EEL-5840 / EEL-4930 Elements of Machine Intelligence

Clustering

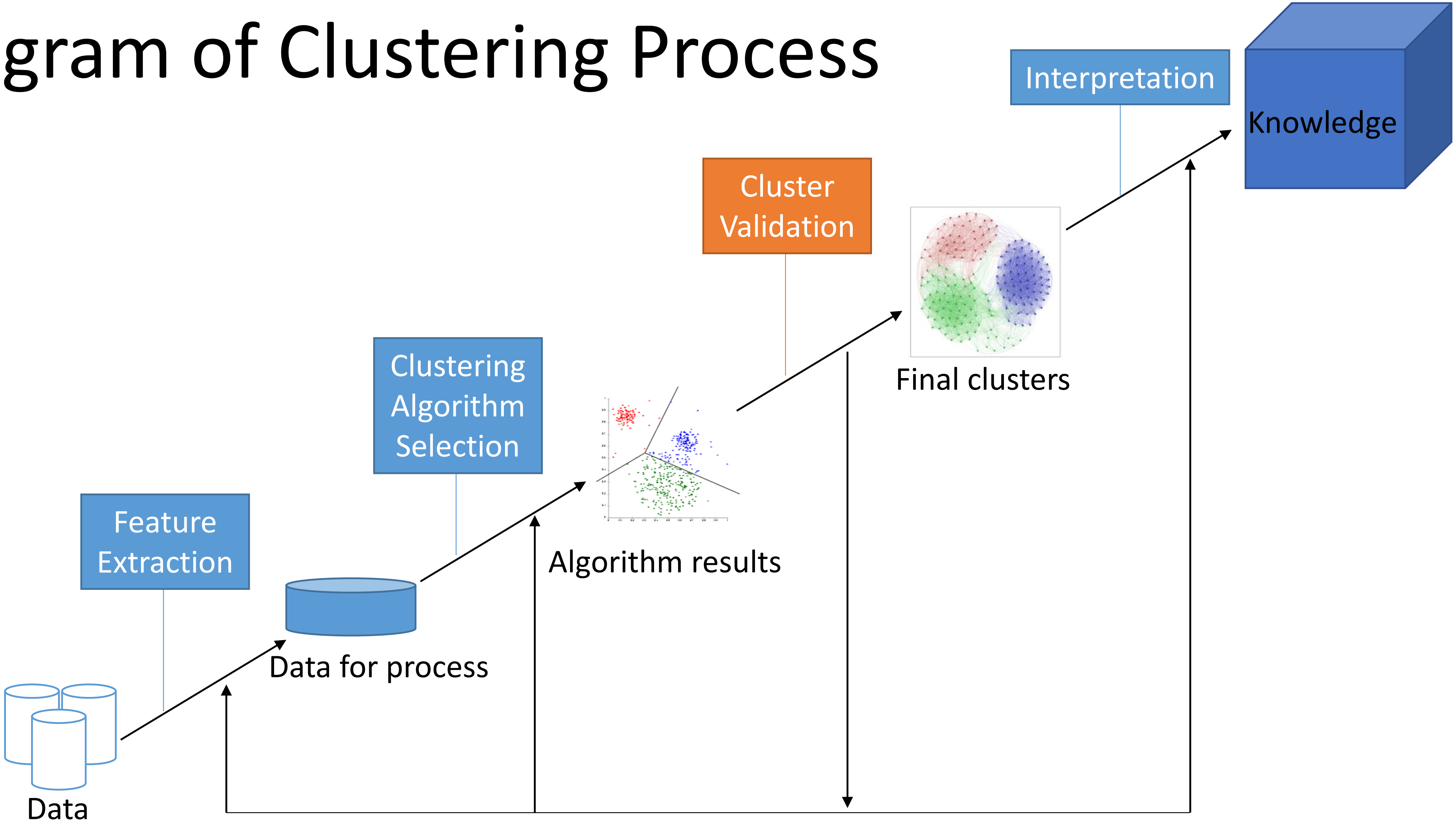
Unsupervised Clustering

- In the clustering process, there are **no predefined classes** and **no examples** that would show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process.
- On the other hand, **classification** is a procedure of **assigning** a data item to a predefined set of categories.
- **Clustering** produces **initial categories** in which values of a data set are classified during the classification process.

