ML w8 Unsupervised learning

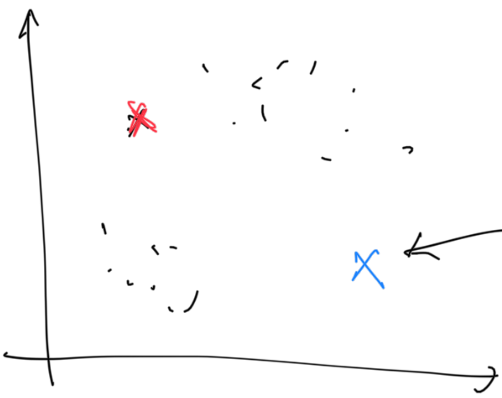
Dataset don't have "y" (label).

Training set $\{x^{(1)}, x^{(2)}, x^{(3)} \ldots x^{(m)}\}$

example.

- Market segmentation
- SN analysis.
- Organize computing clusters
- Astronomical data analysis.

K-Means Algorithm.



cluster center.

1) randomly pick

2) closer point classification

3) calculate mean for classified points

K-means algorithm

Input:

- K (number of cluster)
- Training set

μ_1 μ_2 k

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$

Unsupervised Learning

合計点数 5

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

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

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 $x^{(i)} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. After a cluster assignment step, what will $c^{(i)}$ be? **1点**

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

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 $x^{(i)} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. After a cluster assignment step, what will $c^{(i)}$ be? **0 / 1点**

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

不正解
$x^{(i)}$ is closest to μ_2, so $c^{(i)} = 2$, not 1

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

1点

- [x] The cluster assignment step, where the parameters $c^{(i)}$ are updated.
- [ ] Randomly initialize the cluster centroids.
- [ ] Test on the cross-validation set.
- [x] Move the cluster centroids, where the centroids μ_k are updated.

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?

1点

- ( ) 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.
- (•) Compute the distortion function $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_k)$, and pick the one that minimizes this.
- ( ) Use the elbow method.

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

1点

- [ ] The standard way of initializing K-means is setting $\mu_1 = \cdots = \mu_k$ to be equal to a vector of zeros.
- [ ] Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible.
- [x] For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.
- [x] If we are worried about K-means getting stuck in bad local optima, one way to ameliorate (reduce) this problem is if we try using multiple random initializations.

Dimentionality Reduction,

- Reduce data from 2D to 1D

inch centimeter

Round error

2D→1D

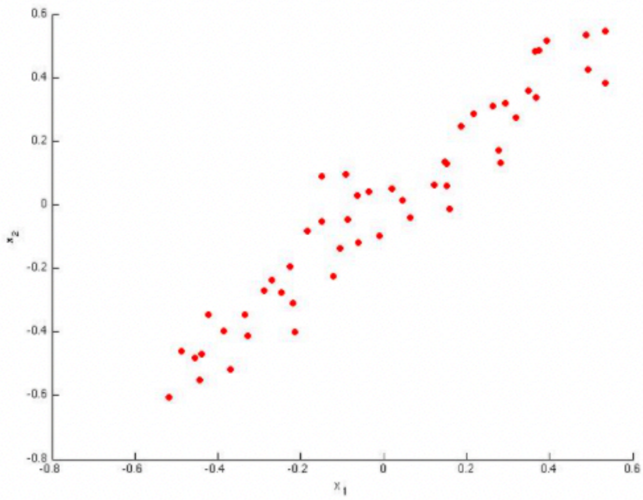
Reduce from 2D to 1D.

