数理解析研究所講究録 1320 巻 2003 年 239-249

Penalized Logistic Regression Machines and Related Linear Numerical Algebra

統計数理研究所　田邊国士 (Kunio Tanabe)
The Institute of Statistical Mathematics
and
Graduate University for Advanced Studies

1 Introduction

In the previous papers [1,2], the author introduced *Penalized Logistic Regression Machines* for multiclass discrimination. The machines are intended to handle noisy stochastic data and it was shown that *by penalizing the likelihood in a specific way, we can intrinsically combine the logistic regression model with the kernel methods.* In particular, *a new class of penalty functions and associated normalized projective kernels were introduced to gain a versatile induction power of the learning machines.* The purpose of this note is to show the use of the conjugate gradient methods for solving the nonlinear equations arises in the learning process by Newron's method. In the section 2 through 8, we summarize the previous papers[1,2] and the new CG-Newton methods are introduced in the section 6 and 8.

2 Multiclass Discrimination Problem

Let us consider the problem of multiclass discrimination given a finite number of training data set $\{(\mathbf{x}_i, c_i)\}_{i=1,...,N}$, where $\mathbf{x}_i$ is a column vector of size n whose elements may be both continuous and discrete numbers and c_i takes a value in the finite set $\{1, 2, ..., K\}$ of classes. We are concerned with the construction of i) a conditional multinomial distribution $\mathcal{M}(\mathbf{p}^*(\mathbf{x}))$ of c given $\mathbf{x} \in \mathbf{R}^n$, where $\mathbf{p}^*(\mathbf{x})$ is a predictive probability vector whose k-th element $p_k^*(\mathbf{x})$ indicates the probability of c taking the value k, and also ii) prediction function $c = d^*(\mathbf{x}) \in \{1, 2, ..., K\}$, which are used respectively as stochastic and deterministic prediction of c given $\mathbf{x}$. Let $\mathbf{e}_k \equiv (0, ..., 0, 1, 0, ..., 0))^t$ be the k-th unit column vector of size K, and let the $K \times N$ constant matrix $\mathbf{Y}$ be defined by $\mathbf{Y} \equiv [\mathbf{y}_1;\ \mathbf{y}_2;\ \cdots;\ \mathbf{y}_N] \equiv [\mathbf{e}_{c_1};\ \mathbf{e}_{c_2};\ \cdots;\ \mathbf{e}_{c_N}]$ whose j-th column vector $\mathbf{y}_j \equiv \mathbf{e}_{c_j}$ indicates which class the data c_j belongs to.

3 Normalized Projective Kernels

Following the idea of the Support Vector Machines, we introduce a map

$$\phi(\mathbf{x}, \lambda) \equiv (\phi_1(\mathbf{x}, \lambda), \phi_2(\mathbf{x}, \lambda), ..., \phi_m(\mathbf{x}, \lambda))^t \tag{1}$$

from $\mathbf{R}^n$ into $\mathbf{R}^m$, where λ is a hyperparamter vector. For $\phi(\mathbf{x}, \lambda)$, we can choose a set $\{\phi_m(\mathbf{x}, \lambda)\}_{i=1,2,\ldots,m}$ of arbitrarily many functions. We will drop the argument λ for notational simlicity. Let $\mathbf{\Phi}$ be the $m \times N$ constant matrix defined by

$$\mathbf{\Phi} \equiv [\,\phi(\mathbf{x}_1);\quad \phi(\mathbf{x}_2);\quad \cdots;\quad \phi(\mathbf{x}_N)\,] \tag{2}$$

whose j-th column vector is $\phi(\mathbf{x}_j)$. In order to gain a versatile induction power of the resulting method we always include the constant function $\phi_0(\mathbf{x}) \equiv \omega$ as a member of the regressors, and let the associated augmented map $\overline{\phi}(\mathbf{x})$ and the $(m+1) \times N$ constant matrix $\overline{\mathbf{\Phi}}$ be defined respectively by

$$\overline{\phi}(\mathbf{x}) \equiv \begin{pmatrix} \omega \\ \phi(\mathbf{x}) \end{pmatrix}, \textit{and} \quad \overline{\mathbf{\Phi}} \equiv \begin{bmatrix} \omega \mathbf{1}_N^t \\ \mathbf{\Phi} \end{bmatrix} \tag{3}$$