Unsupervised Clustering

- The clustering process may result in **different partitioning** of a data set, depending on the **specific criterion** used for clustering. Thus, there is a need of preprocessing before we assume a clustering task in a data set.
- The basic steps to develop clustering process are presented in the following diagram

Diagram of Clustering Process



Feature Extraction

Clustering Algorithm Selection

Cluster Validation

Interpretation

Knowledge

Data

Data for process

Algorithm results

Final clusters

Basic Steps

i. **Feature Selection:** The goal is to select properly the features on which clustering is to be performed so as to encode as much information as possible concerning the task of our interest. Thus, preprocessing of data may be necessary prior to their utilization in clustering task.

ii. **Clustering Algorithm:** This step refers to the choice of an algorithm that results in the definition of a good clustering scheme for a data set. A proximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set.

Basic Steps

iii. **Validation of the Results:** The correctness of clustering algorithm results is verified using appropriate criteria and techniques. Since clustering algorithms define clusters that are not known a priori, irrespective of the clustering methods, the final partition of data requires some kind of evaluation in most applications.

iv. **Interpretation of the results:** In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.

Clustering Applications

- Data Reduction
- Hypothesis generation
- Hypothesis testing
- Prediction based on groups
- Business
- Biology
- Spatial data analysis
- Web mining
- Biometrics

Issues for Clustering

- Representation
  - Vector space?
  - Normalization?
- Notion of similarity/distance
- How many clusters
  - Fixed a priori?
  - Completely data driven?

Interval-valued variables

- Standardize data
  - Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf}).$$

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

where *i* = (x_{i1}, x_{i2}, …, x_{ip}) and *j* = (x_{j1}, x_{j2}, …, x_{jp}) are two *p*-dimensional data objects, and *q* is a positive integer

- If *q = 1, d* is Manhattan distance

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

Similarity and Dissimilarity Between Objects (Cont.)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - *d(i,j)* $\geq$ 0
    - *d(i,i)* = 0
    - *d(i,j)* = *d(j,i)*
    - *d(i,j)* $\leq$ *d(i,k)* + *d(k,j)*
- Also one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures.

Clustering Algorithms

- Flat algorithms
    - Usually start with a random (partial) partitioning
    - Refine it iteratively
        - *K* means clustering
        - (Model based clustering)
- Hierarchical algorithms
    - Bottom-up, agglomerative
    - (Top-down, divisive)

Partitioning Algorithms

- Partitioning method: Construct a partition of *n* vectors into a set of *K* clusters

- Given: a set of vectors and the number *K*

- Find: a partition of *K* clusters that optimizes the chosen partitioning criterion

K-Means

- Assumes objects are real-valued vectors.
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, *c*:

$$\vec{\mu}(\mathrm{c}) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.
    - (Or one can equivalently phrase it in terms of similarities)

K-Means Algorithm

```
Select $K$ random vectors $\{s_1, s_2,\dots s_K\}$ as seeds.
Until clustering converges (or other stopping criterion):
    For each vector $d_i$:
        Assign $d_i$ to the cluster $c_j$ such that $dist(x_i, s_j)$ is minimal.
    (Next, update the seeds to the centroid of each cluster)
    For each cluster $c_j$
        $s_j = \mu(c_j)$
```

Termination conditions

Several possibilities, e.g.,

- A fixed number of iterations

- Vector partition unchanged

- Centroid positions don’t change

Convergence

- Why should the *K*-means algorithm ever reach a *fixed point*?
  - A state in which clusters don’t change.

- *K*-means is a special case of a general procedure known as the *Expectation Maximization (EM) algorithm*.
  - EM is known to converge.
  - Number of iterations could be large.

