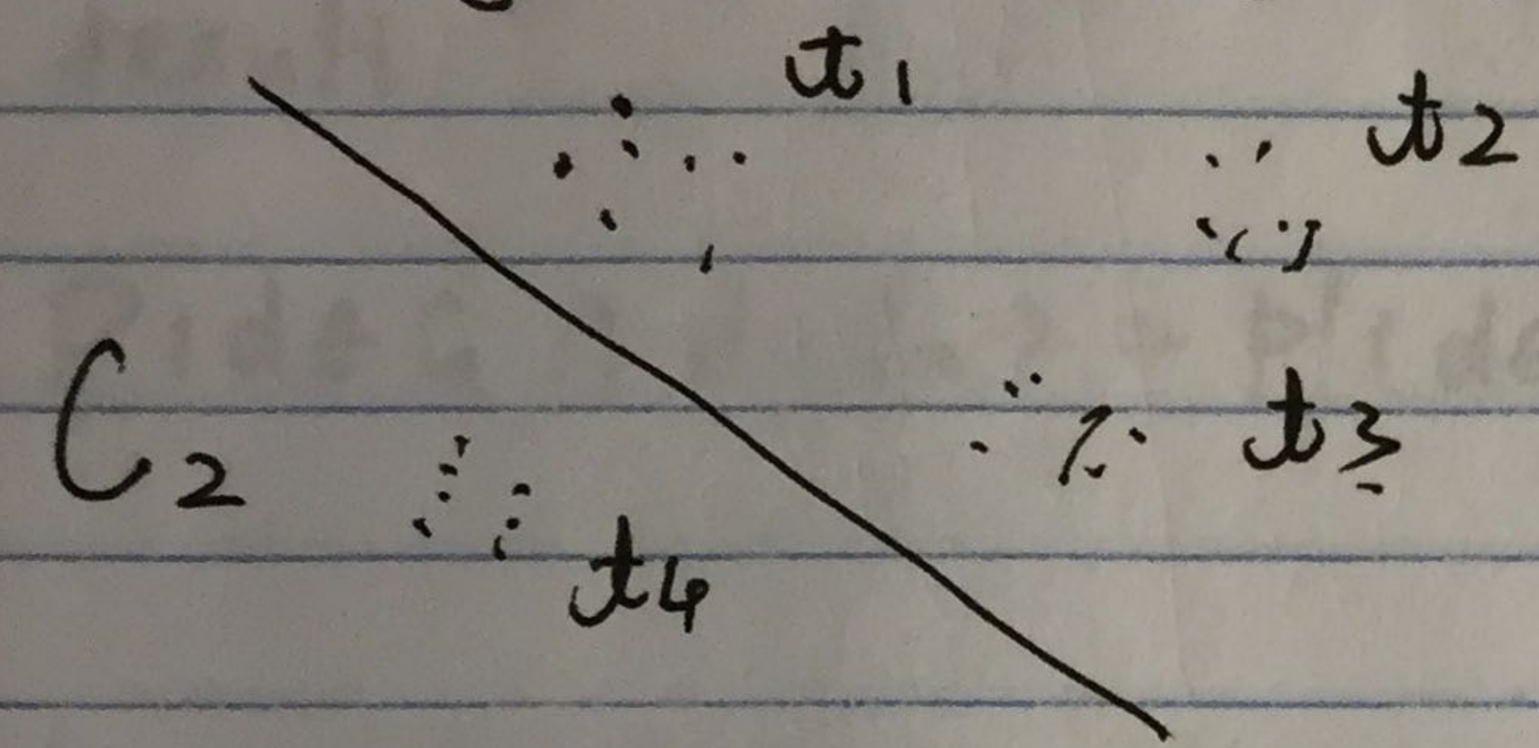
Model:

The data are clustered in nature, the observed categories are groups of the underlying clusters.



t_1, t_2, t_3, t_4 are the actual clusters,

C_1 and C_2 are the observed categories.

Taking naive bayes as an example, the probability for a data point d to be in C_1 is

$$P(d \in C_1) = \sum_{i=1}^{3} p(d \in t_i) = \sum_{i=1}^{3} \prod_{w \in d} p(w|t_i) \cdot P(t_i)$$

The directly trained model is:

$$p'(d \in C_1) = \prod_{w \in d} P(w|C_1) \cdot P(C_1)$$

$$= \prod_{w \in d} \sum_{i=1}^{3} p(w|t_i) P(t_i) \cdot \left(\sum_{i=1}^{3} P(t_i)\right)$$

These two can be different.

The previously proposed cross domain classification:

The previously proprosed model gives the followiy result:

$$P(d \in C_1 \cap d \in C_2) = P'(d \in C_1) P'(d \in C_2) \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

it is hard to show that this is less biased.

Thus, it is better to discover the latent clusters directly.

But there is an issue that $P(t_i \in C_2) = 0$,

this will cause problems in EM, I haven't figure out how to treat it properly.

But if we want some preliminary results, we may try the followj two algorithms first.

(Detailed direvation will not be shown..)

Algorithm 1:

Input: $D = \{d_1, d_2, \cdots\}$. $C = \{C_1, C_2 \cdots\}$

$\{C_d = C_i, d \in C_i\}$

Initialized $t = C$, $flag = true$.

while (flag):

$$p(t) = \frac{\mathbb{1}\{t_d = t\}}{M} ; \quad p(w|t) = \frac{\sum \mathbb{1}\{t_d = t_i, w \in d\}}{\sum \mathbb{1}\{t_d = t\}}$$

E-M iterations:

$$E: \quad Q(C, t, d) = \frac{p(t)\, p(C|t) \prod_w p(w|t)}{\sum_{\hat{c},\hat{t}} p(\hat{t})\, p(\hat{c}|\hat{t}) \prod_w p(w|\hat{t})}$$

$$M: \quad p(t_i) = \frac{\sum Q(C, t_i | d)}{\sum Q(C, t | d)}$$

$$p(w|t_i) = \frac{\sum Q(C, t_i | d)\, \mathbb{1}\{w \in d\}}{\sum Q(C, t_i | d)}$$

flag = false:

For C_i in C:

$IG = H(C_i) - H(C_i | t_c)$

If $IG > IG_{pre}$:

$IG_{pre} = IG$

flag = true

T_{C_i}.add t_{new}

$t_{new} = \{d_i ; t_d \notin T_{C_i}, d \in C_i\}$

Algorithm 2:

$t \leftarrow C.$

while $|\{t\}| < maxNum$:

$$p(t_i) = \frac{\sum \mathbb{1}\{t_d = t_i\}}{|\{t\}|}.$$

$$p(w|t_i) = \frac{\sum \mathbb{1}\{t_d = t_i, w \in d\}}{\sum \mathbb{1}\{t_d = t_i\}}$$

$$P(c|t) = \frac{\sum \mathbb{1}\{t_d = t_i, d \in C\} + 1}{\sum \mathbb{1}\{t_d = t_i\} + |\{t\}|}$$

E-M iteration.

$Q(t, c|d)$. $p(t_i)$, $P(w|t_i)$, $P(c|t)$.

For t_i in $\{t\}$:

$t_i \in C$ if $p(C|t_i) \geq p(c'|t_i)$ $\forall c'$.

Record $\{t\}$.

$t_{new} = \{d;\ t_d \neq T_{C_i},\ d \in C_i\}$.

Choose the best $\{t\}$ manually through validation.