Locality Preserving Multi-nominal Logistic Regression

Kenji Watanabe[1] and Takio Kurita[2]
[1] Department of Computer Science, Graduate School of Systems and Information Engineering,
University of Tsukuba
[2] National Institute of Advanced Industrial Science and Technology (AIST)
[1] kenji-watanabe@aist.go.jp, [2] takio-kurita@aist.go.jp

Abstract

In this paper, we propose a novel algorithm of multi-nominal logistic regression in which the locality regularization term is introduced. The locality is defined by the neighborhood information of the data set and is preserved in the mapped feature space. By using the standard benchmark datasets, it was shown that the proposed algorithm gave higher recognition rates than the linear SVM in binary classification problems. The recognition rates for multi-class classification problem were also better than the general multi-nominal logistic regression.

1. Introduction

Logistic regression (LR) is one of the well-known binary classification methods and is often used for biological signals, such as electro encephalography (EEG) [1]. Multi-nominal logistic regression (MLR) is a natural extension of LR to multi-class classification problems. To improve the generalization performance of these methods, the regularization term is often introduced by giving penalty to unnecessary growth of the parameter values [2, 3].

He et al. [4, 5] proposed locality preserving projections (LPP) and applied it for face recognition because the structure in the original feature space should be reflected in the mapped space as much as possible. In LPP, the manifold structure is modeled by a nearest-neighbor graph which preserves the local structure of the original feature space.

In classification design, the structure in the original feature space should be considered. In this paper, we introduce the locality in the regularization term of MLR and the regularized MLR by shrinkage. We call these algorithms Locality Preserving Multi-nominal Logistic Regression (LPMLR) and Multi-nominal Logistic Regression regularized by Locality Preserving and Shrinkage (LPSMLR). By using the standard benchmark datasets, it was shown that the proposed algorithm gave higher recognition rates than the linear SVM in binary classification problems. The recognition rates for multi-class classification problem were also better than the general MLR.

2. Multi-nominal Logistic Regression

LR is a model used for prediction of the probability of occurrence of an event. It makes use of several predictor variables that may be either numerical or categories. Its natural extension to multi-class classification problems is MLR. In this section, we represented MLR and the general regularized MLR.

2.1. Multi-nominal Logistic Regression

For K-class classification problem, let $D = \{(\boldsymbol{x}_i, \boldsymbol{u}_i)\}_{i=1}^{N}$ be a given training data, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{im}) \in \mathbf{X} \in \Re^{N \times m}$ is the i-th input vector, and $\boldsymbol{u}_i \in \boldsymbol{U} = \{\boldsymbol{u} \mid \boldsymbol{u} \in \{0,\ 1\}^k,\ \|\boldsymbol{u}\|_{L1} = 1\}$ is the k-th class label vector for the i-th input vector. The outputs of MLR estimate the posterior probabilities $p(u_i^k \mid \boldsymbol{x}_i)$. They are defined as follows:

$$p(u_i^k \mid \boldsymbol{x}_i) = y_i^k = \frac{exp(\eta_i^k)}{\sum_{j=1}^{K} exp(\eta_i^j)} = \frac{exp(\eta_i^k)}{1 + \sum_{j=1}^{K-1} exp(\eta_i^j)}, \qquad (1)$$

$$\eta_i^k = \boldsymbol{x}_i \hat{\boldsymbol{w}}^k + b_k = \hat{\boldsymbol{x}}_i \boldsymbol{w}^k, \qquad (2)$$

