# DATA REGULARIZATION

**Norbert Jankowski**[1]

Department of Computer Methods
Nicholas Copernicus University
ul. Grudziądzka 5, 87-100 Toruń, Poland

**Abstract:**
Quite often real-world data set contain errors and inaccuracies. Most classification models are trained using crisp, sharply classified (*black* and *white*) examples only. In many real world problems the soft class labels (shades of *gray*) are quite natural. In this paper data regularization method has been presented. The method may help to strengthen the confidence in a given data set. Further data processing (learning) may become more stable and may lead to more reliable results.

## 1 INTRODUCTION

The dilemma that the adaptive models face is: the system *must* believe in the training data, but the data may not be dependable. Most classification methods do not solve adequately problems related to wrong data, sharp decisions borders caused by *black* and *white* class labeling, or by overlapping clusters. Some models try to solve this problem using several kinds of regularization methods during the learning process. Most regularization methods add a penalty term to the error function, for example regularization proposed by Poggio and Girosi [3], Hinton's *weight decay* [2] and *weight elimination* proposed by Weigend [4]. However, even using regularization methods problems mentioned above do not vanish. One of the reasons is that regularization methods have (almost) the same sensitivity in whole input space.

Moreover, even experts in most cases are not able to check the data vectors and assign to each case uncertainty which could help to add powerful information to the learning process. Well-known data preprocessing methods do not repair the data too.

In the next section the data regularization method is presented with a few variants which may be useful for different models based on the main concept. The regularization scheme gives in a natural way the measure of uncertainty of original data. In the last section empirical examples are shown.

## 2 DATA REGULARIZATION

The typical goal of classification is to find an underlaying mapping:

$$f(\mathbf{x}_i) = y_i, \quad i = 1, 2, \ldots, N \tag{1}$$

for given data set $\mathcal{S}$:

$$\mathcal{S} = \{\langle \mathbf{x}_i, y_i \rangle : 1 \le i \le N\} \tag{2}$$

[1] E-mail address: Norbert.Jankowski@phys.uni.torun.pl, and www is: http://www.phys.uni.torun.pl/˜norbert

where each pair $\langle \mathbf{x}_i, y_i \rangle$ consists of input vector $\mathbf{x}_i$ and class label $y_i$. It is welcome for some classification models to have the class label $y_i$ represented by vector $\mathbf{v}_i$ with 1 on position numer equal to the class label $y_i$ and rests equal to 0 (for example multi-layered perceptron networks):

$$\mathbf{v}_i = [v_1, v_2, \ldots, v_d]^T \quad \text{and} \quad v_k = \begin{cases} 1 & k = y_i \\ 0 & k \neq y_i \end{cases} \tag{3}$$

then data set consists of pairs of vectors:

$$\mathcal{S}^v = \{\langle \mathbf{x}_i, \mathbf{v}_i \rangle : 1 \leq i \leq N\} \tag{4}$$

Basing on data set $\mathcal{S}$ it is possible to define model $\mathcal{P}$ using renormalized Gaussian function:

$$\bar{G}_i(\mathbf{x}; \mathbf{x}_i) = \frac{G(\mathbf{x}; \mathbf{x}_i, \sigma)}{\sum_{j=1}^{N} G(\mathbf{x}; \mathbf{x}_j, \sigma)} \tag{5}$$

where $G(\mathbf{x}; \mathbf{x}_i, \sigma)$ ($\sigma$ is constant) is defined by

$$G(\mathbf{x}; \mathbf{x}_i, \sigma) = e^{-\frac{||\mathbf{x}-\mathbf{x}_i||^2}{\sigma}} \tag{6}$$

then model $\mathcal{P}$ may be defined by

$$P(k|\mathbf{x}, \mathcal{S}) = \sum_{i \in I^k} \bar{G}_i(\mathbf{x}; \mathbf{x}_i) \tag{7}$$

where $I^k = \{i : \langle \mathbf{x}_i, y_i \rangle \in \mathcal{S} \wedge y_i = k\}$. We can see that

$$\sum_{i=1}^{K} P(i|\mathbf{x}, \mathcal{S}) = 1 \tag{8}$$

$K$ is equal to the class number. Then $P(k|\mathbf{x}, \mathcal{S})$ may be interpreted as probability that given vector $\mathbf{x}$ belong to class $k$ for data set $\mathcal{S}$.