where ω is a fixed nonnegative constant to be considered as a hyperparameter. s In order to prepare for the arguments in the subsequent sections, we need to introduce a *Normalized Projective Kernel* function $\mathcal{K}_\omega(\mathbf{x}, \mathbf{y})$ of two arguments $(\mathbf{x}, \mathbf{y}) \in \mathbf{R}^n \times \mathbf{R}^n$, defined by the bilinear form

$$\begin{aligned} \mathcal{K}_\omega(\mathbf{x}, \mathbf{y}) \equiv \overline{\phi}^t(\mathbf{x}) \overline{\mathbf{\Sigma}}^{-1} \overline{\phi}(\mathbf{y}) &\equiv \frac{1}{\sigma^2}(\omega^2 + (\sigma\phi(\mathbf{x}) - \omega\mu)^t \mathbf{\Sigma}^{-1}(\sigma\phi(\mathbf{y}) - \omega\mu)), \\ &\equiv \phi_0^2 + (\phi(\mathbf{x}) - \phi_0\mu)^t \mathbf{\Sigma}^{-1}(\phi(\mathbf{y}) - \phi_0\mu)), \end{aligned} \tag{4}$$

of the maps $\overline{\phi}(\mathbf{x})$ and $\overline{\phi}(\mathbf{y})$, where $\overline{\phi}^t(\mathbf{x})$ is the transpose of the column vector $\overline{\phi}(\mathbf{x})$ and $\overline{\mathbf{\Sigma}}$ is an $(m+1) \times (m+1)$ positive definite matrix which is parametrized as

$$\overline{\mathbf{\Sigma}} \equiv \begin{pmatrix} \sigma \\ \mu \end{pmatrix} \begin{pmatrix} \sigma \\ \mu \end{pmatrix}^t + \begin{bmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \equiv \begin{bmatrix} \sigma^2 & \sigma\mu^t \\ \sigma\mu & \mathbf{\Sigma} + \mu\mu^t \end{bmatrix}. \tag{5}$$

by a scalar σ, a column vector μ and an $m \times m$ positive definite matrix $\mathbf{\Sigma}$, which are considered to be hyperparameters, and $\phi_0 \equiv \omega/\sigma$. Note that any positive definite matrix $\overline{\mathbf{\Sigma}}$ can be put in this form, because the middle part of Eq.(5) is an m-step-premature Cholesky decomposition of $\overline{\mathbf{\Sigma}}$.

It will be shown in Sections 6 and 7 that with a penalized logistic regression model given in the next section we could work directly with the kernel function without resorting explicitly to the map $\overline{\phi}(\mathbf{x})$ and the parameter $\overline{\mathbf{\Sigma}}$ themselves. In fact, we only need the $N \times N$ constant matrix

$$\mathcal{K}_\omega^d \equiv [\mathcal{K}_\omega(\mathbf{x}_i, \mathbf{x}_j)] \equiv \overline{\mathbf{\Phi}}^t\, \overline{\mathbf{\Sigma}}^{-1} \overline{\mathbf{\Phi}} \equiv \phi_0^2 \mathbf{1}_N \mathbf{1}_N^t + (\mathbf{\Phi} - \phi_0\mu\mathbf{1}_N^t)^t\, \mathbf{\Sigma}^{-1}(\mathbf{\Phi} - \phi_0\mu\mathbf{1}_N^t) \tag{6}$$

and the map $\kappa_\omega(\mathbf{x})$ from $\mathbf{R}^n$ into $\mathbf{R}^N$ defined by

$$\kappa_\omega(\mathbf{x}) \equiv \overline{\mathbf{\Phi}}^t\, \overline{\mathbf{\Sigma}}^{-1} \overline{\phi}(\mathbf{x}) \equiv \phi_0^2 \mathbf{1}_N + (\mathbf{\Phi} - \phi_0\mu\mathbf{1}_N^t)^t\, \mathbf{\Sigma}^{-1}(\phi - \phi_0\mu) \tag{7}$$