where $\hat{\boldsymbol{w}}^{k^{\mathrm{T}}} = (w_{1k}, \cdots, w_{mk})$ and b_k are the weight vector and the bias term of k-th class, respectively. To simplify the notation, we include the bias term in the vectors as $\boldsymbol{w}^{k^{\mathrm{T}}} = (w_{1k}, \cdots, w_{mk}, b_k)$ and $\hat{\boldsymbol{x}}_i = (x_{i1}, \cdots, x_{im}, 1)$. In matrix notation, we use $\boldsymbol{W}^{\mathrm{T}} = \left(\boldsymbol{w}^{1^{\mathrm{T}}}, \cdots, \boldsymbol{w}^{K-1^{\mathrm{T}}}\right) \in \Re^{(K-1)M}$ and $\hat{\boldsymbol{X}}^{\mathrm{T}} = \left(\hat{\boldsymbol{x}}_1^{\mathrm{T}}, \cdots, \hat{\boldsymbol{x}}_N^{\mathrm{T}}\right) \in \Re^{N \times M}$. The optimal parameters of MLR are obtained by minimizing the negative log-likelihood

$$\boldsymbol{W} = arg\,\min_{\boldsymbol{w}} E_D, \tag{3}$$

$$E_D = \sum_{i=1}^{N} \sum_{j=1}^{K-1} \left\{ u_i^j\, log\left(1 + \sum_{l=1}^{K-1} exp\left(\eta_i^l\right) \right) - u_i^j \eta_i^j \right\}. \tag{4}$$

Equation (3) represents a convex optimization problem and it has only a single, global minimum. Again the optimal parameter $\boldsymbol{W}$ can be efficiently found using Newton-Raphson method or an iterative re-weighted least squares (IRLS) procedure. Here we show IRLS for MLR. In each iteration step, $\boldsymbol{W}$ is updated by

$$\boldsymbol{W}^{t+1} = \boldsymbol{H}^{-1} \boldsymbol{G}^{\mathrm{T}} \boldsymbol{Z}, \tag{5}$$

where $\boldsymbol{H} = \boldsymbol{G}^T \boldsymbol{R} \boldsymbol{G} \in \Re^{(K-1)M \times (K-1)M}$ is the block Hessian matrix, and $\boldsymbol{H}^{-1}$ is the inverse matrix of $\boldsymbol{H}$. $\boldsymbol{G} = diag\left(\hat{\boldsymbol{X}}^1, \cdots, \hat{\boldsymbol{X}}^{K-1}\right) \in \Re^{(K-1)N \times (K-1)M}$ is the block diagonal matrix of $\hat{\boldsymbol{X}}$, and $\hat{\boldsymbol{X}}^k = \hat{\boldsymbol{X}}$. $\boldsymbol{R} \in \Re^{(K-1)N \times (K-1)N}$ is the block matrix defined as follows:

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \cdots & \boldsymbol{R}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{R}_{(K-1)1} & \cdots & \boldsymbol{R}_{(K-1)(K-1)} \end{pmatrix}, \tag{6}$$

$$\boldsymbol{R}_{jk} = diag\left(r_1^{jk}, \ \cdots, \ r_N^{jk}\right), \tag{7}$$

$$r_n^{jk} = y_n^j\left(\delta_{jk} - y_n^k\right), \quad \delta_{jk} = \begin{cases} 1 & if \quad j = k \\ 0 & otherwise \end{cases}. \tag{8}$$

$\boldsymbol{Z} \in \Re^{(K-1)N}$ is the block vector with elements

$$z_k = \sum_{j=1}^{K-1} \boldsymbol{R}_{kj} \boldsymbol{\eta}^j - \left(\boldsymbol{y}^k - \boldsymbol{u}^k\right). \tag{9}$$

Equation (5) is repeated until convergence.

2.2. Regularization by shrinkage

In general, the regularization term is introduced to control the over-fitting. In shrinkage method, unnecessary growth of the parameters is penalized by introducing the regularization term E_W defined as follows:

$$E_W = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} = \sum_{j=1}^{K-1} \sum_{k=1}^{K-1} \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{w}_k. \tag{10}$$

In this case, the regularized MLR is trained by minimizing the negative log-likelihood as

$$\boldsymbol{W} = arg\,\min_{\boldsymbol{w}} \left(E_D + \lambda_{\boldsymbol{w}} E_W\right). \tag{11}$$