Principal Component Analysis

合計点数 5

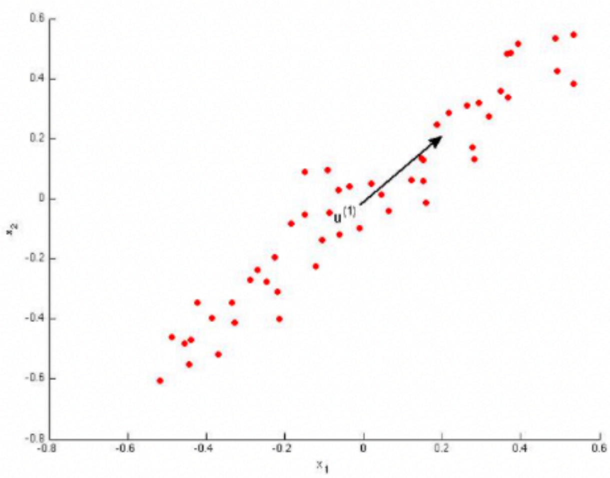
1. Consider the following 2D dataset:

1点

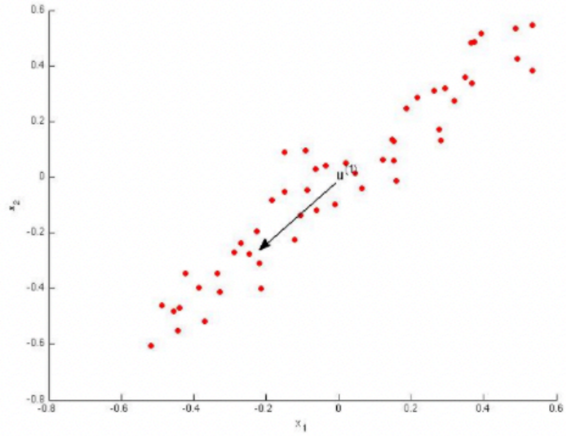


Which of the following figures correspond to possible values that PCA may return for $u^{(1)}$ (the first eigenvector / first principal component)? Check all that apply (you may have to check more than one figure).

- [x]

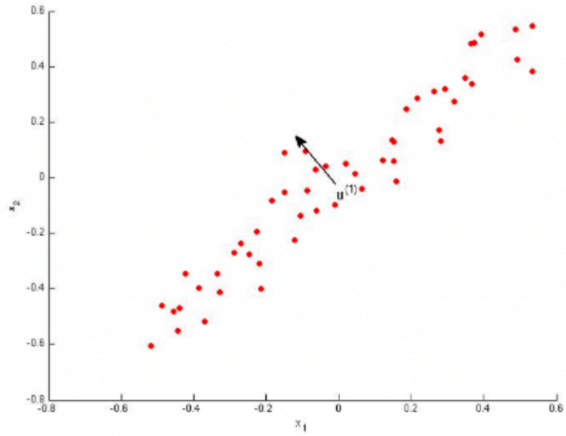


- [x]



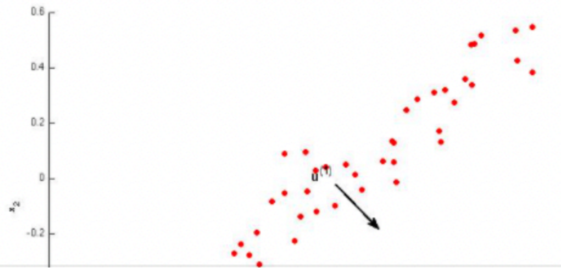
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
x_1
x_2
$u^{(1)}$

- [ ]



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
$u^{(1)}$

- [ ]



0.6
0.4
0.2
0
-0.2
x_2
$u^{(1)}$

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.)

- ( ) Use the elbow method.
- ( ) Choose k to be the largest value so that at least 99% of the variance is retained
- ( ) Choose k to be 99% of m (i.e., $k = 0.99 * m$, rounded to the nearest integer).
- (•) Choose k to be the smallest value so that at least 99% of the variance is retained.

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?

