Noisy-Or Classifier

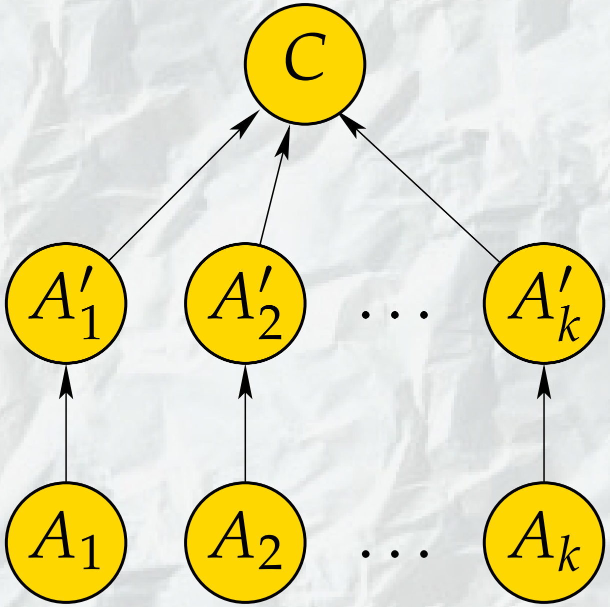
Jiří Vomlel

Laboratory for Intelligent Systems
and
Institute of Information Theory and Automation,

Prague, Czech Republic

http://www.utia.cas.cz/vomlel/

Model of a Noisy-Or Classifier



$$\begin{aligned} P_M(C=0 \mid \mathbf{A}=\mathbf{a}) &= \prod_j P_M(A'_j = 0 \mid A_j = a_j) \\ P_M(C=1 \mid \mathbf{A}=\mathbf{a}) &= 1 - \prod_j P_M(A'_j = 0 \mid A_i = a_j) \end{aligned}$$

Using a threshold $0 \le t \le 1$ all data vectors $\mathbf{a} = (a_1, \dots, a_k)$ such that

$$P_M(C=0 \mid \mathbf{A}=\mathbf{a}) < t$$

are classified to class $C = 1$.

Learning from the training data - I

Let $D = \{\mathbf{e}^1, \ldots, \mathbf{e}^n\}$ be the training data, where the instances are

$$\mathbf{e}^i = \{c^i, \mathbf{a}^i\} = \{c^i, a_1^i, \ldots, a_k^i\}, \quad \text{for i=1,...,n.}$$

and $n(c, \mathbf{a})$ is the occurrence of $(c, \mathbf{a})$ in D.

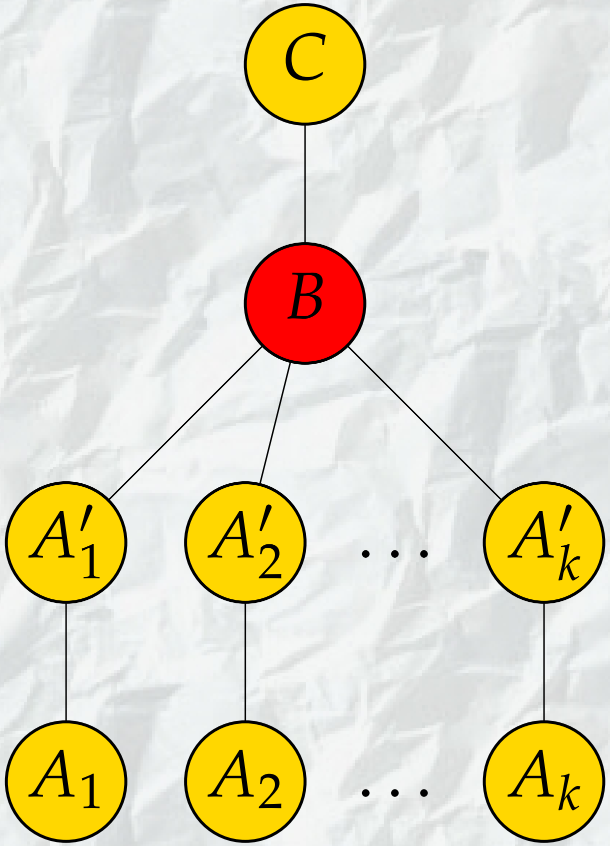
The learning process aims at a model P_M maximizing the ability to correctly predict class C. Assuming i.i.d. data D

$$\begin{aligned} CLL(P_M \mid D) \ &= \\ &= \sum_{i=1}^{n} \log P_M(C = c^i \mid \mathbf{A} = \mathbf{a}^i) \\ &= \sum_{\mathbf{a}} \sum_{c} n(c, \mathbf{a}) \cdot \log P_M(C = c \mid \mathbf{A} = \mathbf{a}) \\ &= n \cdot \sum_{\mathbf{a}} P_D(\mathbf{A} = \mathbf{a}) \cdot \sum_{c} P_D(C = c \mid \mathbf{A} = \mathbf{a}) \cdot \log P_M(C = c \mid \mathbf{A} = \mathbf{a}) \ . \end{aligned}$$