Equation (11) represents a convex optimization problem, and λ_w is the regularization parameter of E_W. IRLS to calculate $\boldsymbol{W}$ are modified as follows:

$$\boldsymbol{W}^{t+1} = \boldsymbol{H}^{-1} \boldsymbol{G}^{\mathrm{T}} \boldsymbol{Z}, \tag{12}$$

where $\boldsymbol{H} \in \Re^{(K-1)M \times (K-1)M}$ is the block Hessian matrix defined as follows:

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_{11} & \cdots & \boldsymbol{H}_{1(K-1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_{(K-1)1} & \cdots & \boldsymbol{H}_{(K-1)(K-1)} \end{pmatrix}, \tag{13}$$

$$\boldsymbol{H}_{jk} = \begin{cases} \hat{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{R}_{jk} \hat{\boldsymbol{X}} + 2\lambda_w \boldsymbol{I} & if \quad j = k \\ \hat{\boldsymbol{X}}^{\boldsymbol{T}} \boldsymbol{R}_{jk} \hat{\boldsymbol{X}} + \lambda_w \boldsymbol{I} & otherwise \end{cases}. \tag{14}$$

$\boldsymbol{I} \in \Re^{M \times M}$ is the identity matrix. $\boldsymbol{R} \in \Re^{(K-1)N \times (K-1)N}$ is the block matrix similar to Equation (6), (7) and (8). $\boldsymbol{Z} \in \Re^{(K-1)N}$ is the block vector with elements

$$\boldsymbol{Z} = \boldsymbol{R}\boldsymbol{\eta} - \left(\boldsymbol{y} - \boldsymbol{u}\right). \tag{15}$$

3. Locality preserving mapping

Locality preserving projection (LPP) was proposed to model the local manifold structure [5]. LPP is a new linear dimensionality reduction algorithm, and it builds a graph incorporating neighborhood information of the data set. He et al. [6] showed that locality is effective for face recognition.

3.1. Locality

LPP is a linear approximation of the nonlinear Laplacian Eigenmap [7]. In this paper, we use the same pair-wise locality with LPP to design the regularization term of MLR. The pair-wise locality Q_{ij}^{kl} is defined as follows:

$$Q_{ij}^{kl} = u_i^k u_j^l\, exp\left(\frac{-\left\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j\right\|_{L2}^2}{\tau} \right), \tag{16}$$

where τ is the hyper parameter and it was tuned by the grid search. The pair-wise locality Q_{ij}^{kl} is equal to zero when $k \neq l$. In matrix notation, we define the block locality matrix $\boldsymbol{Q} \in \Re^{(K-1)N \times (K-1)N}$ as follows:

$$\boldsymbol{Q} = \mathrm{diag}\left(\boldsymbol{Q}^{1}, \ \cdots, \ \boldsymbol{Q}^{K-1}\right), \qquad (17)$$

$$\boldsymbol{Q}^{k} = \begin{pmatrix} Q_{11}^{kk} & \cdots & Q_{1N}^{kk} \\ \vdots & \ddots & \vdots \\ Q_{N1}^{kk} & \cdots & Q_{NN}^{kk} \end{pmatrix}. \qquad (18)$$

3.2. Locality Preserving Multi-nominal Logistic Regression

In this paper, we introduced the regularization term E_{LP} by using the pair-wise localities as follows:

$$E_{LP} = \sum_{i}^{N}\sum_{j}^{N}\sum_{k}^{K-1}\left(\eta_i^k - \eta_j^k\right)^2 Q_{ij}^{kk} \ . \qquad (19)$$

In this case, the parameters of this regularized MLR, namely locality preserving multi-nominal logistic regression (LPMLR), is trained by minimizing the negative log-likelihood

$$\boldsymbol{W} = arg\,\underset{\boldsymbol{w}}{min}\left(E_D + \lambda_{LP}E_{LP}\right), \qquad (20)$$

where λ_{LP} is the regularization parameter of E_{LP} . We can calculate the optimal $\boldsymbol{W}$ by using IRLS using the update rule

$$\boldsymbol{W}^{t+1} = \boldsymbol{H}^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{Z} \ . \qquad (21)$$

