A GENERALIZATION OF THE NOISY-OR MODEL

Jiří Vomlel

In this paper we generalize the noisy-or model. The generalization is two-fold. First, we allow parents to be multivalued ordinal variables. Second, parents can have both positive and negative influences on their common child. Our generalization has several advantages: it requires only one parameter per parent, the child variable is still binary, and each inhibition probability depends on the value of the corresponding parent. We show that our generalized noisy-or model belongs to the family of Generalized Linear Models, namely, it is a subfamily of Poisson Regression models. We suggest a method for learning the models and report results of experiments on the Reuters text classification data. The multivalued noisy-or achieved better performance than standard noisy-or for classes with lower number of documents.

Keywords: Bayesian Networks, Noisy-or Model, Classification, Generalized Linear Models, Poisson Regression

Classification: 68T37, 68T30

1. INTRODUCTION

Conditional probability tables (CPTs) that are the basic building blocks of Bayesian networks [10, 7] have, generally, an exponential size with respect to the number of variables of the CPT. This has two unpleasant consequences. First, during the elicitation of model parameters one needs to estimate an exponential number of parameters. Second, in case of a high number of parent variables the exact probabilistic inference may become intractable.

On the other hand real implementations of Bayesian networks (see e.g. [9]) often have a simple local structure of the CPTs. The noisy-or model [10] is a popular model for describing relations between variables in one CPT of a Bayesian network. Noisy-or is member of a family of models called models of independence of causal influence [5] or canonical models [3]. The advantage of these models is that the number of parameters required for their specification is linear with respect to the number of variables in CPTs and that they allow applications of efficient inference methods, see for example [4, 12].

In this paper we propose a generalization of the noisy-or model to multivalued parent variables. Our proposal differs from the noisy-max model [6] since we keep the child variable binary no matter what the number of states of the parent variables are. Also we have only one parameter for each parent no matter what is the number of states of the parent variables. Our generalization also differs from the generalization of the noisy-or model proposed by Srinivas [13] since in his model the inhibition probabilities cannot depend on the state of the parent variables if the state differs from the state of the child. We consider this to be a quite restrictive requirement for some applications.

We will show that our proposal is closely connected with the Poisson Regression of Generalized Linear Models [8]. We discuss methods one can use to learn parameters of the generalized noisy-or model from data. In the second part of the paper we present results of numerical experiments on the well-known Reuters text classification data. We use this dataset to compare the performance of suggested generalizations of multivalued and binary noisy-or models. We have made the source code and datasets used in experiments available at the worldwide web – see Section 5.

2. MULTIVALUED NOISY-OR

In this section we propose a generalization of noisy-or for multivalued parent variables. Let Y be a binary variable taking states $y \in \{0,1\}$ and $X_i, i = 1,\ldots,n$ be multivalued discrete variables taking states $x_i \in \{0,1,\ldots,m_i\}$, $m_i \in \mathbb{N}^+$. The local structure of both the standard (see, e.g., [3]) and the multivalued generalization of the noisy-or can be made explicit as it is shown in Figure 1.

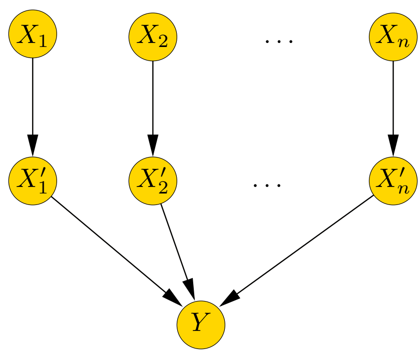

Fig. 1. Noisy-or model with the explicit deterministic (OR) part.

The conditional probability table $P(Y|X_1,\ldots,X_n)$ is defined using CPTs $P(X'_i|X_i)$ as

$$P(X'_i = 0|X_i = x_i) \quad = \quad (p_i)^{x_i} \tag{1}$$
$$P(X'_i = 1|X_i = x_i) \quad = \quad 1 - (p_i)^{x_i} \quad , \tag{2}$$

where (for $i = 1,\ldots,n$) $p_i \in [0,1]$ is the parameter that defines the probability that the positive value x_i of variable X_i is inhibited. In the formula, we use parenthesis to emphasize that x_i is an exponent, not an upper index of p_i. The CPT $P(Y|X'_1,\ldots,X'_n)$ is deterministic and represents the logical OR function.