    But in practice usually isn’t

Convergence of *K*-Means

- Define goodness measure of cluster *k* as sum of squared distances from cluster centroid:
  - $G = \Sigma_k G_k$
  - $G_k = \Sigma_i (d_i - c_k)^2$ (sum over all d_i in cluster *k*)
- Reassignment monotonically decreases G since each vector is assigned to the closest centroid.
- K-Means typically converges quickly

Time Complexity

- Computing distance between two docs is O*(M)* where *M* is the dimensionality of the vectors.
- Reassigning clusters: O*(KN)* distance computations, or O*(KNM).*
- Computing centroids: Each vector gets added once to some centroid: O*(NM).*
- Assume these two steps are each done once for *I* iterations: O*(IKNM).*

Seed Choice

- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
  - Select good seeds using a heuristic: Try out multiple starting points
  - Initialize with the results of another method.

K-means issues, variations, etc.

- Recomputing the centroid after every assignment (rather than after all points are re-assigned) can improve speed of convergence of *K*-means
- Assumes clusters are spherical in vector space
    - Sensitive to coordinate changes, weighting etc.
- Disjoint and exhaustive
    - Doesn’t have a notion of “outliers” by default
    - But can add outlier filtering

How Many Clusters?

- Number of clusters *K* is given
  - Partition *n* docs into predetermined number of clusters
- Finding the “right” number of clusters is part of the problem
- Can usually take an algorithm for one flavor and convert to the other.

K not specified in advance

- Solve an optimization problem: penalize having lots of clusters

- Tradeoff between having more clusters (better focus within each cluster) and having too many clusters

K-means: summary

- Algorithmically, very simple to implement

- *K*-means converges, but it finds a local minimum of the cost function

- Works only for numerical observations

- Outliers can considerable trouble to *K*-means

Quote

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes

Optimal Clustering Scheme

As the goal of clustering is to make objects within the same cluster similar and objects in different clusters distinct, cluster validity measures are defined by combining compactness and separability.

- **Compactness (or intra-distance or within-cluster scatter)**: The members of each cluster should be as close to each other as possible. A common measure of compactness is the variance, which should be minimized.
- **Separation (or inter-distance or between-cluster scatter)**: This indicates how distinct two clusters are. It computes the distance between two different clusters. There are three common approaches measuring the distance between two different clusters.
  - Single linkage: It measures the distance between the closest members of the clusters.
  - Complete linkage: It measures the distance between the most distant members.
  - Comparison of centroids: It measures the distance between the centers of the clusters.

Validation of Clusters – Fundamental Concepts

In general terms, there are three *index criteria* to investigate cluster validity:

1. **Internal criteria**. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g. proximity matrix).
2. **External criteria**. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set.
3. **Relative criteria**. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values. In practice, relative criteria are a combination on internal and external criteria.

Internal Validity Criteria – BIC Index

- **Bayesian Information Criterion (BIC) Index** which basically determines which probability-based mixture is the most appropriate, is based on Bayes Theorem. This index is developed to remove the problem of overfitting. It is defined as follows:

$$\boldsymbol{BIC = -ln\left(\widehat{L}\right) + v\ ln(n)}$$

Where

$\boldsymbol{\widehat{L}}$ is the maximized value of the likelihood function of the model M, i.e., $\boldsymbol{\widehat{L} = p(x|\widehat{\theta}, M)}$, where $\boldsymbol{\widehat{\theta}}$ are the parameter values that maximize the likelihood function;

$\boldsymbol{x}$ is the observed data

$\boldsymbol{\theta}$ is the parameters of the model (e.g., mean and standard deviation in the case of a Gaussian model)

$\boldsymbol{n}$ is the number of data points in x, the number of observations, or equivalently, the sample size;

$\boldsymbol{v}$ is the number of free parameters to be estimated. If the model under consideration is a linear regression, v is the number of regressors.

- Thus, the BIC index is devised in such a way that it takes into account both the fit of the model to the data and the complexity of the model. Smaller value of BIC indicates a better model.