Where $\boldsymbol{H} \in \Re^{(K-1)M\times(K-1)M}$ is the block Hessian matrix defined by

$$\boldsymbol{H} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{G} + 4\lambda_{LP}\boldsymbol{G}^{\mathrm{T}}\left(\boldsymbol{S} - \boldsymbol{Q}\right)\boldsymbol{G} \ . \qquad (22)$$

Where $\boldsymbol{R} \in \Re^{(K-1)N\times(K-1)N}$ is the block matrix similar to Equation (6), (7) and (8). $\boldsymbol{S} \in \Re^{(K-1)N\times(K-1)N}$ is the block diagonal matrix obtained from $\boldsymbol{Q}$ defined as follows:

$$\boldsymbol{S} = \mathrm{diag}\left(\boldsymbol{S}^{1}, \ \cdots, \ \boldsymbol{S}^{K-1}\right), \qquad (23)$$

$$\boldsymbol{S}^{k} = \mathrm{diag}\left(s_1^k, \ \cdots, \ s_N^k\right), \qquad (24)$$

$$s_i^k = \sum_{j=1}^{N} Q_{ij}^{kk} \ . \qquad (25)$$

$\boldsymbol{Z} \in \Re^{(K-1)N}$ is the block vector with elements similar to Equation (15).

3.3. Regularization by locality and shrinkage

We also introduced the regularization term E_{LP} to the regularized MLR. In this case, the parameters of this regularized MLR, namely multi-nominal logistic regression regularized by locality preserving and shrinkage (LPSMLR), is trained by minimizing the criterion

$$\boldsymbol{W} = arg\,\underset{w}{min}\left(E_D + \lambda_w E_W + \lambda_{LP}E_{LP}\right), \qquad (26)$$

where λ_w and λ_{LP} are the regularization parameters of E_W and E_{LP} , respectively. We can calculate the optimal $\boldsymbol{W}$ by using IRLS defined as the update rule

$$\boldsymbol{W}^{t+1} = \boldsymbol{H}^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{Z} \ . \qquad (27)$$

Where $\boldsymbol{H} \in \Re^{(K-1)M\times(K-1)M}$ is the block Hessian matrix defined by

$$\boldsymbol{H}_{jk} = \begin{cases} \hat{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_{jk}\hat{\boldsymbol{X}} + 2\lambda_w\boldsymbol{I} & \\ \quad + 4\lambda_{LP}\hat{\boldsymbol{X}}^{\mathrm{T}}\left(\boldsymbol{S}^k - \boldsymbol{Q}^k\right)\hat{\boldsymbol{X}} & if \quad j = k \\ \hat{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{R}_{jk}\hat{\boldsymbol{X}} + \lambda_w\boldsymbol{I} & otherwise \end{cases}, \qquad (28)$$

where $\boldsymbol{R} \in \Re^{(K-1)N\times(K-1)N}$ is the block matrix similar to Equation (6), (7) and (8). $\boldsymbol{Q}^k$ and $\boldsymbol{S}^k$ are the locality matrix similar to Equation (18) and the diagonal matrix similar to Equation (24), respectively. The vector $\boldsymbol{Z} \in \Re^{(K-1)N}$ is the block vector with elements similar to Equation (15).

4. Experiments

To confirm the effectiveness of the proposed algorithm, the recognition rates are compared using the standard benchmark datasets for binary classification and multi-class classification. Table 1 shows a summary of these datasets. They are German, Heart, Satimage, Segment and Ionosphere [9]. The training and test samples were randomly selected except for Satimage. For binary classification problems, we evaluated the recognition rates by Linear Support vector machine (SVM), Linear Discriminant Analysis (LDA), LR and the regularized LRs. On the other hand, LDA and the regularized MLRs are compared for multi-class classification problems. The hyper parameters of the regularized MLRs were tuned by the grid search. The parameter λ_w of LPSMLR was set to the best value obtained by the grid search for the regularized MLR, because we are investigating the effect of locality as the regularization.