Remark. Note that the higher is the value x_i of X_i the lower the probability of $X'_i = 0$, which is a desirable property in many applications.

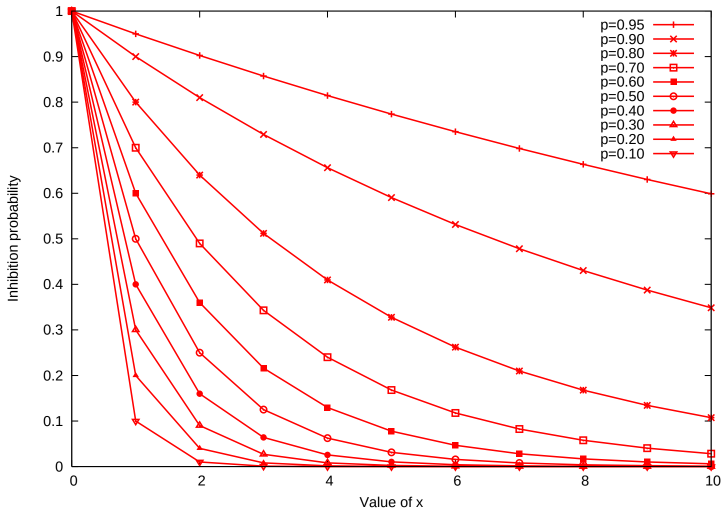

Fig. 2. The dependence of $P(X'_i = 0|X_i = x_i)$ on parameter $p_i = p$ and the variable value $x_i = x$.

The conditional probability table $P(Y|X_1, \ldots, X_n)$ is then defined as

$$\begin{aligned} P(Y = 0|X_1 = x_1, \ldots, X_n = x_n) &= \prod_{i=1}^{n} P(X'_i = 0|X_i = x_i) \\ &= \prod_{i=1}^{n} (p_i)^{x_i} \quad (3) \\ P(Y = 1|X_1 = x_1, \ldots, X_n = x_n) &= 1 - \prod_{i=1}^{n} (p_i)^{x_i} \ . \quad (4) \end{aligned}$$

Remark. Note that if $m_i = 1$, i.e. the values x_i of X_i are either 0 or 1, then we get the standard noisy-or model [10, 3].

In Figure 2 dependence of the inhibitory probability $P(X' = 0|X = x)$ on the value x of a variable X is depicted for different values of the parameter p.

It is important to note that contrary to the definition of causal noisy-max [3, Section 4.1.6] we have only one parameter p_i for each parent X_i of Y no matter what is the number of states of X_i. This implies that our model is more restricted. But, on the other hand, the suggested simple parametrization guarantees ordinality, which is in many application a desirable property (as it is also discussed in [3]). Also, since fewer parameters are estimates or elicited from domain experts the estimates might be more reliable.

3. RELATION TO THE POISSON REGRESSION

In this section we will describe the relation of the generalized noisy-or model and the Poisson Regression of Generalized Linear Models [8].

By taking the logarithm of both sides of equation (3) we get

$$\log P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad \sum_{i=1}^{n} x_i \cdot \log p_i \quad .$$

Define a new parameter $\beta_i = \log p_i$. Note that since p_i is probability it holds that $\beta_i \in (-\infty, 0]$. Then we can write

$$\log P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad \sum_{i=1}^{n} x_i \cdot \beta_i \quad = \quad \boldsymbol{x}^T \boldsymbol{\beta} \quad ,$$

where $\boldsymbol{x}$ denotes vector $(x_1,\ldots,x_n)$ and $\boldsymbol{\beta}$ denotes vector $(\beta_1,\ldots,\beta_n)$. The above formula is the formula of the Poisson Regression of the binary variable $1-Y$. Since the expected value $E(1-Y|x_1,\ldots,x_n) = P(Y=0|X_1=x_1,\ldots,X_n=x_n)$ it holds that

$$\log E((1-Y)|x_1,\ldots,x_n) \quad = \quad \boldsymbol{x}^T \boldsymbol{\beta} \quad .$$

However, as it is noted above, in the generalized noisy-or model we require non-positive values of $\boldsymbol{\beta}$. There is not any such constraint in the Poisson Regression. In this sense, the generalized noisy-or models are a subfamily of the Poisson Regression models.

