The Coders' Club

Machine Learning: G1

Week 8: Assignment

Topics:

Unsupervised Learning
Principal Component Analysis

Some Additional Courses:

- Machine Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/machine-learning-onramp.html

- Deep Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/deep-learning-onramp.html

- AI From the Data Center to the Edge – An Optimized Path Using Intel® Architecture (Intel AI)
  https://software.intel.com/en-us/ai/courses/data-center-to-edge

- Machine Learning (Intel)
  https://software.intel.com/en-us/ai/courses/machine-learning

- Deep Learning (Intel)
  https://software.intel.com/en-us/ai/courses/deep-learning

Unsupervised Learning

Q.1. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

- Given historical records, predict if tomorrow's weather will be sunny or rainy.
- From the user usage patterns on a website, figure out what different groups of users exist.
- Given many emails, you want to determine if they are spam or non-spam emails.
- Given a set of new articles from many different news websites, find out what are the main topics covered.

Q.2. Suppose we have three cluster centroids

$\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$ and $\mu_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$

Furthermore, we have a training example

$x^{(i)} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$.

After a cluster assignment step, what will $c^{(i)}$ be?

- $c^{(i)} = 1$
- $c^{(i)} = 3$
- $c^{(i)}$ is not assigned
- $c^{(i)} = 2$

Q.3. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

- Feature scaling, to ensure each feature is on a compatible scale to the others.
- Move the cluster centroids, where the centroids μ_k are updated.
- Using the elbow method to choose K.
- The cluster assignment step, where the parameters $c^{(i)}$ are updated.

Q.4. Suppose you have an unlabeled dataset $\{x^{(1)}, \dots , x^{(m)}\}$. 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?

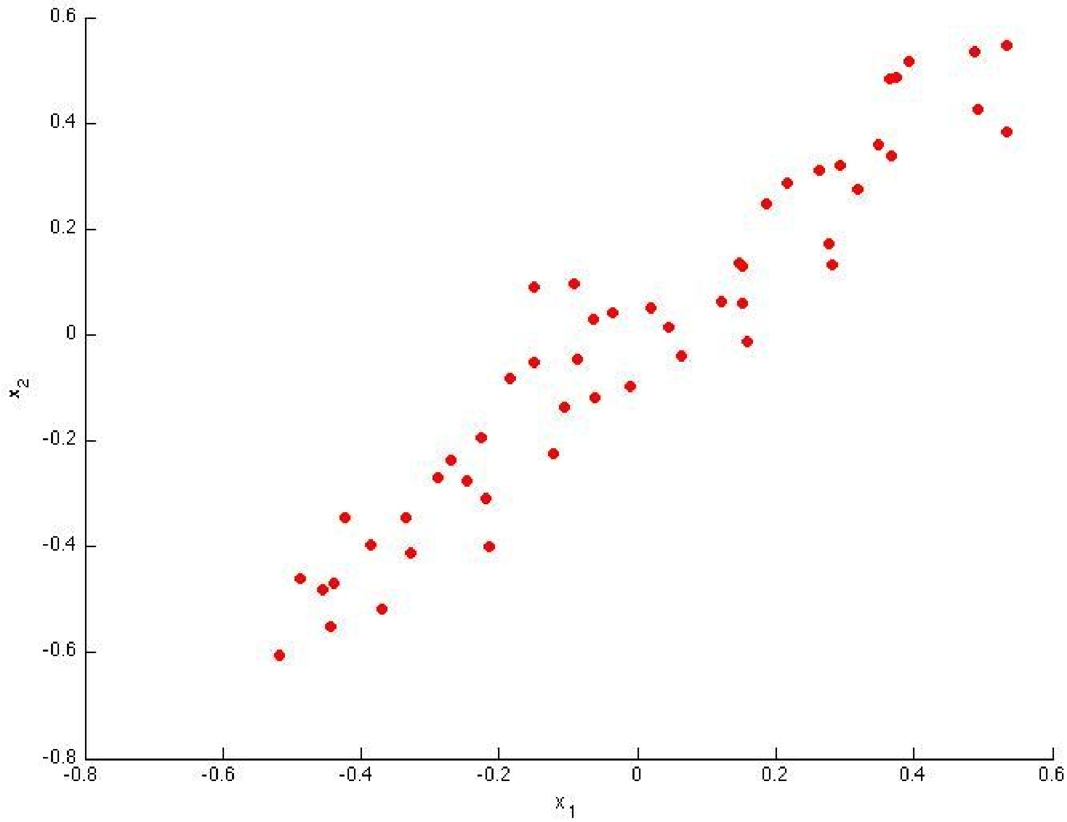
- Use the elbow method
- Compute the distortion function $J(c^{(1)}, \dots , c^{(m)}, \mu_1, \dots , \mu_k)$, and pick the one that minimizes this.
- Plot the data and the cluster centroids, and pick the clustering that gives the most “coherent” cluster centroids.
- Manually examine the clusterings, and pick the best one.