Table 1. Summary of the benchmark datasets

	Class	# of training	# of test	# of feature
German	2	400	600	24
Heart	2	140	130	13
Ionosphere	2	180	170	34
Satimage	6	4435	2000	36
Segment	7	1400	910	19

4.1. Logistic Regression

The recognition rates of the two-class benchmark datasets are shown in Table 2. For SVM we used libSVM [8]. The recognition rates of LDA were calculated by using *k* nearest neighbor (*k*-NN) classifier in LDA subspace. The cost parameter of SVM was tuned by the grid search.

From Table 2, it is noticed that the recognition rates obtained by LDA and MLR without regularization are comparable to SVM. The proposed LPMLR and LPSMLR give the better recognition rates than SVM. Especially LPSMLR gives the best recognition rate for Heart, and the regularization term by locality is effective for Ionosphere datasets. These results suggest that LPMLR and LPSMLR can give better recognition performance than other methods. This means that the locality can be useful for regularization in classifier design.

4.2. Multi-nominal Logistic Regression

When solving a multi-class classification problem, general linear discriminative models, such as LDA and MLR, are able to treat multi-classes by one model. The recognition rates of the multi-class benchmark datasets are shown in Table 3. The regularized MLR gives the highest recognition rate for Satimage. The recognition rates by the proposed LPMLR and LPSMLR are better than MLR, and LPMLR gives the best recognition rate for Segment. These show that the effectiveness of locality is affected to the properties of datasets. Since the parameter λ_w of LPSMLR was fixed to the best value obtained by the grid search for the regularized MLR, the recognition rate by LPSMLR becomes lower than the regularized MLR and LPMLR for Segment.

5. Conclusion and Feature works

This paper proposed a novel algorithm of multi-nominal logistic regression in which the locality regularization term is introduced. By using the standard benchmark datasets, it was shown that the proposed algorithm gave higher recognition rates than the other methods. But we find that the effectiveness of locality is probably affected to the properties of the datasets. As future works, we should investigate in which types of datasets the regularization by locality preserving is effective.

References

[1] R. Tomioka, K. Aihara and K.-R. Müller. Logistic Regression for Single Trial EEG Classification. *Advances in Neural Information Processing Systems*, 19: 1377-1384, 2007.

[2] G.C. Cawley, N.L.C. Talbot and M. Girolami. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. *Advances in Neural Information Processing Systems*, 19: 209-216, 2007.

[3] B. Krishnapuram and M.A.T. Figueiredo. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(6): 957-968, 2005.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using Laplacianfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3): 328–340, 2005.

[5] X. He and P. Niyogi. Locality Preserving Projections. *Advances in Neural Information Processing Systems*, 16:153–160, 2004.

[6] H. Wang, W. Zheng, Z. Hu and S. Chen. Local and Weighted Maximum Margin Discriminant Analysis. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[7] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Advances in Neural Information Processing Systems*, 14, 2002.

[8] C.C. Chang and C.J. Lin. LIBSVM : a library for support vector machines. *http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[9] A. Asuncion and D.J. Newman. UCI Machine Learning Repository. *http://www.ics.uci.edu/~mlearn/MLRepository.html*. 2007.

Table 2. Recognition rates of the two-class benchmark datasets

	SVM	LDA	MLR	Regularized MLR	LPMLR	LPSMLR
German	0.7067	0.7267	0.7233	**0.7317**	0.7233	**0.7317** *
Heart	0.8385	0.8385	0.8462	0.8538	0.8538	**0.8615**
Ionosphere	0.9064	0.9123	0.8947	0.9006	**0.9474**	0.9240

*: $\lambda_{LP} = 0$

Table 3. Recognition rates of the multi-class benchmark datasets

	LDA	MLR	Regularized MLR	LPMLR	LPSMLR
Satimage	0.8395	0.8375	**0.8435**	0.8385	**0.8435** *
Segment	0.9077	0.8198	0.9176	**0.9220**	0.8879

*: $\lambda_{LP} = 0$