In the Poisson Regression the first value of vector $\boldsymbol{\beta}$ is usually the intercept. Often, it is included in the model as β_0 with corresponding entry x_0 in $\boldsymbol{x}$ being fixed to 1. In the generalized noisy-or this corresponds to a leaky cause [3], i.e. an auxiliary cause that is always in state 1. This allows the probability

$$P(Y=0|X_1=x_1,\ldots,X_n=x_n) \quad = \quad p_L \quad < \quad 1$$

even if $\boldsymbol{x} = (x_1,\ldots,x_n) = \boldsymbol{0}$. The value $p_L = \exp(\beta_0)$ is called leaky probability and usually it is close (but not equal) to one. This allows to model unobserved or unknown causes of $Y=1$.

4. LEARNING PARAMETERS OF THE GENERALIZED NOISY-OR

The relation to the Poisson Regression allows us to apply standard maximum likelihood estimation methods for Poisson Regression models. A method typically used to learn the generalized linear models is the Iteratively Reweighted Least Squares (IRLS) method [8]. By use of the standard Poisson Regression methods some values $\beta_i, i=1,\ldots,n$ may be learned to be positive. We can give this a quite natural interpretation that the higher values of X_i imply higher inhibitory probability. In the generalized noisy-or model we can treat the parents with positive values of β_i by relabeling their values $x_i \in \{0,1,\ldots,m_i\}$

to $(m_i - x_i) \in \{m_i, \ldots, 1, 0\}$. In this way the generalized noisy-or is now capable to treat not only positive influences (presence of X_i increases probability of $Y = 1$) but also negative influences (presence of X_i decreases probability of $Y = 1$).

To guarantee that for all possible values of $\boldsymbol{x}$ we get probability values (i.e. values from $[0, 1]$) we need to use learning methods with the constraints $\boldsymbol{\beta} \leq \mathbf{0}$. In the experiments reported in the next section we used a quasi-Newton method with box constraints [2] implemented in R [11]. When searching for the minimum of the negative log likelihood we used the formula (6) for the gradient derived in Appendix A. To avoid infinite numbers we added a small positive value (10^{-10}) to the denominator in formula (6). The algorithm was started from ten different initial values of $\boldsymbol{\beta}$ generated randomly from interval $[-1, +1]$.

In the Poisson Regression each parent has either positive (+) or a negative influence (−) on $Y = 1$ (we exclude parents with no influence). Thus, we can get two versions of generalized binary noisy-or and two versions of generalized multivalued noisy-or that either have:

- all parents $X_i, i = 1, \ldots, n$ of Y have positive influence (+) on probability of $Y = 1$ or
- all parents from $X_i, i = 1, \ldots, n$ of Y that have negative influence (−) in the Poisson Regression have their states renumbered reversely (as discussed above) and then included with a positive influence.

Remark. Note that the generalized binary noisy-or with all parents having a positive influence (+) is the standard noisy-or.

5. EXPERIMENTS

In this section we describe experiments we performed with the well known Reuters-21578 collection (Distribution 1.0) of text documents. The text documents from this dataset appeared on the Reuters newswire in 1987 and were manually classified by personnel from Reuters Ltd. and Carnegie Group, Inc. to several classes according to their topic. In the test we used the split of documents to training and testing sets according to Apté et al. [1]. We performed experiments with preprocessed data for eight largest classes[1]. To allow interested readers to replicate our experiments easily we have made our R code and the datasets used in experiments available on the web [2].

