Part II: Web Content Mining
Chapter 4: Evaluating Clustering

- Approaches to Evaluating Clustering
- Similarity-Based Criterion Functions
- Probabilistic Criterion Functions
- MDL-Based Model and Feature Evaluation
- Classes to Clusters Evaluation
- Precision, Recall and F-measure
- Entropy

Approaches to Evaluating Clustering

- Criterion functions evaluate clustering models *objectively*, i.e. using only the document content.
  - Similarity-based functions: *intracluster similarity* and *sum of squared errors*
  - Probabilistic functions: *log-likelihood* and *category utility*
  - MDL-based evaluation
- Document labels (if available) may also be used for evaluation of clustering models
  - If labeling is correct and reflects accurately the document content we can evaluate the quality of clustering.
  - If clustering reflects accurately the document content we can evaluate the quality of labeling.
  - Criterion function based on labeled data: *classes to clusters evaluation*

Similarity-Based Criterion Functions (distance)

- Basic idea: the cluster center m_i (*centroid* or *mean* in case of numeric data) best represents cluster D_i if it minimizes the sum of the lengths of the "error" vectors $x - m_i$ for all $x \in D_i$.

$$J_e = \sum_{i=1}^{k} \sum_{x \in D_i} \left\| x - m_i \right\|^2$$

$$m_i = \frac{1}{\left| D_i \right|} \sum_{x \in D_i} x$$

- Alternative formulation based on *pairwise distance* between cluster members

$$J_e = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{\left| D_i \right|} \sum_{x_j, x_k \in D_i} \left\| x_j - x_k \right\|^2$$

Similarity-Based Criterion Functions (cosine similarity)

- For document clustering the *cosine similarity* is used

$$J_s = \sum_{i=1}^{k} \sum_{d_j \in D_i} sim(c_i, d_j)$$

$$sim(c_i, d_j) = \frac{c_i \bullet d_j}{\|c_i\| \|d_j\|} \qquad c_i = \frac{1}{|D_i|} \sum_{d_j \in D_i} d_j$$

- Equivalent form based on pairwise similarity

$$J_s = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{|D_i|} \sum_{d_j, d_k \in D_i} sim(d_j, d_k)$$

- Another formulation based on *intracluster similarity* (used to controls merging of clusters in hierarchical agglomerative clustering)

$$J_s = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{|D_i|} \sum_{d_j, d_k \in D_i} sim(d_j, d_k) = \frac{1}{2} \sum_{i=1}^{k} |D_i| \, sim(D_i)$$

Similarity-Based Criterion Functions (example)