- ( ) $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2} \geq 0.95$
- (•) $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \leq 0.05$
- ( ) $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \geq 0.95$
- ( ) $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \geq 0.05$

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

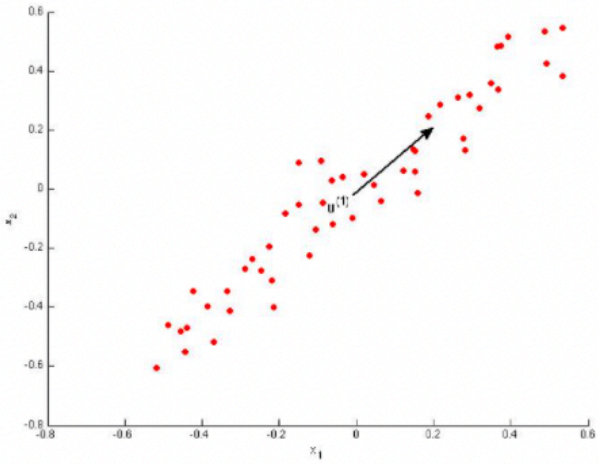
- [ ] Given only $z^{(i)}$ and U_{reduce}, there is no way to reconstruct any reasonable approximation to $x^{(i)}$.
- [x] Even if all the input features are on very similar scales, we should still perform mean normalization (so that each feature has zero mean) before running PCA.
- [ ] PCA is susceptible to local optima; trying multiple random initializations may help.
- [x] Given input data $x \in \mathbb{R}^n$, 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.)

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

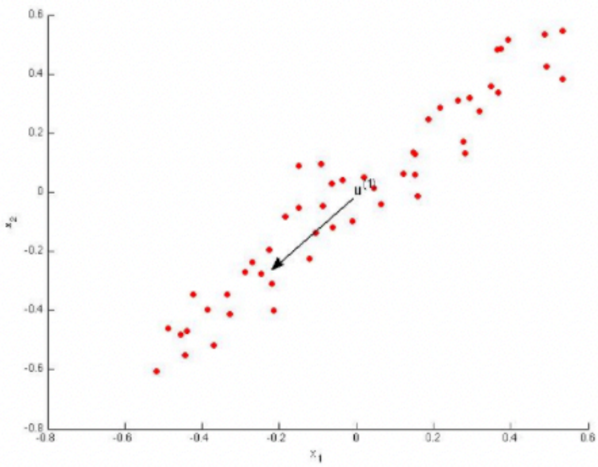
1点

- [ ] 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).
- [x] Data visualization: To take 2D data, and find a different way of plotting it in 2D (using k=2).
- [x] Data compression: Reduce the dimension of your data, so that it takes up less memory / disk space.
- [ ] As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.

—
Answer



正解
The maximal variance is along the y = x line, so this option is correct.



正解
The maximal variance is along the y = x line, so the negative vector along that line is correct for the first principal component.

2. Which of the following is a reasonable way to select the number of principal components k? 1 / 1点

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

- ◯ Use the elbow method.
- ◯ Choose k to be the largest value so that at least 99% of the variance is retained
- ◯ Choose k to be 99% of m (i.e., $k = 0.99 * m$, rounded to the nearest integer).
- ◉ Choose k to be the smallest value so that at least 99% of the variance is retained.

正解
This is correct, as it maintains the structure of the data while maximally reducing its dimension.

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? 1 / 1点

- ◯ $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2} \geq 0.95$
- ◉ $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \leq 0.05$
- ◯ $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \geq 0.95$
- ◯ $\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{\text{approx}}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \geq 0.05$

正解
This is the correct formula.

4. Which of the following statements are true? Check all that apply. 1 / 1点

- ☐ 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.

正解
If you do not perform mean normalization, PCA will rotate the data in a possibly undesired way.

- ☐ PCA is susceptible to local optima; trying multiple random initializations may help.
- ☑ Given input data $x \in \mathbb{R}^n$, 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.)

正解
The reasoning given is correct: with $k = n$, there is no compression, so PCA has no use.

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

- [ ] 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).

- [x] Data visualization: To take 2D data, and find a different way of plotting it in 2D (using k=2).

これを選択しないでください
You should use PCA to visualize data with dimension higher than 3, not data that you can already visualize.

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

正解
If memory or disk space is limited, PCA allows you to save space in exchange for losing a little of the data's information. This can be a reasonable tradeoff.

- [ ] As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.