Internal Validity Criteria – Calinski-Harabasz Index

- The **Calinski-Harabasz (CH) index** is computed by

$$CH = \frac{trace(S_B)}{trace(S_W)} \frac{n_p - 1}{n_p - K}$$

Where

$\boldsymbol{S_B}$ is the between-cluster scatter matrix (inter-distance)

$\boldsymbol{S_W}$ is the internal scatter matrix (intra-distance)

$\boldsymbol{n_p}$ is the number of clustered samples

$\boldsymbol{K}$ is the number of clusters

- Higher value of the CH-index indicates a better partitioning.

Internal Validity Criteria – Silhouette Width Criterion Index

- The **Silhoute Width Criterion (SWC) index** is a cluster validity index that is used to judge the quality of any clustering solution. Suppose a represents the average distance of a point from the other points of the cluster to which is assigned, and b represents the minimum of the average distances of the point from the points of the other clusters. Now the silhouette width s of the point is defined as

$$s = \frac{b - a}{\max\{a, b\}}$$

- Silhouette index is the average silhouette width of all data points and it reflects the compactness and separation of clusters.
- The value of silhouette index varies from -1 and 1 and higher indicates better clustering result.

Internal Validity Criteria – Davies-Bouldin Index

- The **Davies-Bouldin (DB) index** is a function of the ration of the sum of *within-cluster scatter* to *between-cluster separation*. The scatter within the i th cluster is computed as

$$S_{i,q} = \left(\frac{\sum_{\bar{x} \in X_i} |\bar{x} - \overline{c_i}|^q}{|X_i|} \right)^{\frac{1}{q}}$$

And the distance between cluster X_i and X_j is defined as

$$d_{i,j,t} = \left\| \overline{c_i} - \overline{c_j} \right\|_t$$

$S_{i,q}$ is the q th moment of the $|X_i|$ points in cluster X_i with respect to their mean $\overline{c_i}$ and is a measure of the dispersion of the points in the cluster.

Internal Validity Criteria – Davies-Bouldin Index (cont.)

Generally, $S_{i,1}$ is used, which is the average Euclidean distance of the vectors in class i to the centroid of class i. $d_{i,j,t}$ is the Minkowski distance of order t between the centroids $\overline{c_i}$ and $\overline{c_j}$ that characterize clusters X_i and X_j. Subsequently, one computes

$$R_{i,q,t} = \max_{j,j \neq i} \frac{S_{i,q} + S_{j,q}}{d_{i,j,t}}$$

The **Davies-Bouldin (DB) index** is then defined as

$$DB = \frac{1}{K} \sum_{i=1}^{K} R_{i,q,t}$$

Here K denotes the number of clusters. The objective is to minimize the DB index for achieving the proper clustering.

Internal Validity Criteria – Dunn's Index

- Let S and T be two non-empty subsets of $\mathbb{R}^N$. Then the diameter Δ of S and set distance δ between S and T are

$$\Delta(S) = \max_{\bar{x},\bar{y} \in S} d(\bar{x}, \bar{y})$$

And

$$\delta(S,T) = \min_{\bar{x} \in S, \bar{y} \in T} d(\bar{x}, \bar{y})$$

Where $d(\bar{x}, \bar{y})$ is the distance between points $\bar{x}$ and $\bar{y}$. For any partition, Dunn defined the following index:

$$Dunn = \min_{1 \le i \le K} \left\{ \min_{1 \le j \le K, j \ne i} \left\{ \frac{\delta\left(C_i, C_j\right)}{\max_{1 \le k \le K} \Delta(C_k)} \right\} \right\}$$

Larger values of Dunn correspond to good clusters, and the number of clusters that maximizes the Dunn index is taken as the optimal number of clusters.

