Writing exercise 5 in machine learning

1. Describe the three steps of anormally detection with Gaussian distribution.
2. Given the following dataset

|  |  |
| --- | --- |
| x1 | x2 |
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

* 1. Plot the dataset in a two dimentional space.
  2. Perform the mean normalization for the dataset.
  3. Calculate the covariance matrix
  4. Calculate the eigenvectors and eigenvalues of the covariance matrix
  5. Calcualte the percentage of variance that has been kept if we only keep the first principal component for all data examples.
  6. Calculate the first components for all data examples.

1. Collaborative filtering methods build a model based on users past behavior (items previously purchased, movies viewed and rated, etc) and use decisions made by current and other users. This model is then used to predict items (or ratings for items) that the user may be interested in.

Content-based filtering methods use features of an item to recommend additional items with similar properties. These approaches are often combined in Hybrid Recommender Systems.

In the following example, which one is content-based method and which one is collaborative filtering method? Explain why.

1. Last.fm creates a "station" of recommended songs by observing what bands and individual tracks the user has listened to on a regular basis and comparing those against the listening behavior of other users. Last.fm will play tracks that do not appear in the user's library, but are often played by other users with similar interests.
2. Pandora uses the properties of a song or artist (a subset of the 400 attributes provided by the Music Genome Project) in order to seed a "station" that plays music with similar properties. User feedback is used to refine the station's results, deemphasizing certain attributes when a user "dislikes" a particular song and emphasizing other attributes when a user "likes" a song.
3. At the first step of collaborative learning is that “Initialize  to small random values.” Please explain why we had better to start from small random values.
4. Suppose you have a PhotoOCR system, where you have the following pipeline:



Each module is treated as a classification task in the pipeline. Please describe how we convert the each module as a classification task.