Sum of centroid similarity evaluation of four clusterings

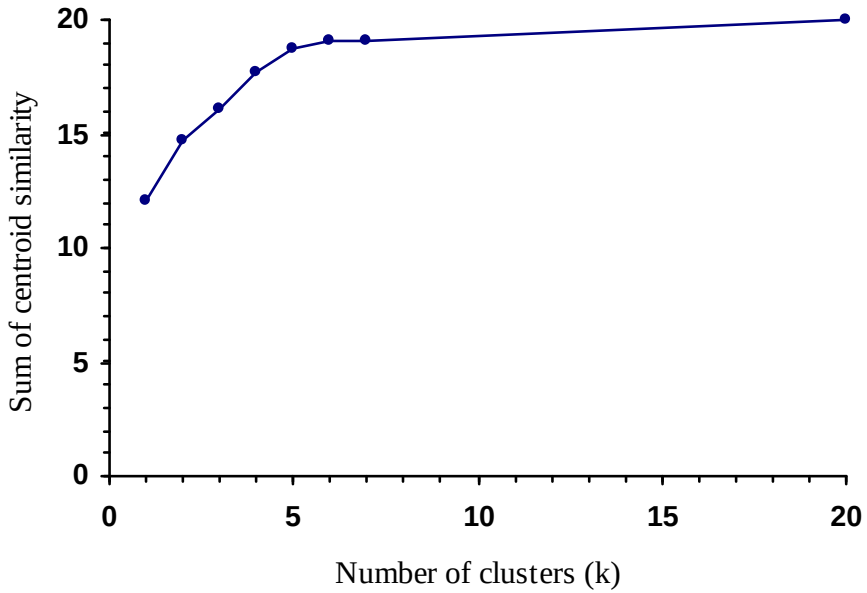
Agglomerative	k-means (k=2)	k-means (k=3)	k-means (k=4)
1 [12.0253] **2 [9.43932]** **3 [5.64819]** **4 [4.6522]** **5 [3.8742]** **6 [2.95322]** **7 [1.99773]** Chemistry Computer Political Geography Anthropology **8 [1.98347]** Justice Theatre **9 [5.44416]** **10 [2.81679]** **11 [1.97333]** Psychology Sociology Math **12 [2.90383]** **13 [1.96187]** Biology Economics Physics **14 [5.40061]** **15 [2.83806]** **16 [1.98066]** History Music Philosophy **17 [3.81771]** **18 [2.97634]** **19 [1.99175]** English Languages Art Communication	**1 [12.0253]** **2 [8.53381]** Anthropology Biology Chemistry Computer Economics Geography Math Physics Political Psychology Sociology **3 [6.12743]** Art Communication Justice English History Languages Music Philosophy Theatre	**1 [12.0253]** **2 [2.83806]** History Music Philosophy **3 [6.09107]** Anthropology Biology Chemistry Computer Geography Math Political **4 [7.12119]** Art Communication Justice Economics English Languages Physics Psychology Sociology Theatre	**1 [12.0253]** **2 [3.81771]** Art Communication English Languages **3 [5.44416]** Biology Economics Math Physics Psychology Sociology **4 [2.83806]** History Music Philosophy **5 [5.64819]** Anthropology Chemistry Computer Justice Geography Political Theatre
14.83993 (clusters 2+14)	**14.6612**	**16.05032**	**17.74812**

Finding the Number of Clusters

Agglomerative clustering

Partitioning	Sum of centroid similarity
2,14	14.8399
3,9,14	16.493
2,15,17	16.0951
4,8,9,14	17.4804
3,9,15,17	17.7481
4,8,9,15,17	18.7356
3,10,12,15,17	18.0246
4,8,10,12,14	17.7569
4,8,10,12,15,17	19.0121

K-means clustering



Probabilistic Criterion Functions

Document is a *random event*

- Probability of document

$$P(d) = \sum_A P(d \mid A)P(A)$$

- Probability of sample (assuming that documents are independent events)

$$P(d_1, d_2, ..., d_n) = \prod_{i=1}^{n} \sum_A P(d_i \mid A)P(A)$$

- Log-likelihood (log of probability of sample):

$$L = \sum_{i=1}^{n} \log \sum_A P(d_i \mid A)P(A)$$

- Category Utility (based on probability of attributes)

$$\sum_k \sum_i \sum_j P(a_j = v_{ij} \mid C_k)\, P(C_k \mid a_j = v_{ij})\, P(a_j = v_{ij})$$

Category Utility (basic idea)

$$\sum_k \sum_i \sum_j P(a_j = v_{ij} \mid C_k) P(C_k \mid a_j = v_{ij}) P(a_j = v_{ij})$$

- $P(a_j = v_{ij} \mid C_k)$ is the probability that an object has value v_{ij} for its attribute a_j, given that it belongs to category C_k. The higher this probability, the more likely two objects in a category share the same attribute values.

- $P(C_k \mid a_j = v_{ij})$ is the probability that an object belongs to category C_k, given that it has value v_{ij} for its attribute a_j. The greater this probability, the less likely objects from different categories have attribute values in common.

- $P(a_j = v_{ij})$ is a weight coefficient assuring that frequent attribute values have stronger influence on the evaluation.