In the experiments we compare the generalized binary noisy-or classifier and the generalized multivalued noisy-or classifier. Both models were learned using the iteratively reweighted least squares method [8] implemented in R – a language and environment for statistical computing [11]. We performed experiments with three versions of both types of classifiers (with binary features and with multivalued features, respectively):

(a) features X_i with a positive (+) or a negative influence (−) were both allowed – this is the standard Poisson Regression,

(b) only features X_i with a positive influence (+) were included.

[1]The preprocessed dataset is available at `http://web.ist.utl.pt/acardoso/datasets/`.
[2]The code and data are available at `http://www.utia.cas.cz/vomlel/generalized-noisy-or/`

(c) features X_i with a positive influence (+) were included and features X_i with negative influence (−) have their states renumbered in the reverse order and then they are also included.

Remark. Note that the binary classifier consisting of binary features X_i with positive influence (+) on probability of $Y = 1$ only is the standard noisy-or classifier [14].

We decided to include into the models all features that were not rejected as irrelevant at the significance level 0.1. We performed the experiments also with the significance level increased to 0.3. In this way, we increased the number of features, but since for most classes there was no significant increase in the accuracy we prefer simpler models. However, it may be topic for a future research to apply exhaustive feature selection methods that would find optimal models for the families of our interest.

The results of experiments are summarized in Tables 1 and 2. The accuracy is reported using the percentage scale, it is the relative proportion of correctly classified documents either as belonging to the given class or not. From Tables 1 and 2 we can see that often binary noisy-or performs better for larger models, while multivalued noisy-or is often better at smaller models. The models for the classes *ship* and *grain* are very small, they have at most three features only and there is no difference between the models' accuracy.

The most general model that is not restricted performs in average slightly better than those two restricted to positive features or positive and reversed features respectively. In the case of binary noisy-or there is no difference in performance of the two restricted models. In the case of multivalued noisy-or the model with positive and reversed features is slightly better than the one with positive features only.

6. AN EXAMPLE

In this section we use the class *interest* to illustrate the benefits of treating the features as multivalued. In the first example we present the binary noisy-or model and in the second the multivalued noisy-or model. Both models contain only significant features for the significance level 0.1.

Example 1 (The noisy-or model for the interest class)**.** In Figure 3 the structure of the noisy-or model for the interest class is presented (in the examples we do not make the deterministic part explicit). All variables are binary, taking values 0 or 1. The leaky cause has a fixed value 1. The conditional probability fo the class interest is defined as

$$P(Class.interest = 0|Rate = r, Fed = f, Prime = p) \quad = \quad (p_1)^r \cdot (p_2)^f \cdot (p_3)^p \cdot p_0 \ ,$$

where $r \in \{0, 1\}$ is the state of feature *Rate*, $f \in \{0, 1\}$ is the state of feature *Fed*, and $p \in \{0, 1\}$ is the state of feature *Prime*. The values of parameters p_1, p_2, p_3 were estimated to be

$$\begin{aligned} p_1 &= \exp(\beta_1) = \exp(-0.387625201) \doteq 0.6786667 \\ p_2 &= \exp(\beta_2) = \exp(-0.377786535) \doteq 0.6853768 \\ p_3 &= \exp(\beta_3) = \exp(-0.412698847) \doteq 0.6618616 \end{aligned}$$

Tab. 1. Comparisons of the accuracy of the generalized noisy-or. The best achieved accuracy is printed boldface and framed if it is the best when considering also multivalued generalizations from Table 2.

Class	# test documents	(+ and -)	(only +)	(only + & rev.)	nr. features
earn	1083	**93.74143**	93.55870	93.55870	17
acq	696	$\boxed{\mathbf{94.01553}}$	91.73138	91.73138	28
crude	121	$\boxed{\mathbf{97.57880}}$	$\boxed{\mathbf{97.57880}}$	$\boxed{\mathbf{97.57880}}$	4
money-fx	87	96.66514	$\boxed{\mathbf{96.80219}}$	$\boxed{\mathbf{96.80219}}$	4
interest	81	**96.71083**	**96.71083**	**96.71083**	3
trade	75	**97.35039**	**97.35039**	**97.35039**	5
ship	36	$\boxed{\mathbf{99.04066}}$	$\boxed{\mathbf{99.04066}}$	$\boxed{\mathbf{99.04066}}$	2
grain	10	$\boxed{\mathbf{99.90863}}$	$\boxed{\mathbf{99.90863}}$	$\boxed{\mathbf{99.90863}}$	1
total	2189				

Tab. 2. Comparisons of the accuracy of the multivalued generalization of the noisy-or. The best achieved accuracy is printed boldface and framed if it is the best when considering also binary generalizations from Table 1.