Note that parameter $\sigma$ from Eq. 6 defines the smoothness of model $\mathcal{P}$. Assuming that $\sigma$ is sufficiently small

$$P(i|\mathbf{x}_i, \mathcal{S}) \approx 1 \tag{9}$$

Suppose data set $\mathcal{S}$ is not very *fragile* (is sufficiently dense) and removing a single pair from data set $\mathcal{S}$ model $\mathcal{P}$ should not change *crucially* for most pairs. Let $\mathcal{S}^j$ design set $\mathcal{S}$ with subtracted pair $\langle \mathbf{x}_j, y_j \rangle$ ($\mathcal{S}^j = \{\langle \mathbf{x}_k, y_k \rangle : \langle \mathbf{x}_k, y_k \rangle \in \mathcal{S} \wedge k \neq j\}$).

Now using probability

$$P(i|\mathbf{x}_i, \mathcal{S}^i) \tag{10}$$

the certainty that vector $\mathbf{x}_i$ is consistent with set $\mathcal{S}$ may be measured as a *consistence test*. Factor $\sigma$ (Eq. 6) which defines the smoothness of Gauss function may be used to control the regularization strength of model $\mathcal{P}$. The choosing of $\sigma$ should depend on the pre-uncertainty for set $\mathcal{S}$ or may be set to $D^2/N$ ($D$ is equal to the maximal distance between two vectors from set $\mathcal{S}$).

Consistence test may be used in several ways in data regularization. Two types of regularization arise from below sets as extension of set $\mathcal{S}$:

$$\mathcal{S}^P \quad = \quad \{\langle\langle \mathbf{x}_i, y_i\rangle, P(y_i|\mathbf{x}_i, \mathcal{S}^i)\rangle : 1 \le i \le N\} \tag{11}$$

$$\mathcal{S}^{Pv} \quad = \quad \{\langle\langle \mathbf{x}_i, y_i\rangle, P(1|\mathbf{x}_i, \mathcal{S}^i), \dots, P(K|\mathbf{x}_i, \mathcal{S}^i)\rangle : 1 \le i \le N\} \tag{12}$$

**Shades of gray.** Data set $\mathcal{S}$ consists from *black* nad *white* examples only. Now basing on above sets $\mathcal{S}^P$ and $\mathcal{S}^{Pv}$ data set with *shades of gray* may be produced:

$$\mathcal{S}^I = \{\langle \mathbf{x}_i, \langle y_i, P(y_i|\mathbf{x}_i, \mathcal{S}^i)\rangle\rangle : 1 \le i \le N\} \tag{13}$$

or in multi-non-zero output mode:

$$\mathcal{S}^{II} = \{\langle \mathbf{x}_i, \mathbf{p}_i\rangle : 1 \le i \le N\} \tag{14}$$

where

$$\mathbf{p}_i = [P(1|\mathbf{x}_i, \mathcal{S}^i), \dots, P(K|\mathbf{x}_i, \mathcal{S}^i)]^T \tag{15}$$

**Wrong pair elimination and class relabeling.** It is possible that for some vectors $P(y_i|\mathbf{x}_i, \mathcal{S}^i)$ is considerably smaller than $P(j|\mathbf{x}_i$ ($j \neq y_i$), what mean that pair $\langle \mathbf{x}_i, y_i\rangle$ is not consistent (wrong) with original set $\mathcal{S}$. Another possibility is to remove such wrong vectors from sets $\mathcal{S}^I$ and $\mathcal{S}^{II}$ (Eq. 13 and 14).

Wrong vector will not be labeled with original class label because of small probability value $P(y_i|\mathbf{x}_i, \mathcal{S}^i)$). And for set $\mathcal{S}^{II}$ each wrong vector $\mathbf{x}_i$ will be relabeled with more certain class:

$$\max_{j \neq i} P(j|\mathbf{x}_i, \mathcal{S}^i) \tag{16}$$

than with the original one.

In the case of a method must be used with *black* and *white* data the information from sets $\mathcal{S}^P$ and $\mathcal{S}^{Pv}$ may help to be excluded or relabeled wrong original pairs from set $\mathcal{S}$, for example to relabel wrong data set $\mathcal{S}^{III}$ may be useful:

$$\mathcal{S}^{III} = \{\langle \mathbf{x}_i, k\rangle : 1 \le i \le N\} \tag{17}$$

where $k = \arg\max_j P(j|\mathbf{x}_i, \mathcal{S}^i)$.

Such regularized data sets may be used to learning with different artificial neural networks (MLP, RBF, etc.) may be used to used in costs functions in many machine learning methods to add certainty weighting for each vectors (for example in CART model [ 1]).

# 3 EXAMPLE OF DATA REGULARIZATION

Simple and fruitful example may concern on regularization of two class data generated independently with Gaussian distribution.

Figures 1 and 2 presents data before (triangles — lower for class I and upper for class II) and after regularization (circles for class I, and crosses for class II). Two solid lines presents probability of model $\mathcal{P}$ for two classes defined by Eq. 7 on original set $\mathcal{S}$. Successive subfigures presents results for different dispersions and centers placements.