Category Utility Function

$$CU = \sum_k \sum_i \sum_j P(a_j = v_{ij} \mid C_k)\, P(C_k \mid a_j = v_{ij})\, P(a_j = v_{ij})$$

$$P(C_k \mid a_j = v_{ij}) = \frac{P(a_j = v_{ij} \mid C_k \mid)\, P(C_k)}{P(a_j = v_{ij})}$$ Bayes rule

$$CU = \sum_k P(C_k) \sum_i \sum_j P(a_j = v_{ij} \mid C_k)^2$$

$\sum_i \sum_j P(a_j = v_{ij} \mid C_k)^2$ Expected number of attribute values of a member of C_k correctly guessed using a *probability matching strategy*

$\sum_i \sum_j P(a_j = v_{ij})^2$ Expected number of correctly guessed attribute values without knowing the categories (clusters) in the sample

$$CU(C_1, C_2, ..., C_n) = \frac{1}{n} \sum_k P(C_k) \sum_i \sum_j \left(P(a_j = v_{ij} \mid C_k)^2 - P(a_j = v_{ij})^2 \right)$$

$$CU(C_1, C_2, ..., C_n) = \frac{1}{n} \sum_{k=1}^{n} P(C_k) \frac{1}{2\sqrt{\pi}} \sum_i \left(\frac{1}{\sigma_{ik}} - \frac{1}{\sigma_i} \right)$$ (normal distribution)

Category Utility (example)

$A = \{1,3,4,6,8,10,12,16,17,18,19\}$ $B = \{2,5,7,9,11,13,14,15,20\}$

Document	No.	history	science	research	offers	students	hall
		a_1	a_2	a_3	a_4	a_5	a_6
Anthropology	1	0	1	1	0	1	1
Art	2	0	0	0	1	1	1
Biology	3	0	1	1	0	1	1
Chemistry	4	0	1	0	0	1	1
Communication	5	0	0	0	1	1	0
Computer	6	0	1	0	0	1	1
Justice	7	0	0	0	0	1	0
Economics	8	0	0	1	0	0	0
English	9	0	0	0	1	0	1
Geography	10	0	1	0	0	1	0
History	11	1	0	0	1	0	0
Math	12	0	1	1	1	1	1
Languages	13	0	0	0	1	0	1
Music	14	1	0	0	0	1	1
Philosophy	15	1	0	0	1	0	1
Physics	16	0	0	1	0	1	1
Political	17	0	1	1	0	1	1
Psychology	18	0	0	1	1	1	1
Sociology	19	0	0	1	1	1	1
Theatre	20	0	0	0	0	1	1

$$CU(A,B) = \frac{1}{2}\left(CU(A) + CU(B)\right)$$

$$CU(A) = P(A)\left(\sum_{j=1}^{6}\left(P(a_j = 0 \mid A)^2 - P(a_j = 0)^2\right) + \sum_{j=1}^{6}\left(P(a_j = 1 \mid A)^2 - P(a_j = 1)^2\right)\right)$$

$$CU(B) = P(B)\left(\sum_{j=1}^{6}\left(P(a_j = 0 \mid B)^2 - P(a_j = 0)^2\right) + \sum_{j=1}^{6}\left(P(a_j = 1 \mid B)^2 - P(a_j = 1)^2\right)\right)$$

$$P(A) = 11/20 = 0.55 \quad P(B) = 9/20 = 0.45$$
$$P(a_1 = 0) = 17/20 = 0.85$$
$$P(a_1 = 1) = 3/20 = 0.15$$
$$P(a_1 = 0 \mid A) = 1$$
$$P(a_1 = 0 \mid B) = 6/9 = 0.667$$
$$CU(A) = 4.87295 \quad CU(B) = 1.80028$$
$$CU(A,B) = \frac{1}{2}(4.87295 + 1.80028) = 3.3366$$