Class	# test documents	(+ and -)	(only +)	(only + & rev.)	nr. features
earn	1083	$\boxed{\mathbf{93.96985}}$	93.78712	93.83280	13
acq	696	**92.37095**	92.18821	92.23390	23
crude	121	**97.25902**	**97.25902**	**97.25902**	3
money-fx	87	**96.75651**	96.71083	96.71083	4
interest	81	97.62449	$\boxed{\mathbf{97.85290}}$	$\boxed{\mathbf{97.85290}}$	2
trade	75	$\boxed{\mathbf{98.49246}}$	98.44678	98.44678	4
ship	36	$\boxed{\mathbf{99.04066}}$	$\boxed{\mathbf{99.04066}}$	$\boxed{\mathbf{99.04066}}$	3
grain	10	$\boxed{\mathbf{99.90863}}$	$\boxed{\mathbf{99.90863}}$	$\boxed{\mathbf{99.90863}}$	1
total	2189				

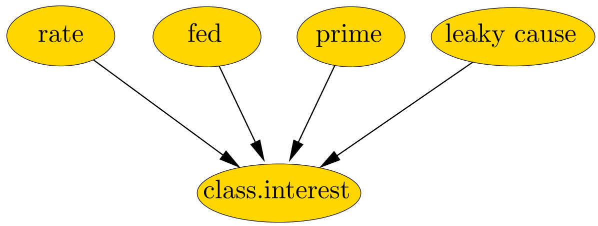

Fig. 3. Binary noisy-or model for the interest class.

and the leaky parameter $p_0 = \exp(\beta_0)$ was estimated to be

$$p_0 \quad = \quad \exp(\beta_0) \quad = \quad \exp(-0.001797324) \doteq \ 0.9982043 \ .$$

This model has accuracy 96.71%.

Example 2 (The multivalued noisy-or model for the interest class)**.** In Figure 4 the structure of the multivalued noisy-or model for the interest class is presented. The

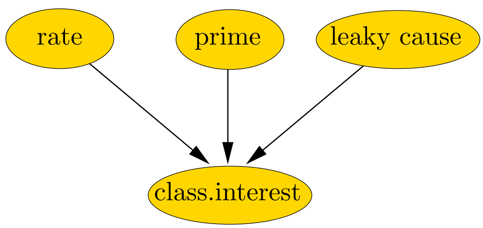

Fig. 4. Multivalued noisy-or model for the interest class.

variable *Rate* takes values from the set $\{0, 1, \ldots, 34\}$, variable *Prime* takes values from the set $\{0, 1, \ldots, 17\}$. The leaky cause has fixed state 1. The conditional probability for the class interest is defined as

$$P(Class.interest = 0 | Rate = r, Prime = p) \quad = \quad (p_1)^r \cdot (p_2)^p \cdot p_0 \ ,$$

where $r \in \{0, 1, \ldots, 34\}$ is the state of feature *Rate* and $p \in \{0, 1, \ldots, 17\}$ is the state of feature *Prime*. The values of parameters p_1 and p_2 were estimated to be

$$\begin{aligned} p_1 &= \exp(\beta_1) = \exp(-0.174466109) \doteq 0.8399053 \\ p_2 &= \exp(\beta_2) = \exp(-0.326624088) \doteq 0.7213549 \end{aligned}$$

and the leaky parameter $p_0 = \exp(\beta_0)$ was estimated to be

$$p_0 \quad = \quad \exp(\beta_0) \quad = \quad \exp(-0.004392911) \doteq \ 0.9956167 \ .$$

Despite this model has less features than the noisy-or model from Example 1 its accuracy is higher (97.85% > 96.71%). This example illustrates how we can benefit from multivalued features.