# 4 CONCLUSIONS

Data regularization method described in this paper may be successfully used in many different models used for classification. Such data regularization may assist the learning process, especially when the data requires sharp decision borders. Regularization is able to remove wrong data or to relabel some vectors. Data set class labels transformed to a set with *shades of gray* may stabilize the learning process. It may also be used to weight each vector's contribution to the cost function depending on the uncertainty of this vector.

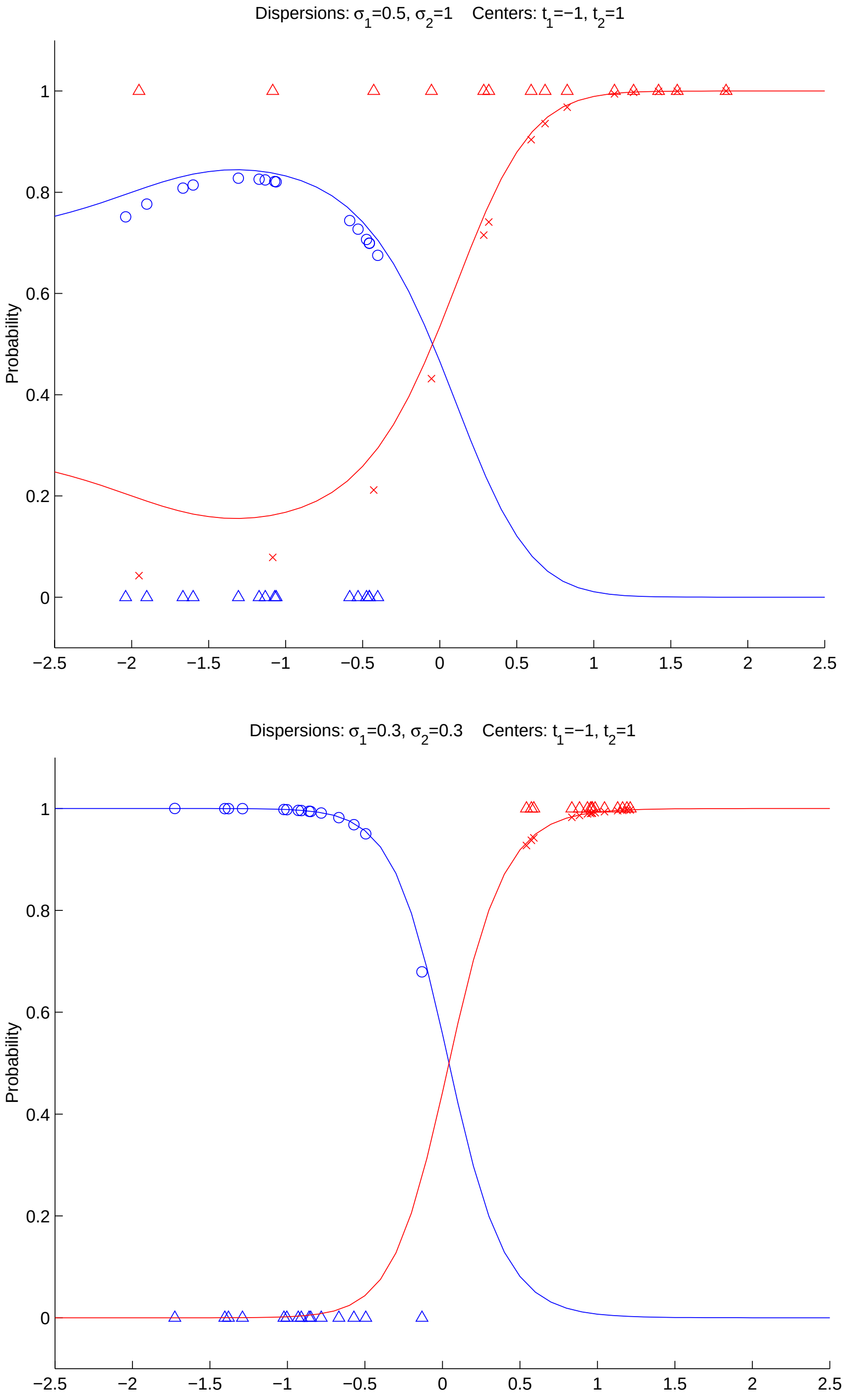
# REFERENCES

[1] L. Breiman, J. H. Friedman, A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth, Belmont, CA, 1984.

[2] G. E. Hinton. Learning translation invariant recognition in massively parallel networks. In J. W. de Bakker, A. J. Nijman, and P. C. Treleaven, editors, *Proceedings PARLE Conference on Parallel Architectures and Languages Europe*, pages 1–13, Berlin, 1987. Springer-Verlag.