Finding Regularities in Data

- Describe the cluster as a pattern of attribute values that repeats in data
- Include attribute values that are the same for all members of the cluster and omit attributes that have different values (*generalization* by *dropping conditions*)
- Hypothesis 1 (using attributes *science* and *research):*

$$H_1 = \begin{cases} R_1: \text{IF (science = 0) AND (research = 1) THEN Class = A} \\ R_2: \text{IF (science = 1) AND (research = 0) THEN Class = A} \\ R_3: \text{IF (science = 1) AND (research = 1) THEN Class = A} \\ R_4: \text{IF (science = 0) AND (research = 0) THEN Class = B} \end{cases}$$

$$A = \{1,3,4,6,8,10,12,16,17,18,19\}$$
$$B = \{2,5,7,9,11,13,14,15,20\}$$

- Hypothesis 2 (using attribute *offers*):

$$H_2 = \begin{cases} R_1: \text{If offers} = 0 \text{ THEN Class} = A \\ R_2: \text{If offers} = 1 \text{ THEN Class} = B \end{cases} \qquad \begin{aligned} A &= \{1, 3, 4, 6, 7, 8, 10, 14, 16, 17, 20\} \\ B &= \{2, 5, 9, 11, 12, 13, 15, 18, 19\} \end{aligned}$$

- Which one is better, H_1 or H_2?

Occam's Razor

- "*Entities are not to be multiplied beyond necessity*" (William of Occam, 14th century)
- *Among several alternatives the simplest one is usually the best choice*.
- H_2 looks simpler (shorter) than H_1, so H_2 may be better than H_1
- How do we measure simplicity?
- Dropping more conditions produces simpler (shorter) hypotheses, the simplest one been the empty rule. The latter however is a single cluster including the whole dataset (overgeneralization).
- The most complex (longest) hypothesis has 20 clusters and is equivalent to the original dataset (overfitting)
- How do we find the right balance?
- The answer to both questions is MDL

Minimum Description Length (MDL)

- Given a data set D (e.g. our document collection) and a set of hypotheses $H_1, H_2, ..., H_n$ each one describing D.
  Find the most likely hypothesis $H_i = \arg\max_i P(H_i | D)$
- Direct estimation of $P(H_i | D)$ is difficult, so we apply Bayes

$$P(H_i | D) = \frac{P(H_i)P(D | H_i)}{P(D)}$$

- Take –log of both sides

$$-\log_2 P(H_i | D) = -\log_2 P(H_i) - \log_2 P(D | H_i) + \log_2 P(D)$$

- Consider hypotheses and data as messages and apply Shannon's information theory, which defines information in a message as a negative logarithm of its probability. Then estimate the number of bits (L) needed to encode the messages.

$$L(H_i | D) = L(H_i) + L(D | H_i) - L(D)$$

Minimum Description Length Principle

- $L(H_i)$ and $L(D)$ are the minimum number of bits needed to encode the hypothesis and data.
- $L(D \mid H_i)$ is the number of bits needed to encode D if we know H.
- If we think of H as a pattern that repeats in D we don't have to encode all its occurrences, rather we encode only the pattern itself and the differences that identify each individual instance in D. Thus the more regularity in data the shorter description length $L(D \mid H_i)$.
- We need of good balance between $L(D \mid H_i)$ and $L(H_i)$ because if H describes the data *exactly* then $L(D \mid H_i) = 0$, but $L(H_i)$ will be large.
- We can exclude $L(D)$ because it does not depend on the choice of hypotheses.
- *Minimum Description Length (MDL) principle:*

$$H_i = \arg\min_i L(H_i) + L(D \mid H_i)$$

MDL-based Model Evaluation (Basics)

- Choose a description language (e.g. rules)
- Use the same encoding scheme for both hypotheses and data given the hypotheses
- Assume that hypotheses and data are uniformly distributed
- Then probability of occurrence of an item out of n alternatives is $1/n$.
- And the minimum code length of the message informing us that a particular item has occurred is $-\log_2 1/n = \log_2 n$

MDL-based Model Evaluation (Example)

- Description language: 12 attribute-value pairs (6 attributes each with two possible values)
- Rule R_1 covers documents 8, 16, 18 and 19
- There are 9 different attribute-value pairs that occur in these documents: {history=0}, {science=0}, {research=1}, {offers=0}, {offers=1}, {students=0}, {students=1}, {hall=0}, {hall=1}.
- Specifying rule R_1 is equivalent to choosing 9 out of 12 attribute-value pairs, which can be done in $\binom{12}{9}$ different ways.
- Thus $\log_2 \binom{12}{9}$ bits are needed to encode the right-hand side of R_1.
- In addition we need one bit (a choice of one out of two cluster labels) to encode the choice of the class
- Thus the code length of R_1 is

$$L(R_1) = \log_2 \binom{12}{9} + 1 = \log_2 220 + 1 = 8.78136$$

MDL-based Model Evaluation ($L(H_1)$)