Internal Validity Criteria – Xie-Beni Index

- The **Xie-Beni (XB) index** focuses on two properties: compactness and separation. According to the definition of the XB-index, the numerator measures the compactness of the obtained fuzzy partition while the denominator measures the separation between clusters. In general, a good partition has a small value for the compactness, and well-separated centers will produce a high value for the separation. Hence, the most desirable partition is obrained by minimizing the XB-index for $k = 1, \ldots, K_{max}$.

$$XB = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n} \mu_{ij}^{2} \left\| \overline{x_j} - \overline{c_i} \right\|^{2}}{n \left(\min_{i \neq k} \| \overline{c_i} - \overline{c_k} \|^{2} \right)}$$

Internal Validity Criteria – I-Index

- The **I-index** is defined as follows:

$$I = \left(\frac{1}{K}\frac{E_1}{E_k}D_k\right)^2$$

Where E_1 is constant for a given data set. Here, $E_k = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\|\overline{x_i} - \overline{c_k}\|^2$ and $D_k = \max_{i,j=1}^{K}\left\|\overline{c_i} - \overline{c_j}\right\|^2$, where n_i is the number of points in cluster C_i and $\overline{c_k}$ is the center of the k th cluster.

The best partitioning occurs at the maximum value of the function.

Internal Validity Criteria – CS-Index

- The **CS-index** is defined as follows:

$$CS = \frac{\sum_{i=1}^{K}\left[\frac{1}{N_i}\sum_{\overline{x_i}\in X_i}\max_{\overline{x_q}\in X_i} d(\overline{x_i},\overline{x_q})\right]}{\sum_{i=1}^{K}\left[\min_{j\in K, j\neq i} d(\overline{c_i},\overline{c_q})\right]}$$

Where $c_i, i = 1, \dots, K$ are the cluster centers. The CS-index is more efficient in tackling clusters of different densities and/or sizes than the other popular validity measures, the price being paid in terms of high computational load with increasing K and n.

External Validity Criteria – F-measure

- The **F-measure index** combines the precision and recall concepts from information retrieval.

$$\text{Recall}(i,j)=\frac{n_{ij}}{n_i}$$

And

$$\text{Precision}(i,j)=\frac{n_{ij}}{n_j}$$

Where n_{ij} is the number of objects of class i that are in cluster j, n_j is the number of objects in cluster j, and n_i, is the number of objects in class i. The **F-measure** of cluster j and class i is given by

$$F(i,j)=\frac{2\ \text{Recall}(i,j)\ \text{Precision}(i,j)}{\text{Precision}(i,j)+\text{Recall}(i,j)}$$

The F-measure values are within the interval $[0,1]$ and larger values indicate higher clustering quality.

External Validity Criteria – Entropy

- The **Entropy** measures the putiry of the clusters class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy increases. To compute the entropy of a dataset, we need to calculate the class distribution of the objects in each cluster as follows

$$H(X_j) = \sum_i p_{ij} \log p_{ij}$$

Where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the weighted sum of the entropies of all clusters, that is,

$$H\left(\{X_j\}_{j=1}^{K}\right) = \sum_{j=1}^{K} \frac{n_j}{n} H(X_j)$$

Where n_j is the size of cluster j, K is the number of clusters, and n is the total number of data points.

External Validity Criteria – Purity

- **Purity** is very similar to entropy, in the sense that, with both measures, we want to calculate the putiry of a set of clusters.
- For each cluster, the purity

$$P_j = \frac{1}{n_j} \max_i n_j^i$$

Is the number of objects in j with class label i. In other words, P_j is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual and given as

$$\text{Purity} = \sum_{j=1}^{K} \frac{n_j}{n} P_j$$

Where n_j is the size of cluster j, K is the number of clusters, and n is the total number of objects.

External Validity Criteria – NMI Measure

- The **Normalized Mutual Information(NMI) measure** of two labeled objects can be measured as

$$\mathrm{NMI}(\mathrm{X},\mathrm{Y}) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$$

Where $I(X,Y)$ denotes the mutual information between two random variables X and Y and $\mathrm{H}(X)$ denotes the entropy of X, X will be consensus clustering while Y will be true labels.

Questions?