There are many possibilities for choosing the hyperparameter μ. A typical choice is $\mu = N^{-1}\mathbf{\Phi}^t\mathbf{1}_N = N^{-1}\sum_{j=1}^N \phi(\mathbf{x}_j)$. We briefly touch on the relationship of the normalized projective kernel to the ordinary (but normalized) kernel $\mathcal{K}_0(\mathbf{x}, \mathbf{y}) = \phi^t(\mathbf{x})\, \mathbf{\Sigma}^{-1}\phi(\mathbf{y})$. In

particular, we relate the matrix in Eq.(6) and the function in Eq.(7) to the existing ones. If μ is chosen as above, then

$$\mathcal{K}_\omega^d = N\phi_0^2\mathbf{\Pi} + (\mathbf{I}_N - \phi_0\mathbf{\Pi}_N)\,\mathcal{K}_0^d(\mathbf{I}_N - \phi_0\mathbf{\Pi}_N),$$

$$\kappa_\omega(\mathbf{x}) = \phi_0^2\mathbf{1}_N + (\mathbf{I}_N - \phi_0\mathbf{\Pi}_N)(\kappa_0(\mathbf{x}) - N^{-1}\phi_0\mathcal{K}_0^d\mathbf{1}_N)$$

where $\mathbf{\Pi}_N \equiv N^{-1}\mathbf{1}_N\mathbf{1}_N^t$ is the orthogonal projector onto the space spanned by $\mathbf{1}_N$. These equations can be used as formulas for converting the existing kernels to normalized projective kernels.

4 Penalized Logistic Regression Models

In order to solve the problem mentioned earlier, we introduce the penalized logistic regression model. We assume that the joint probability distribution $\zeta(\mathbf{x},\mathbf{y})$ of $(\mathbf{x},\mathbf{y})$ from which the training data is sampled is unknown and that the conditional distribution $\zeta(\mathbf{y}|\mathbf{x})$ of $\mathbf{y}$ given $\mathbf{x}$, follows the multinomial distribution $\mathcal{M}(\mathbf{p}(\mathbf{x}))$ specified by the probability vector

$$\mathbf{p}(\mathbf{x}) \equiv (p_1(\mathbf{x}), p_2(\mathbf{x}), ..., p_K(\mathbf{x}))^t \equiv \hat{\mathbf{p}}(\mathbf{f}(\mathbf{x})), \tag{8}$$

which is parametrized by the logistic transform $\hat{\mathbf{p}}(\mathbf{f})$, due to Leonard(1973), of the affine transformation $\mathbf{f}(\mathbf{x})$ of the map $\phi(\mathbf{x})$, where

$$\hat{\mathbf{p}}(\mathbf{f}) \equiv (\hat{p}_1(\mathbf{f}), \hat{p}_2(\mathbf{f}), ..., \hat{p}_K(\mathbf{f}))^t, \quad \text{where } \hat{p}_k(\mathbf{f}) \equiv expf_k(\sum_{i=1}^{K} expf_i)^{-1}, \tag{9}$$

$$\mathbf{f}(\mathbf{x}) \equiv (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_K(\mathbf{x}))^t \equiv \omega\mathbf{w}_0 + \mathbf{W}\phi(\mathbf{x}) \equiv \overline{\mathbf{W}}\,\overline{\phi}(\mathbf{x}), \tag{10}$$

$\mathbf{w}_0$ and $\mathbf{W}$ are respectively an $K \times 1$ parameter vector and an $K \times m$ parameter matrix to be estimated from the given training data, $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,...,N}$ as $\mathbf{y}_i$ given in Eq.(??), and the $K \times (m+1)$ augmented parameter matrix $\overline{\mathbf{W}} \equiv [\mathbf{w}_0\ ; \mathbf{W}]$ is introduced for notational convenience.

If the data is completely separable by the map $\mathbf{f}(\mathbf{x})$, there exists no maximum likelihood estimate of $\overline{\mathbf{W}}$ which maximizes the likelihood function

$$L(\overline{\mathbf{W}}) \equiv \prod_{j=1}^{N} p_{c_j}(\mathbf{x_j}) \equiv \prod_{j=1}^{N} \hat{p}_{c_j}(\mathbf{f}(\mathbf{x_j})) \equiv \prod_{j=1}^{N} \hat{p}_{c_j}(\overline{\mathbf{W}}\,\overline{\phi}(\mathbf{x})) \tag{11}$$

with respect to $\overline{\mathbf{W}}$. Besides, even in the cases where the maximum likelihood estimate $\overline{\mathbf{W}}^{**}$ exists, overfitting could occur with $\overline{\mathbf{W}}^{**}$. If this is the case, the learning process of maximizing the likelihhod suffers from the phenomenon called 'overlearning'. In order to avoid it and obtain a *due induction(generalization) capacity to the size and the quality of an available training data set*, we introduce a penalty function

