Major issue in clustering – labeling

- After a clustering algorithm finds a set of clusters: how can they be useful to the end user?
- We need a pithy label for each cluster.
- For example, in search result clustering for "jaguar", The labels of the three clusters could be "animal", "car", and "operating system".
- Topic of this section: How can we automatically find good labels for clusters?

Exercise

- Come up with an algorithm for labeling clusters
- Input: a set of documents, partitioned into K clusters (flat clustering)
- Output: A label for each cluster
- Part of the exercise: What types of labels should we consider? Words?

Discriminative labeling

- To label cluster ω, compare ω with all other clusters
- Find terms or phrases that distinguish ω from the other clusters
- We can use any of the feature selection criteria used in text classification to identify discriminating terms:
  (i) mutual information, (ii) χ^2, (iii) frequency
  (but the latter is actually not discriminative)

Non-discriminative labeling

- Select terms or phrases based solely on information from the cluster itself
- Terms with high weights in the centroid (if we are using a vector space model)
- Non-discriminative methods sometimes select frequent terms that do not distinguish clusters.
- For example, MONDAY, TUESDAY, ... in newspaper text

Using titles for labeling clusters

- Terms and phrases are hard to scan and condense into a holistic idea of what the cluster is about.
- Alternative: titles
- For example, the titles of two or three documents that are closest to the centroid.
- Titles are easier to scan than a list of phrases.

Feature selection

- In text classification, we usually represent documents in a high-dimensional space, with each dimension corresponding to a term.
- In this lecture: axis = dimension = word = term = feature
- Many dimensions correspond to rare words.
- Rare words can mislead the classifier.
- Rare misleading features are called noise features.
- Eliminating noise features from the representation increases efficiency and effectiveness of text classification.
- Eliminating features is called feature selection.

Example for a noise feature

- Let's say we're doing text classification for the class *China*.
- Suppose a rare term, say ARACHNOCENTRIC, has no information about *China* . . .
- . . . but all instances of ARACHNOCENTRIC happen to occur in *China* documents in our training set.
- Then we may learn a classifier that incorrectly interprets ARACHNOCENTRIC as evidence for the *China*.
- Such an incorrect generalization from an accidental property of the training set is called overfitting.
- Feature selection reduces overfitting and improves the accuracy of the classifier.

Different feature selection methods

A feature selection method is mainly defined by the feature utility measures it employs.

Feature utility measures:

- Frequency – select the most frequent terms
- Mutual information – select the terms with the highest mutual information (mutual information is also called information gain in this context)
- χ^2 (Chi-square)

Information

- $H[p] = \sum_{i=1,n} -p_i \log_2 p_i$ measures information uncertainty
- has maximum $H = \log_2 n$ for all $p_i = 1/n$

Consider two probability distributions:
$p(x)$ for $x \in X$ and $p(y)$ for $y \in Y$

- MI: $I[X;Y] = H[p(x)] + H[p(y)] - H[p(x,y)]$ measures how much information $p(x)$ gives about $p(y)$ (and vice versa)
- MI is zero iff $p(x,y) = p(x)p(y)$, i.e., x and y are independent for *all* $x \in X$ and $y \in Y$
- can be as large as $H[p(x)]$ or $H[p(y)]$

$$I[X;Y] = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Mutual information

- Compute the feature utility as the expected mutual information (MI) of term t and class c.
- MI tells us "how much information" the term contains about the class and vice versa.
- For example, if a term's occurrence is independent of the class (same proportion of docs within/without class contain the term), then MI is 0.
- Definition:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

$$= p(t,c) \log_2 \frac{p(t,c)}{p(t)p(c)} + p(\bar{t},c) \log_2 \frac{p(\bar{t},c)}{p(\bar{t})p(c)}$$

$$+ p(t,\bar{c}) \log_2 \frac{p(t,\bar{c})}{p(t)p(\bar{c})} + p(\bar{t},\bar{c}) \log_2 \frac{p(\bar{t},\bar{c})}{p(\bar{t})p(\bar{c})}$$

Consider a set of $N = 100$ articles, 10 of which contain the word *export*, 20 of which are in class POULTRY, and 5 of which both contain the word *export* and are in class POULTRY. (In N_{tc} notation, that's $N_{1.} = 10$, $N_{.1} = 20$, $N_{11} = 5$.)

Estimate the probabilities $p(e)$, $p(P)$, $p(\overline{e})$, $p(\overline{P})$, and joint probabilities $p(e, P)$, $p(e, \overline{P})$, $p(\overline{e}, P)$, $p(\overline{e}, \overline{P})$, to calculate the sum of the four terms in the mutual information

$$MI(export; POULTRY) = \sum_{t=e,\overline{e};\ c=P,\overline{P}} p(t, c) \log_2 \frac{p(t, c)}{p(t)p(c)}$$

and thereby infer the number of bits of information that the term and class contain about one another.