Q.5. Which of the following statements are true? Select all that apply.

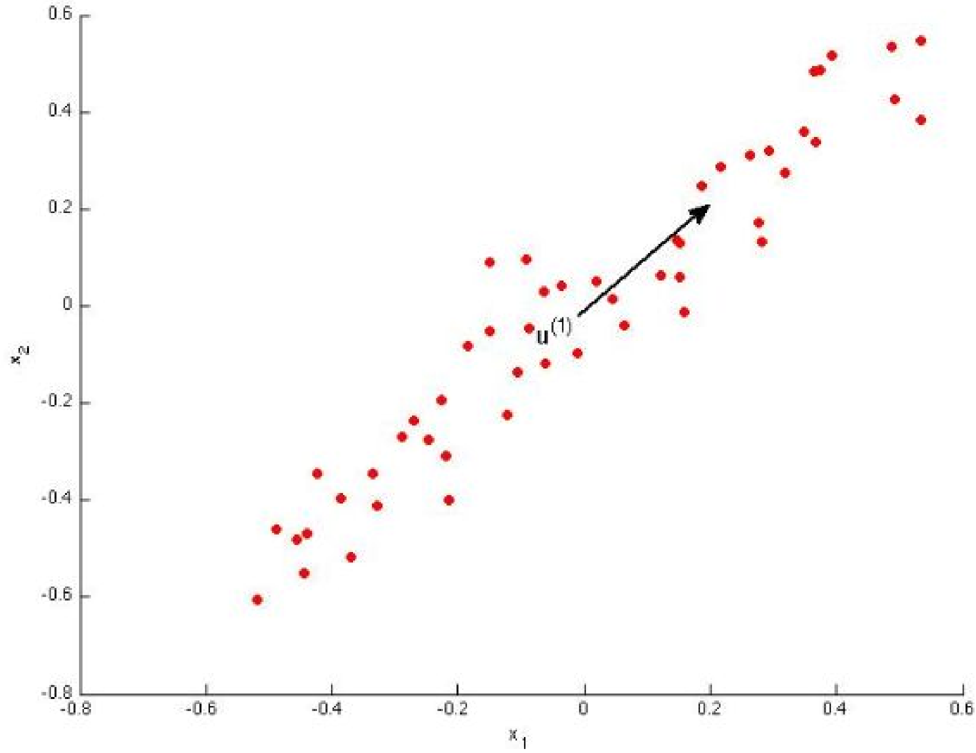
- Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid.
- K-means will always give the same results regardless of the initialization of the centroids.
- On every iteration of K-means, the cost function (distortion function) $J(c^{(1)}, \dots , c^{(m)}, \mu_1, \dots , \mu_k)$ should either stay the same or decrease; in particular, it should not increase.
- 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.