Similarly we compute the code lengths of R_2, R_3, and R_4 and obtain:

$$L(R_2) = \log_2 \binom{12}{7} + 1 = \log_2 792 + 1 = 10.6294$$

$$L(R_3) = \log_2 \binom{12}{7} + 1 = \log_2 792 + 1 = 10.6294$$

$$L(R_4) = \log_2 \binom{12}{10} + 1 = \log_2 66 + 1 = 7.04439$$

Using the additivity of information to obtain the code length of *L(H)* we add the code lengths of its constituent rules.

$$L(H_1) = 37.0845$$

MDL-based Model Evaluation ($L(D|R_1)$)

- Consider the message exchange setting, where the hypothesis R_1 has already been communicated.
- This means that the recipient of that message already knows the subset of 9 attribute-value pairs selected by rule R_1.
- Then to communicate each document of those covered by R_1 we need to choose 6 (the pairs occurring in each document) out of the 9 pairs.
- This choice will take $\log_2 \binom{9}{6}$ bits to encode.
- As R_1 covers 4 documents (8, 16, 18, 19) the code length needed for all of them will be

$$L(\{8,16,18,19\} \mid R_1) = 4 \times \log_2 \binom{9}{6} = 4 \times \log_2 84 = 25.5693$$

MDL-based Model Evaluation (MDL(H_1))

- Similarly we compute the code length of the subsets of documents covered by the other rules.

$$L(\{4,6,10\} \mid R_2) = 3 \times \log_2 \binom{7}{6} = 3 \times \log_2 7 = 8.4220$$

$$L(\{1,3,12,17\} \mid R_3) = 4 \times \log_2 \binom{7}{6} = 4 \times \log_2 7 = 11.2294$$

$$L(\{2,5,7,9,11,13,14,15,20\} \mid R_4) = 9 \times \log_2 \binom{10}{6} = 9 \times \log_2 210 = 69.4282$$

- The code length needed to communicate all documents given hypothesis H_1 will be the sum of all these code lengths, i.e.

$$L(D \mid H_1) = 114.649$$

- Now adding this to the code length of the hypothesis we obtain

$$MDL(H_1) = L(H_1) + L(D \mid H_1) = 37.0845 + 114.649 = 151.733$$

MDL-based Model Evaluation (H_1 or H_2?)

- Similarly we compute $MDL(H_2)$

$$MDL(H_2) = L(H_2) + L(D \mid H_2) = 9.16992 + 177.035 = 186.205$$

$$MDL(H_1) = L(H_1) + L(D \mid H_1) = 37.0845 + 114.649 = 151.733$$

- $MDL(H_1) < MDL(H_2) \Rightarrow H_1$ is better than H_2

- Also (very intuitively), $L(H_1) > L(H_2)$ and $L(D|H_1) < L(D|H_2)$

- How about the most general and the most specific hypotheses?

Information (Data) Compression

- The most general hypothesis (the empty rule {}) does not restrict the choice of attribute-value pairs, so it selects 12 out of 12 pairs

$$L(\{\}) = \log_2 \binom{12}{12} + 1 = 1$$

$$L(D \mid \{\}) = L(D) = 20 \times \log_2 \binom{12}{6} = 20 \times \log_2 924 = 197.035$$

- The most specific hypothesis has 20 rules – one for each document

$$L(S) = 20 \times (\log_2 \binom{12}{6} + 1) = 20 \times (\log_2 924 + 1) = 217.035$$

- Both {} and S represent extreme cases, undesirable in learning – *overgeneralization* and *overspecialization*
- Good hypotheses should provide smaller MDL than {} and S, or stated otherwise (*principle of Information Compression*)

$$L(H) + L(D \mid H) < L(D) \quad \text{or} \quad H_i = \arg\max_i \left(L(D) - L(H_i) - L(D \mid H_i) \right)$$

MDL-based Feature Evaluation

- An attribute can split the set of documents into subsets, each one including the documents that share the same value of that attribute.
- Consider this split as a clustering and evaluate its MDL score.
- Rank attributes by their MDL score and select attributes that provide the lowest score.
- *MDL(H_1)* was actually the MDL score of attribute *offers* from our 6-attribute document sample.