From $N_{1.} = 10$, $N_{.1} = 20$, and $N_{11} = 5$:
we infer $N_{10} = 5$, $N_{01} = 15$, and $N_{00} = 75$, so:

$$p(e, P) = N_{11}/N = .05 \qquad p(e, \overline{P}) = N_{10}/N = .05$$
$$p(\overline{e}, P) = N_{01}/N = .15 \qquad p(\overline{e}, \overline{P}) = .N_{00}/N = 75$$

and

$$p(e) = N_{1.}/N = 0.1 \qquad p(\overline{e}) = N_{0.}/N = 0.9$$
$$p(P) = N_{.1}/N = 0.2 \qquad p(\overline{P}) = N_{.0}/N = 0.8$$

Thus

$$MI[e; P] = .05 \cdot \log_2 \frac{.05}{.1 \cdot 0.2} + .05 \cdot \log_2 \frac{.05}{.1 \cdot .8}$$
$$+ .15 \cdot \log_2 \frac{.15}{.9 \cdot .2} + .75 \cdot \log \frac{.75}{.9 \cdot .8} = 0.03691 \text{ bits}$$

If instead there are only 2 articles that both contain the word *export* and are in class POULTRY? (i.e., $N_{11} = 2$, and otherwise still $N = 100$, $N_{1.} = 10$, $N_{.1} = 20$)

For $p(e, P) = .02$, $p(e, \overline{P}) = .08$, $p(\overline{e}, P) = .18$, $p(\overline{e}, \overline{P}) = .72$
$p(e) = 0.1$, $p(\overline{e}) = 0.9$, $p(P) = 0.2$, $p(\overline{P}) = 0.8$
the probabilities are independent,
$p(e, P) = p(e)p(P)$, etc.,
and hence all the logs are zero:

$$MI[e; P] = .02 \cdot \log_2 \frac{.02}{.1 \cdot .2} + .08 \cdot \log_2 \frac{.08}{.1 \cdot .8}$$

$$+.18 \cdot \log_2 \frac{.18}{.9 \cdot .2} + .72 \cdot \log \frac{.72}{.9 \cdot .8} = 0 \text{ bits}$$

How to compute MI values

- Based on maximum likelihood estimates, the formula we actually use is:

$$
\begin{aligned}
I(U;C) \quad = \quad & \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} \\
& + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}
\end{aligned} \qquad (1)
$$

- N_{11}: # of documents that contain t ($e_t = 1$) and are in c ($e_c = 1$)
- N_{10}: # of documents that contain t ($e_t = 1$) and not in c ($e_c = 0$)
- N_{01}: # of documents that don't contain t ($e_t = 0$) and in c ($e_c = 1$)
- N_{00}: # of documents that don't contain t ($e_t = 0$) and not in c ($e_c = 0$)
- $N = N_{00} + N_{01} + N_{10} + N_{11}$
- $p(t,c) \approx N_{11}/N$, $p(\overline{t},c) \approx N_{01}/N$, $p(t,\overline{c}) \approx N_{10}/N$, $p(\overline{t},\overline{c}) \approx N_{00}/N$
- $N_{1.} = N_{10} + N_{11}$: # documents that contain t, $p(t) \approx N_{1.}/N$
- $N_{.1} = N_{01} + N_{11}$: # documents in c, $p(c) \approx N_{.1}/N$
- $N_{0.} = N_{00} + N_{01}$: # documents that don't contain t, $p(\overline{t}) \approx N_{0.}/N$
- $N_{.0} = N_{00} + N_{10}$: # documents not in c, $p(\overline{c}) \approx N_{.0}/N$

MI example for POULTRY/*export* in Reuters

	$e_c = e_{\text{POULTRY}} = 1$	$e_c = e_{\text{POULTRY}} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 141$
$e_t = e_{export} = 0$	$N_{01} = 27{,}652$	$N_{00} = 774{,}106$

Plug these values into formula:

$$\begin{aligned}
I(U;C) &= \frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49+27{,}652)(49+141)}\\
&+\frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141+774{,}106)(49+141)}\\
&+\frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)}\\
&+\frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)}\\
&\approx 0.000105
\end{aligned}$$

MI feature selection on Reuters

Terms with highest mutual information for three classes:

COFFEE

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

SPORTS

soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

POULTRY

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

$I(\textit{export}, \textsc{poultry}) \approx .000105$ not among the ten highest for class POULTRY, but still potentially significant.