Principal Component Analysis

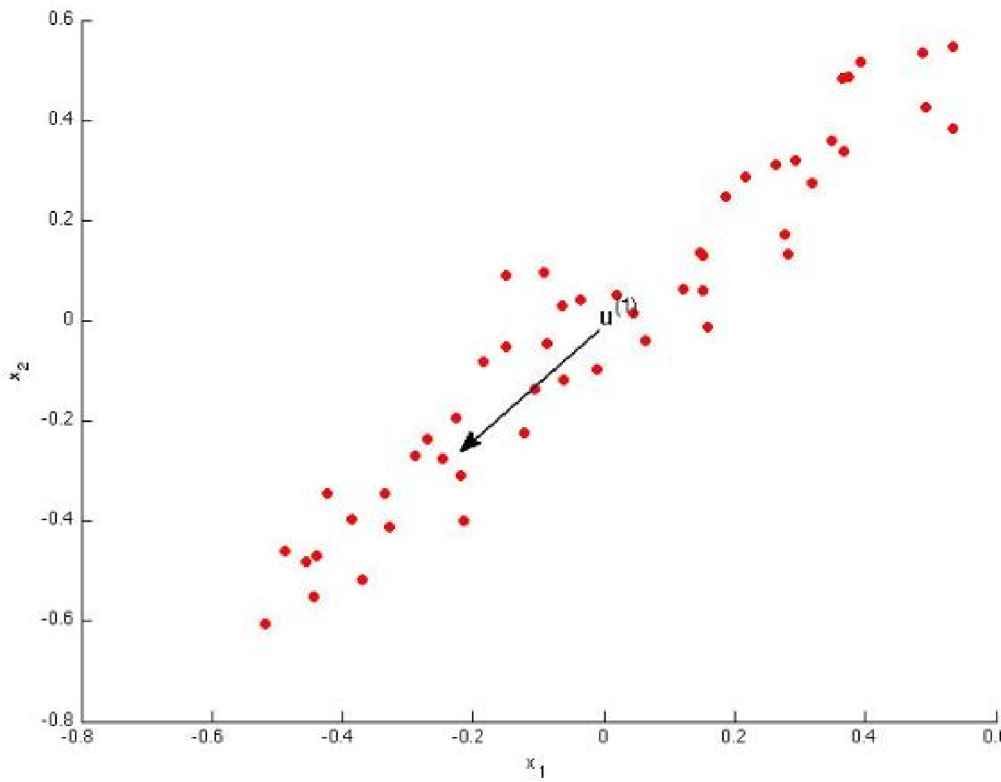
Q.1. Consider the following 2D dataset:



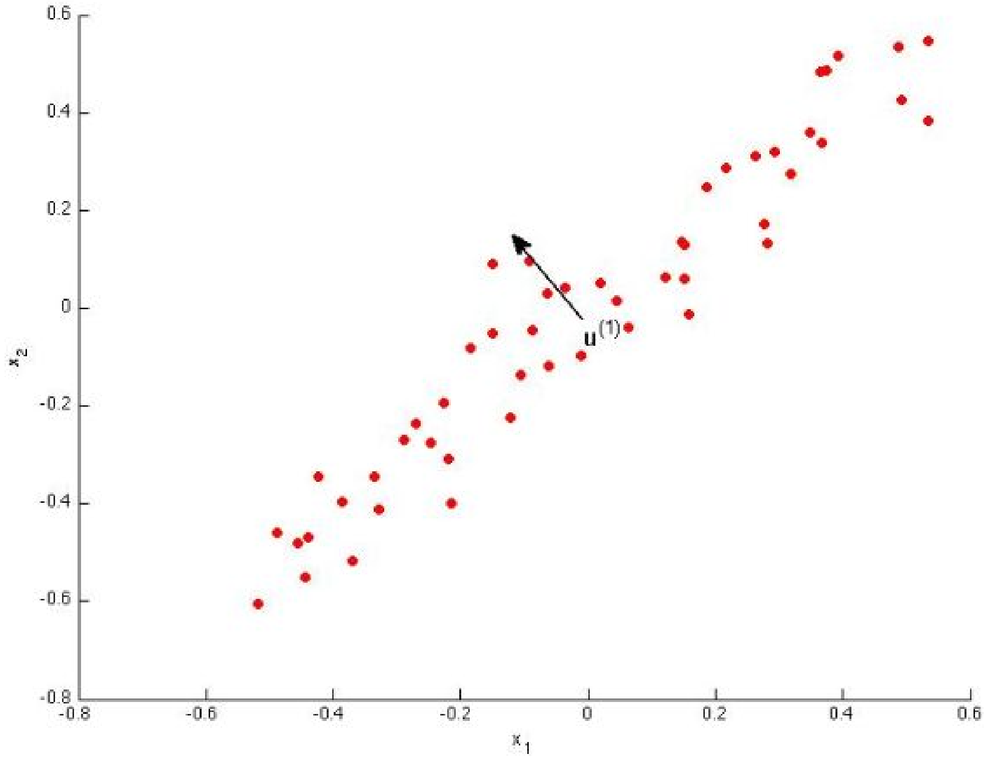
Which of the following figures correspond to possible values that PCA may return for $u^{(1)}$ (the first eigenvector/first principal component)? Check all that apply (you may have to check more than one figure).