[3] T. Poggio and F. Girosi. Network for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

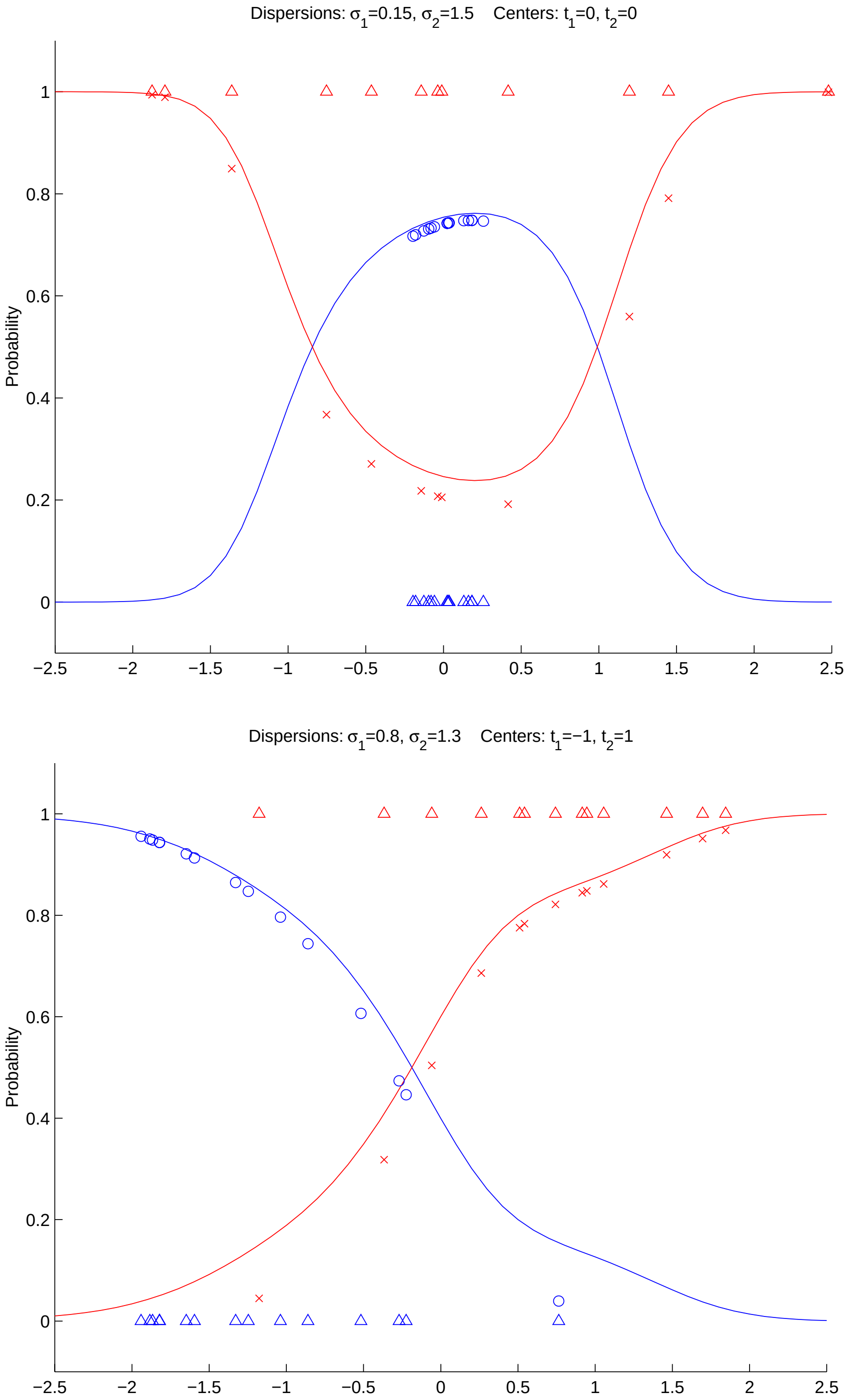
[4] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman. Generalization by weight elimination with application to forecasting. In R. P. Lipmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 875–882, San Mateo, CA, 1991. Morgan Kaufmann.



Dispersions: $\sigma_1$=0.5, $\sigma_2$=1    Centers: $t_1$=−1, $t_2$=1

Dispersions: $\sigma_1$=0.3, $\sigma_2$=0.3    Centers: $t_1$=−1, $t_2$=1

Figure 1: Data regularization I.



Dispersions: $\sigma_1$=0.15, $\sigma_2$=1.5    Centers: $t_1$=0, $t_2$=0

Dispersions: $\sigma_1$=0.8, $\sigma_2$=1.3    Centers: $t_1$=−1, $t_2$=1

Figure 2: Data regularization I.