7. CONCLUSIONS

In this paper we proposed generalizations of the popular noisy-or model to multivalued explanatory variables. We showed the our generalizations are subfamilies of the Poisson family of Generalized Linear Models. We used iteratively reweighted least squares method to learn models from the Poisson family. For the restricted subfamilies corresponding to generalized noisy-or models we used a quasi-Newton method with box constraints. In the experiments with the Reuters text collection the generalized binary noisy-or performed better for larger models, while the generalized multivalued noisy-or performed better for smaller models. They both represent handy generalizations of noisy-or for real applications with multivalued variables and/or with parent variables being a mixture of variables having either positive or negative influence on their child variable.

ACKNOWLEDGEMENT

I am grateful to Remco Bouckaert from The University of Auckland, New Zealand for his suggestion to consider generalizations of noisy-or classifier [14] to multivalued variables. This work was supported by the Czech Science Foundation through the project 13-20012S. A preliminary version of this paper appeared in the proceedings of the 16th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty [15].

REFERENCES

[1] Ch. Apté, F. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251, 1994.

[2] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16:1190–1208, 1995.

[3] F. J. Díez and M. J. Druzdzel. Canonical probabilistic models for knowledge engineering. Technical Report CISIAD-06-01, UNED, Madrid, Spain, 2006.

[4] F. J. Díez and S. F. Galán. An efficient factorization for the noisy MAX. *International Journal of Intelligent Systems*, 18:165–177, 2003.

[5] D. Heckerman and J. Breese. A new look at causal independence. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA*, pages 286–292. Morgan Kaufmann, 1994.

[6] Max Henrion. Practical issues in constructing a Bayes' Belief Network. In *Proceedings of the Third Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 132–139. AUAI Press, 1987.

[7] F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs, 2nd ed.* Springer, 2007.

[8] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, London, 1989.

[9] R. A. Miller, F. E. Fasarie, and J. D. Myers. Quick medical reference (QMR) for diagnostic assistance. *Medical Computing*, 3:34–48, 1986.

[10] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA, 1988.

[11] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[12] P. Savicky and J. Vomlel. Exploiting tensor rank-one decomposition in probabilistic inference. *Kybernetika*, 43(5):747–764, 2007.

[13] Sampath Srinivas. A generalization of the noisy-or model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 208–215. Morgan Kaufmann, 1993.

[14] J. Vomlel. Noisy-or classifier. *International Journal of Intelligent Systems*, 21:381–398, 2006.

[15] J. Vomlel. A generalization of the noisy-or model to multivalued parent variables. In *The Proceedings of the 16th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty*, pages 19–27, 2013.

A. THE LOG LIKELIHOOD AND ITS GRADIENT

Assume data vectors $(y_i, \boldsymbol{x}_i^T), i = 1, \ldots, N$ and recall that boldface symbols denote vectors. Under the generalized noisy-or model the probability of y_i given $\boldsymbol{x}_i$ is defined as:

$$\begin{aligned} \log P(y_i = 0|\boldsymbol{x}_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ \log P(y_i = 1|\boldsymbol{x}_i) &= \log(1 - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})) \ . \end{aligned}$$

The log likelihood of data given this model is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} (1 - y_i)\boldsymbol{x}_i^T \boldsymbol{\beta} + y_i \log(1 - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})) \ . \tag{5}$$

The gradient of the log likelihood is the vector of partial derivatives with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{N} (1 - y_i)\boldsymbol{x}_i - y_i \boldsymbol{x}_i \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \\ &= \sum_{i=1}^{N} \frac{(1 - y_i)\boldsymbol{x}_i - (1 - y_i)\boldsymbol{x}_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) - y_i \boldsymbol{x}_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \\ &= \sum_{i=1}^{N} \boldsymbol{x}_i \frac{(1 - y_i) - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 - \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \ . \end{aligned} \tag{6}$$

(Received ????)

Jiří Vomlel, Institute of Information Theory and Automation of the AS CR, Pod vodárenskou věží 4, 18208 Praha 8, Czech Republic
e-mail: vomlel@utia.cas.cz