$$P_{induct}(\overline{\mathbf{W}}) \equiv exp(-\frac{1}{2}trace\ \mathbf{\Gamma}\overline{\mathbf{W}}\mathbf{\Sigma}\overline{\mathbf{W}}^t) \equiv exp(-\frac{1}{2}(\|\sigma\mathbf{w}_0 + \mathbf{W}\mu\|_{\mathbf{\Gamma}}^2 + trace\ \mathbf{\Gamma}\mathbf{W}\mathbf{\Sigma}\mathbf{W}^t)),$$

where $\mathbf{\Gamma}$ is an $K \times K$ positive definite matrix, $\overline{\mathbf{\Sigma}}$, σ, μ and $\mathbf{\Sigma}$ are given in Section 2, $\| \ \|_F$ is the Frobenius norm, and $\|\mathbf{a}\|^2_{\mathbf{\Gamma}} \equiv \mathbf{a}^t\mathbf{\Gamma}\mathbf{a}$.

We employ the penalized logistic regression(PLR) likelihood $PL_\delta(\overline{\mathbf{W}}) \equiv L(\overline{\mathbf{W}})P^\delta_{induct}(\overline{\mathbf{W}})$ which is to be maximized for obtaining the optimal parameter value $\overline{\mathbf{W}}^*$ of the model, where $\delta \in [0, \infty)$ is a balancing parameter introduced for notational convenience. We have introduced the matrices $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$, the vectors μ and λ, the scalars ω, σ and δ as *hyperparameters* for the model so that we can gain a variety of *induction(generalization) capacity* of the resulting machines by controlling them according to the sampling scheme of training data set and also to the prospective situation in which the predictor is to be used. Generally, we determine the values of the hyperparameters by a statistical criteria such as the empirical Bayes method, the maximum Type II likelihood method and *GIC* method.

The choice of $\mathbf{\Gamma}$ *does not affect the kernel function itself* as was seen in [1,2], but it controls the learning process and hence the induction characteristics of the obtainable predictor. In other words, *the induction penalty does not completely specify the kernel function.* See [1] for further discussions on this point.

5 Maximum PLR Likelihood

The maximum penalized logistic regression likelihood estimate $\overline{\mathbf{W}}^*$ is given by minnimizing the negative-log-penalized-likelihood,

$$pl_\delta(\overline{\mathbf{W}}) \equiv -log\ PL_\delta(\overline{\mathbf{W}}) = -\sum_{j=1}^{N} log\ \hat{p}_{c_j}(\mathbf{f}(\mathbf{x}_j)) + \frac{\delta}{2} trace \mathbf{\Gamma}\overline{\mathbf{W}}\mathbf{\Sigma}\overline{\mathbf{W}}^t. \tag{12}$$

Bishop(1992) gave a set of formulas to be composed for obtaining the derivatives of this function, but his formulas can be further simplified to the following closed formulas, in which many terms in his formulas have been cancelled out one another in the case of the likelihood of the multinomial distribution.

Lemma 1: The following equalities hold for $\delta \geq 0$.

$$\nabla pl_\delta(\overline{\mathbf{W}}) \equiv (\mathbf{P}(\overline{\mathbf{W}}) - \mathbf{Y})\overline{\mathbf{\Phi}}^t + \delta\mathbf{\Gamma}\overline{\mathbf{W}}\overline{\mathbf{\Sigma}}, \tag{13}$$

$$\nabla^2 pl_\delta(\overline{\mathbf{W}}) \equiv \sum_{j=1}^{N}\{(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) \otimes ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)\mathbf{p}^t(\mathbf{x}_j))\} + \delta\overline{\mathbf{\Sigma}} \otimes \mathbf{\Gamma} \tag{14}$$

where $\nabla \equiv \partial/\partial\overline{\mathbf{W}}$ arranged in the same $K \times (m+1)$ matrix form as $\overline{\mathbf{W}}$ itself, $\mathbf{P}(\overline{\mathbf{W}})$ is the $K \times N$ matrix defined by $\mathbf{P}(\overline{\mathbf{W}}) \equiv [\mathbf{p}(\mathbf{x}_1);\ \mathbf{p}(\mathbf{x}_2);\ \cdots; \mathbf{p}(\mathbf{x}_N)]$, whose j-th column vector is given by $\mathbf{p}(\mathbf{x}_j) \equiv \hat{\mathbf{p}}(\mathbf{f}(\mathbf{x}_j)) \equiv \hat{\mathbf{p}}(\overline{\mathbf{W}}\ \overline{\phi}(\mathbf{x}_j))$, $[\mathbf{p}\ ] \equiv daig(\mathbf{p})$ is the diagonal matrix formed from the vector $\mathbf{p}$ and $\otimes$ is the tensor product.

