Universiteti i Prishtinës "Hasan Prishtina" Kosovë

Hyrje në shkencën e të dhënave

Pjesa 10 – Mësimi jo i mbikqyrur (Unsupervised Learning)

Prof. Asoc. Dr. Ermir Rogova

What is unsupervised learning?

- Organize or describe data when no labels are available.
- A major area: clustering

Agglomerative clustering

- Use any computable cluster similarity measure sim(Ci, Cj) e.g., Euclidean distance, cosine similarity etc.
- For n objects v1,..., vn, assign each to a singleton cluster Ci = {vi}
- Repeat {
    - - identify two most similar clusters Cj and Ck (could be ties-chose one pair)
    - - delete Cj and Ck and add (Cj U Ck) to the set of clusters.
    - } until just one cluster.
- Dendrograms diagram the sequence of cluster merges.

Divisive clustering

- Put all objects in one cluster
- Repeat until all clusters are singletons {
- - choose a cluster to split based on some criterion.
- - replace the chosen cluster with sub-clusters.
- }

K-means algorithm

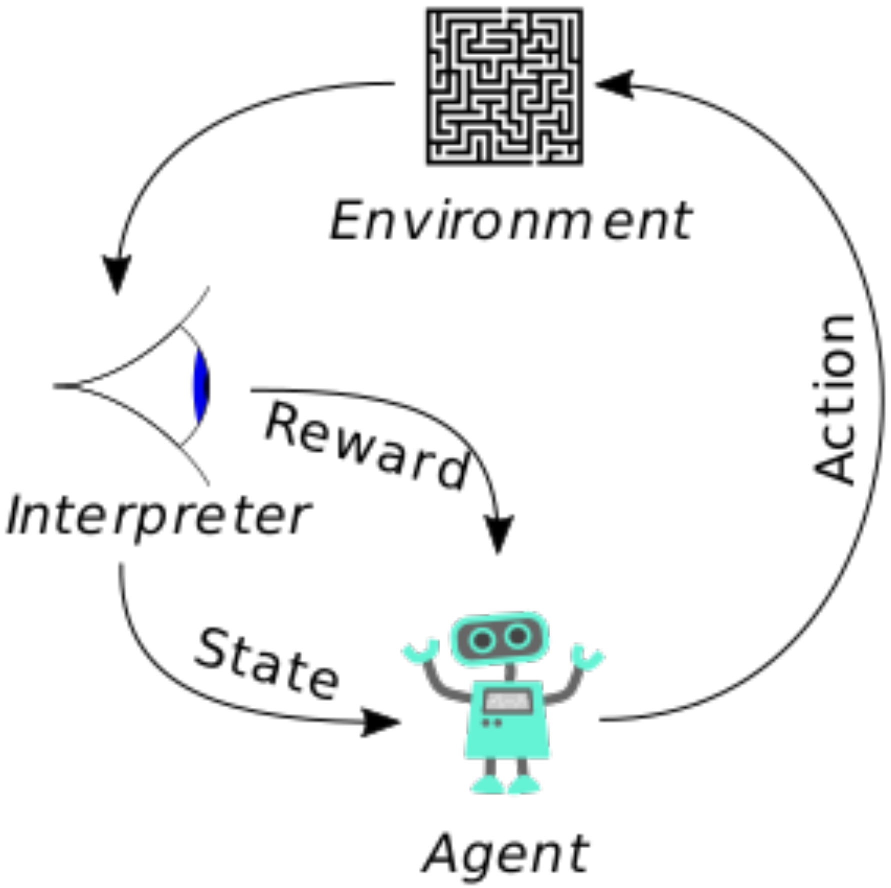
- Begin with a decision on the value of K = number of clusters.
- Put any initial partition that classifies the data into K clusters. You may assign the training samples randomly or systematically.
- Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.
- Repeat the above three steps until convergence is achieved.

Expectation Maximization (EM)

- Enables parameter estimation (learning) in probabilistic models with incomplete data.
- Uses Maximum Likelihood Estimation (MLE).
- MLE is a way to assess the quality of a statistical model based on the probability that model assigns to the observed data. The model that has the highest probability of generating the data is the best one.
- EM algorithm is used to find (locally) MLE parameters of a statistical model in cases where the equations cannot be solved directly.

Universiteti i Prishtinës "Hasan Prishtina" Kosovë

Introduction to reinforcement learning



Branch of machine learning that attempts to model how **agents** should take **actions** in an **environment** that will maximize some form of cumulative **reward**.

Summary

- Unsupervised learning methods help with data exploration and identifying useful patterns in the data.
- Clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense.
- Clustering approaches:
  - Agglomerative
  - Divisive

Pyetje???