MDL-based Feature Evaluation (Example)

Attribute	Split		MDL
	Value = 0	Value = 1	
science	2, 5, 7, 8, 9, 11, 13, 14, 15, 16, 18, 19, 20	1, 3, 4, 6, 10, 12, 17	173.185
research	2, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15, 20	1, 3, 8, 12, 16, 17, 18, 19	179.564
students	8, 9, 11, 13, 15	1, 2, 3, 4, 5, 6, 7, 10, 12, 14, 16, 17, 18, 19, 20	182.977
history	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 16, 17, 18, 19, 20	11, 14, 15	183.023
offers	1, 3, 4, 6, 7, 8, 10, 14, 16, 17, 20	2, 5, 9, 11, 12, 13, 15, 18, 19	186.205
hall	5, 7, 8, 10, 11	1, 2, 3, 4, 6, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20	186.205

Classes to Clusters Evaluation

- Assume that the classification of the documents in a sample is known, i.e. each document has a class label.
- Cluster the sample without using the class labels.
- Assign to each cluster the class label of the majority of documents in it.
- Compute the *error* as the proportion of documents with different class and cluster label.
- Or compute the *accuracy* as the proportion of documents with the same class and cluster label.

Classes to Clusters Evaluation (Example)

history		science		research		offers		students		Hall	
14/20		16/20		17/20		14/20		14/20		12/20	
A (11/17)	**B (3/3)**	**B (9/13)**	**A (7/7)**	**B (9/12)**	**A (8/8)**	**A (8/11)**	**B (6/9)**	**B (4/5)**	**A (10/15)**	**B (3/5)**	**A (9/15)**
1-A	11-B	2-B	1-A	2-B	1-A	1-A	2-B	8-A	1-A	5-B	1-A
2-B	14-B	5-B	3-A	4-A	3-A	3-A	5-B	9-B	2-B	7-B	2-B
3-A	15-B	7-B	4-A	5-B	8-A	4-A	9-B	11-B	3-A	8-A	3-A
4-A		8-A	6-A	6-A	12-A	6-A	11-B	13-B	4-A	10-A	4-A
5-B		9-B	10-A	7-B	16-A	7-B	12-A	15-B	5-B	11-B	6-A
6-A		11-B	12-A	9-B	17-A	8-A	13-B		6-A		9-B
7-B		13-B	17-A	10-A	18-A	10-A	15-B		7-B		12-A
8-A		14-B		11-B	19-A	14-B	18-A		10-A		13-B
9-B		15-B		13-B		16-A	19-A		12-A		14-B
10-A		16-A		14-B		17-A			14-B		15-B
12-A		18-A		15-B		20-B			16-A		16-A
13-B		19-A		20-B					17-A		17-A
16-A		20-B							18-A		18-A
17-A									19-A		19-A
18-A									20-B		20-B
19-A											
20-B											

Counting the cost

- In some cases we need to look into the type of the error (usually called error cost).
- For example, in an e-mail filtering system the cost of classifying non-spam as spam is higher than classifying spam as non-spam.
- As the most common clustering and classification problems involve two classes they are usually called *positive* and *negative*.
- The original class labels are referred to as *actual* and those determined by the clustering algorithm are called *predicted*.