Learning from the training data - II

If a noisy-or classifier P_M maximizes log-likelihood $LL(P_M \mid D)$ then it also maximizes conditional log-likelihood $CLL(P_M \mid D)$.

$$
\begin{aligned}
LL(P_M \mid D) \; &= \\
&= \sum_{i=1}^{n} \log P_M(C = c^i, \mathbf{A} = \mathbf{a}^i) \\
&= \sum_{i=1}^{n} \log P_M(C = c^i | A_1 = a_1^i, \ldots, A_k = a_k^i) + \sum_{i=1}^{n} \log \prod_{j=1}^{k} P_M(A_j = a_j^i) \\
&= CLL(P_M \mid D) + \sum_{i=1}^{n} \sum_{j=1}^{k} \log P_M(A_j = a_k^i) \; .
\end{aligned}
$$



EM-algorithm

The expectation step:

$$n(A'_\ell, A_\ell) \quad = \quad \sum_{i=1}^{n} P_M(A'_\ell, A_\ell \mid \mathbf{e}^i) \ \text{ for all } \ell = 1, \dots k \ ,$$

The maximization step:

$$P^\star_M(A'_\ell \mid A_\ell) \quad = \quad \frac{n(A'_\ell, A_\ell)}{n(A_\ell)}, \ \text{ for all } \ell = 1, \dots k \ .$$

Logistic discrimination

The fundamental assumption of logistic discrimination:

$$\log \frac{P_M(C=0 \mid \mathbf{A}=\mathbf{a})}{1-P_M(C=0 \mid \mathbf{A}=\mathbf{a})} \quad = \quad \beta_0 + \sum_j \beta_j \cdot a_j \;=\; \beta_0 + \sum_{j:a_j=1} \beta_j \ ,$$

which implies $P_M(C=0 \mid \mathbf{A}=\mathbf{a}) \quad = \quad \dfrac{\exp(\beta_0 + \sum_{j:a_j=1} \beta_j)}{1+\exp(\beta_0 + \sum_{j:a_j=1} \beta_j)} \ .$

A data vector $\mathbf{a}$ is classified to $C=1$ provided

$$\log \frac{P_M(C=0 \mid \mathbf{A}=\mathbf{a})}{1-P_M(C=0 \mid \mathbf{A}=\mathbf{a})} \quad < \quad 0 \ .$$

Noisy-or revisited

$$
\begin{aligned}
\log P_M(C = 0 \mid \mathbf{A} = \mathbf{a}) &= \sum_j \log P_M(A'_j = 0 \mid A_j = a_j) \\
&= \sum_{j:a_j=0} p_j(0) + \sum_{j:a_j=1} p_j(1) \\
&= \sum_j p_j(0) + \sum_{j:a_j=1} p_j(1) - p_j(0) \\
&= r_0 + \sum_{j:a_j=1} r_j \; .
\end{aligned}
$$

Let $r_0 = \beta_0 - \log(1 + \exp(\beta_0 + \sum_{j:a_j=1} \beta_j))$ and $r_j = \beta_j, j = 1, \ldots, k$.

$$
P_M(C = 0 \mid \mathbf{A} = \mathbf{a}) = \exp(r_0 + \sum_{j:a_j=1} r_j) = \frac{\exp(\beta_0 + \sum_{j:a_j=1} \beta_j)}{1 + \exp(\beta_0 + \sum_{j:a_j=1} \beta_j)} \; .
$$

This is the formula used in logistic discrimination.

The criteria used in the logistic discrimination is

$$\log \frac{P_M(C = 0 \mid \mathbf{A} = \mathbf{a})}{1 - P_M(C = 0 \mid \mathbf{A} = \mathbf{a})} \quad < \quad 0 \ .$$

$f(x) = \log \frac{x}{1-x}$ is a monotone function of x. Therefore this criteria is equivalent to

$$P_M(C = 0 \mid \mathbf{A} = \mathbf{a}) \quad < \quad 0.5$$

Note that f $\log \frac{t}{1-t} = 0$.

Classification using the noisy-or classifier with the threshold t $= 0.5$ and logistic discrimination provide equivalent results.

Reuters text collection

We have used the Reuters-21578 text categorization collection containing the Reuters new stories.

The split of data to the training and testing sets was according to the time of documents (ModApte).

Classes that contained only one document were eliminated together with the corresponding documents. The resulting **training set** contained 7769 documents and **testing set** 3018 documents.