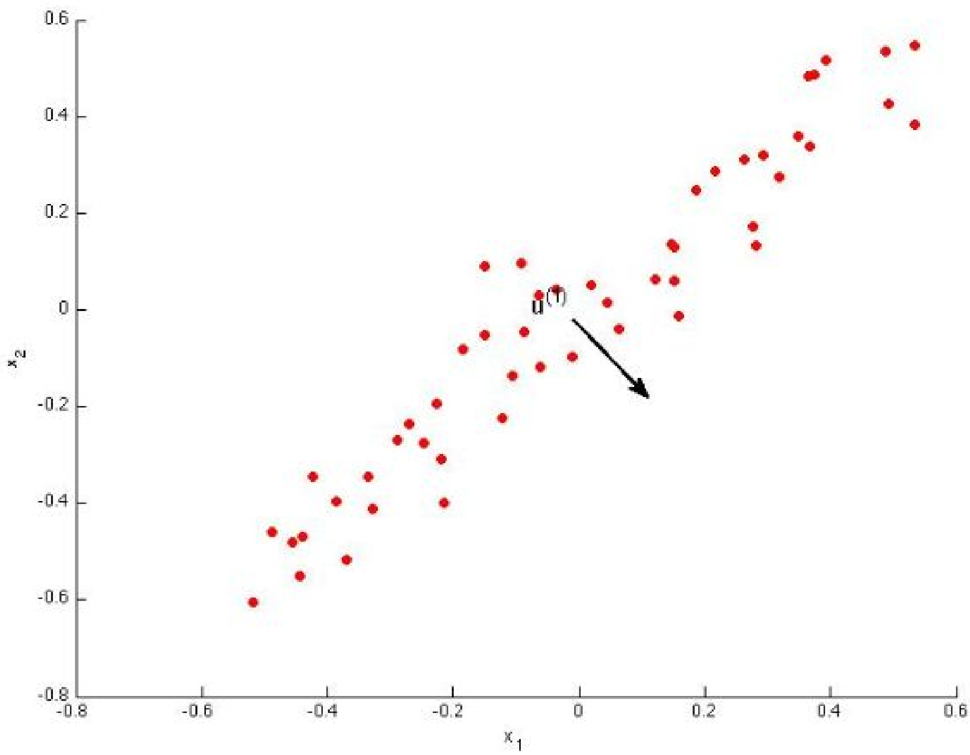


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
x_1
x_2
$\mathbf{u}^{(1)}$



0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
x_1
x_2
$\mathbf{u}^{(1)}$


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
x_1
x_2
$\mathbf{u}^{(1)}$


0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
x_1
x_2
$\mathbf{u}^{(1)}$

Q.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 k to be the smallest value so that at least 1% of the variance is retained.
- Choose k to be the smallest value so that at least 99% of the variance is retained.
- Choose the value of k that minimizes the approximation error $\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - x_{approx}^{(i)}\|^2$
- Choose k to be 99% of n (i.e. k=0.99 * n, rounded to the nearest integer).

Q.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?

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - x_{approx}^{(i)}\|^2} \geq 0.95$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)}\|^2} \geq 0.95$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)}\|^2} \geq 0.05$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m} \|x^{(i)}\|^2} \leq 0.05$$

Q.4. Which of the following statements are true? Check all that apply.

- PCA is susceptible to local optima; trying multiple random initializations may help.
- Given input data $x \in R^n$, it makes sense to run PCA only with values of k that satisfy $k \leq n$. (in particular, running it with $k = n$ is possible but not helpful, and $k > n$ does not make sense).
- Given only $z^{(i)}$ and U_{reduce} , there is no way to reconstruct any reasonable approximation to $x^{(i)}$.
- Even if all the input features are on very similar scales, we should still perform mean normalization (so that each feature has zero mean) before running PCA.

Q.5. Which of the following are recommended applications of PCA? Select all that apply.

- Preventing overfitting: Reduce the number of features (in a supervised learning problem), so that there are fewer parameters to learn.
- Data compression: Reduce the dimension of your data, so that it takes up less memory/disk space.
- To get more features to feed into a learning algorithm
- Data visualization: Reduce data to 2D (or 3D) so that it can be plotted.