Confusion matrix (contingency table)

Actual (classes) \ Predicted (clusters)	Positive	Negative
Positive	*TP*	*FN*
Negative	*FP*	*TN*

TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative)

$$Error = \frac{FP + FN}{TP + FP + TN + FN} \qquad Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision and Recall

Actual (classes) \ Predicted (clusters)	Positive	Negative
Positive	*TP*	*FN*
Negative	*FP*	*TN*

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Attribute *research*		
Actual (classes) \ Predicted (clusters)	A	B
A	8	3
B	0	9

Precision = 1.00
Recall = 0.73

Attribute *hall*		
Actual (classes) \ Predicted (clusters)	A	B
A	9	2
B	6	3

Precision = 0.60
Recall = 0.82

F-Measure

Generalized confusion matrix for *m* classes and *k* clusters

Classes \ Clusters	1	…	*j*	…	*k*
1	n_{11}	…	n_{1j}	…	n_{1k}
…	…	…	…	…	…
i	n_{i1}	…	n_{ij}	…	n_{ik}
…	…	…	…	…	…
m	n_{m1}	…	n_{mj}	…	n_{mk}

Combining precision and recall

$$P(i,j) = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}} \qquad R(i,j) = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}}$$

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)}$$

Evaluating the whole clustering

$$F = \sum_{i=1}^{m} \frac{n_i}{n} \max_{j=1,\ldots,k} F(i,j)$$

$$n_i = \sum_{j=1}^{k} n_{ij}$$

$$n = \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij}$$

F-Measure (Example)

offers	
n_{11}=8	n_{12}=3
n_{21}=3	n_{22}=6

students	
n_{11}=10	n_{12}=1
n_{21}=5	n_{22}=4

Accuracy(offers)=14/20=0.7

Accuracy(students)=14/20=0.7

offers	
$P(1,1)=0.73,\ R(1,1)=0.73$ $F(1,1)=0.73$	$P(1,2)=0.33,\ R(1,2)=0.27$ $F(1,2)=0.30$
$P(2,1)=0.27, R(2,1)=0.33$ $F(2,1)=0.30$	$P(2,2)=0.67, R(2,2)=0.67$ $F(2,2)=0.67$
$F=\frac{11}{20}0.73+\frac{9}{20}0.67=0.70$	

students	
$P(1,1)=0.67,\ R(1,1)=0.91$ $F(1,1)=0.77$	$P(1,2)=0.2,\ R(1,2)=0.09$ $F(1,2)=0.12$
$P(2,1)=0.33, R(2,1)=0.56$ $F(2,1)=0.42$	$P(2,2)=0.8, R(2,2)=0.44$ $F(2,2)=0.57$
$F=\frac{11}{20}0.77+\frac{9}{20}0.57=0.68$	

Entropy

- Consider the class label as a random event and evaluate its probability distribution in each cluster.
- The probability of class *i* in cluster *j* is estimated by the proportion of occurrences of class label *i* in cluster j.

$$p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}}$$

- The entropy is as a measure of "impurity" and accounts for the average information in an arbitrary message about the class label.

$$H_j = -\sum_{i=1}^{m} p_{ij} \log p_{ij}$$

- To evaluate the whole clustering we sum up the entropies of individual clusters weighted with the proportion of documents in each.

$$H = \sum_{j=1}^{k} \frac{n_j}{n} H_j$$

Entropy (Examples)

- A "pure" cluster where all documents have a single class label has entropy of 0.
- The highest entropy is achieved when all class labels have the same probability.
- For example, for a two class problem the 50-50 situation has the highest entropy of $(-0.5\log 0.5 - 0.5\log 0.5) = 1$.
- Compare the entropies of the previously discussed clusterings for attributes *offers* and *students*:

$$H(\textit{offers}) = \frac{11}{20}\left(-\frac{8}{11}\log\frac{8}{11} - \frac{3}{11}\log\frac{3}{11}\right) + \frac{9}{20}\left(-\frac{3}{9}\log\frac{3}{9} - \frac{6}{9}\log\frac{6}{9}\right) = 0.878176$$

$$H(\textit{students}) = \frac{15}{20}\left(-\frac{10}{15}\log\frac{10}{15} - \frac{5}{15}\log\frac{5}{15}\right) + \frac{5}{20}\left(-\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5}\right) = 0.869204$$