4 Principal Component Analysis

4.1 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

4.2 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)

4.3 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]

4.4 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?

4.5 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.

4.6 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).

5 Week 9 Quiz. Anomaly Detection

Question 1.

For which of the following problems would **anomaly detection** be a suitable algorithm?

- From a large set of hospital patient records, predict which patients have a particular disease (say, the flu).
- From a large set of primary care patient records, identify individuals who might have unusual health conditions.
- CORRECT Given data from credit card transactions, classify each transaction according to type of purchase (for example: food, transportation, clothing).
- In a computer chip fabrication plant, identify microchips that might be defective.

Question 2. Variant A

Suppose you have trained an anomaly detection system that flags anomalies when p(x) is less than ϵ, and you find on the cross-validation set that it has too many false negatives (failing to flag a lot of anomalies). What should you do?

- Increase ϵ [CORRECT]
- Decrease ϵ

Question 2. Variant B

Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ϵ, and you find on the cross-validation set that it mis-flagging far too many good transactions as fradulent. What should you do?

- Increase ϵ
- Decrease ϵ [CORRECT]

Question 3.

Suppose you are developing an anomaly detection system to catch manufacturing defects in airplane engines. You model uses

You have two features x_1 = vibration intensity, and x_2 = heat generated.

Both x1 and x2 take on values between 0 and 1 (and are strictly greater than 0), and for most "normal" engines you expect that $x_1 \approx x_2$.

One of the suspected anomalies is that a flawed engine may vibrate very intensely even without generating much heat (large x1, small x2), even though the particular values of x1 and x2 may not fall outside their typical ranges of values. What additional feature x3 should you create to capture these types of anomalies:

- x3=x1/x2 [CORRECT]
- x3=x21 *times* x2
- x3=x1 *times* x2
- x3=x1+x2

Question 4.

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

- When evaluating an anomaly detection algorithm on the cross validation set (containing some positive and some negative examples), classification accuracy is usually a good evaluation metric to use.
- In anomaly detection, we fit a model p(x) to a set of negative (y=0) examples, without using any positive examples we may have collected of previously observed anomalies.
- **CORRECT** When developing an anomaly detection system, it is often useful to select an appropriate numerical performance metric to evaluate the effectiveness of the learning algorithm.
- In a typical anomaly detection setting, we have a large number of anomalous examples, and a relatively small number of normal/non-anomalous examples.

Question 5.

You have a 1-D dataset $\{x(1), \ldots, x(m)\}$ and you want to detect outliers in the dataset. You first plot the dataset and it looks like this:

Suppose you fit the gaussian distribution parameters μ_1 and σ_1^2 to this dataset. Which of the following values for μ_1 and σ_1^2 might you get?

- μ_1=-3,σ_1^2=4
- μ_1=-6,σ_1^2=4
- μ_1=-3,σ_1^2=2
- μ_1=-6,σ_1^2=2

6 Week 9 Quiz. Recommender Systems

Information Filtering System that attempts to recommend information items likely to be of interest to a user.

Commonly used algorithms

- $k-$means clustering
- Pearson's Rho
- Collaborative Filtering

Collaborative Filtering is the process of filtering for information or patterns using collaboration among multiple agents.

Applications: online news aggregation or similar items of clothings

best approached by other methods - prediction

Collaborative Filtering Gradient

$$\frac{\partial J}{\partial X_k^{(i)}} = \sum [] \theta_k^{(j)}$$

$$\frac{\partial J}{\partial \theta_k^{(i)}} = \sum [] X_k^{(j)}$$

No regularization applied

Anomaly Detection Gaussian Distribution Estimate Gaussian Distribution

For n feastures of X , compute the mean and variance for each feature

Selecting Threshold of ϵ.

Implement an algorithm to select the threshold ϵ using an F_i score on a cross validation set.

$P(X) < \epsilon$ is considered to be an anomaly.

6.1 F_1 Score