After removing function words and words that appeared only in one document the **feature set** contained 15715 words. For each class we selected atributes for which the expected information gain

$$IG(A) = \sum_a P(A=a) \cdot (H(P(C)) - H(P(C \mid A=a))) \ \geq \ 0.005 \ ,$$

H denotes the Shannon entropy.

Performance measures

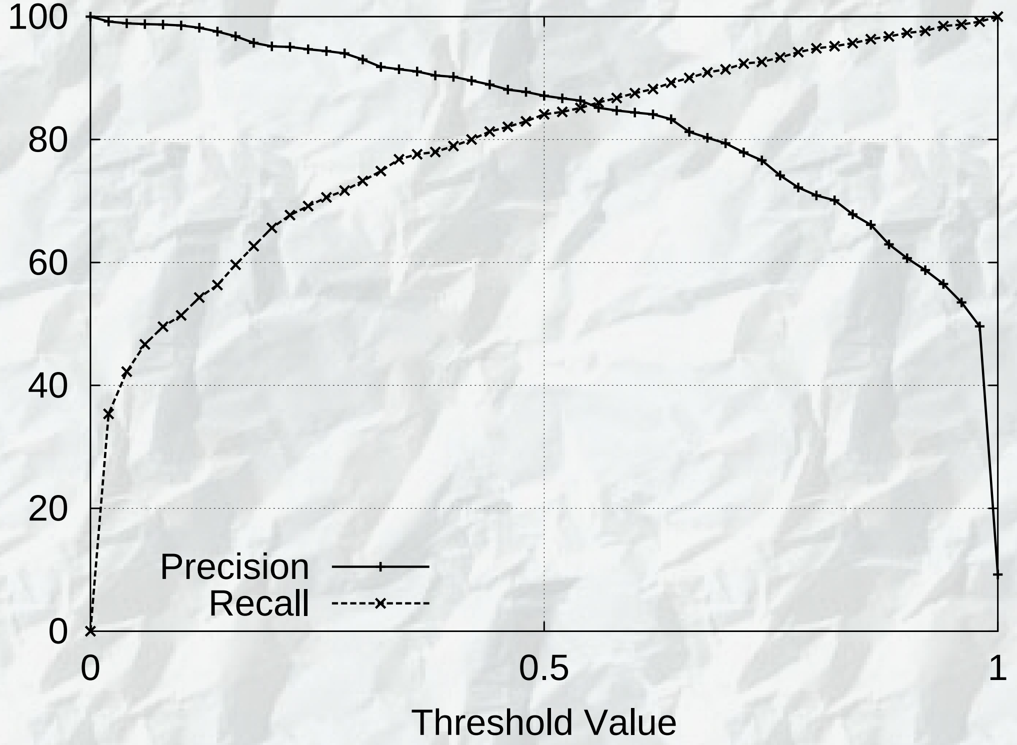
For each class we counted the documents:

- *tp*, true positive ... classified correctly $C = 1$,
- *fp*, false positive ... classified incorrectly $C = 1$,
- *tn*, true negative ... classified correctly $C = 0$,
- *fn*, false negative ... classified incorrectly $C = 0$.

precision $\pi = \frac{tp}{tp+fp}$ recall $\rho = \frac{tp}{tp+fn}$

Macroaverage is average of local values of π and ρ in each class.

Microaverge is computed so that the total sums of *tp*,*fp*, and *fn* are computed over all ten tested classes and then π and ρ are computed from the total sums of *tp*,*fp*, and *fn*.



Microaverage precision and recall

Comparisons of breakeven points

class	*nd*	NB	SVM	N-OR
earn	1087	96.7	**98.1**	95.6
acq	719	89.4	**93.7**	86.1
crude	189	83.1	81.5	**85.2**
money-fx	179	61.3	**66.5**	63.5
grain	149	75.6	85.9	**91.8**
interest	131	58.6	**65.6**	56.4
trade	117	54.7	**70.1**	54.7
ship	89	80.9	69.7	**84.3**
wheat	71	71.5	**85.9**	81.4
corn	56	57.0	83.9	**84.0**
micro-avg.		84.4	**88.9**	85.7
macro-avg.		72.9	**81.8**	78.3

Conclusions and future work

- Noisy-or model can be used to perform classification tasks when the instances are described by a large number of attributes.
- Classification using the noisy-or corresponds to logistic discrimination.
- Efficient version of the EM-algorithm can be used to learn parameters of the noisy-or.
- Efficient version of the EM-algorithm can be easily modified for other models containing functional dependence, for example, noisy-and, noisy-max, noisy-min.
- We plan to analyze differences between the learning methods used in logistic discrimination and the EM-algorithm used for the noisy-or model.