Lemma 2: The following equalities hold for $\delta \geq 0$.

$$\begin{aligned} &\delta\overline{\mathbf{\Sigma}} \otimes \mathbf{\Gamma} \ \prec\ \nabla^2 pl_\delta(\overline{\mathbf{W}}) \ \prec\ (\overline{\mathbf{\Phi}}\overline{\mathbf{\Phi}}^t) \otimes [\vee\mathbf{p}] + \delta\overline{\mathbf{\Sigma}} \otimes \mathbf{\Gamma} \\ &\prec (\overline{\mathbf{\Phi}}\overline{\mathbf{\Phi}}^t) \otimes (\mathbf{I}_K - K^{-1}\mathbf{1}_K\mathbf{1}_K^t) + \delta\overline{\mathbf{\Sigma}} \otimes \mathbf{\Gamma} \prec\ (\overline{\mathbf{\Phi}}\overline{\mathbf{\Phi}}^t) \otimes \mathbf{I}_K + \delta\overline{\mathbf{\Sigma}} \otimes \mathbf{\Gamma}, \end{aligned} \tag{15}$$

where $\mathbf{A} \prec \mathbf{B}$ implies that $\mathbf{B} - \mathbf{A}$ is a nonnegative definite matrix, $\mathbf{I}_K$ is the identity matrix of size K, $\mathbf{\Pi}_N \equiv N^{-1}\mathbf{1}_N\mathbf{1}_N^t$ is the orthogonal projector onto the space spanned by $\mathbf{1}_N$, and $\vee\mathbf{p} \equiv \vee\mathbf{p}(\overline{\mathbf{W}}) \equiv \bigvee_{j=1}^{N} \mathbf{p}(\mathbf{x}_j)$ is the smallest vector such that $\vee\mathbf{p} \geq \mathbf{p}(\mathbf{x}_j)$ for $j = 1, 2, ..., N$ and the inequality is meant elementwise.

Proposition 3: The functions, $pl_0(\overline{\mathbf{W}})$ and $pl_\delta(\overline{\mathbf{W}})$ $(\delta > 0)$ are convex and strictly convex functions respectively with respect to the parameter $\overline{\mathbf{W}}$. The function $pl_\delta(\overline{\mathbf{W}})$ has the unique minimum point $\overline{\mathbf{W}}^*$ which satisfies the condition, $\nabla pl_\delta(\overline{\mathbf{W}}^*) = \mathbf{O}_{K,m+1}$, where $\mathbf{O}_{K,m+1}$ is the $K \times (m+1)$ zero matrix. See [1] for the proof.

PLRP: We adopt $\mathbf{p}^*(\mathbf{x}) \equiv \hat{\mathbf{p}}(\mathbf{f}^*(\mathbf{x})) \equiv \hat{\mathbf{p}}(\overline{\mathbf{W}}^*\overline{\phi}(\mathbf{x}))$, as a our predictive probability vector, and $y^*(\mathbf{x}) \equiv arg\max_k f_k^*(\mathbf{x})$ as our deterministic predicton function, where $\mathbf{f}^*(\mathbf{x}) \equiv \overline{\mathbf{W}}^*\overline{\phi}(\mathbf{x})$.

6 Penalized Logistic Regression Machines

Based on Lemma 1 and 2, we constructed in [1] the *penalized logistic regression machine* (PLRM) for computing $\overline{\mathbf{W}}^*$,.

PLRM-1: Starting with an $K \times (m+1)$ matrix $\overline{\mathbf{W}}^0$, generate a sequence of matrices $\{\overline{\mathbf{W}}^i\}_{i=1,2,...}$ by the iterative formula for i=0,1,2,..., ∞,

$$\overline{\mathbf{W}}^{i+1} = \overline{\mathbf{W}}^i - \alpha_i((\mathbf{P}(\overline{\mathbf{W}}^i) - \mathbf{Y})\overline{\mathbf{\Phi}}^t + \delta\mathbf{\Gamma}\overline{\mathbf{W}}^i\overline{\mathbf{\Sigma}}), \tag{16}$$

where $\overline{\mathbf{\Phi}}$ is the constant matrix given in Eqs. (2).

Theorem 4: The sequence generated by PLRM-1 converges to the unique minimizer $\overline{\mathbf{W}}^*$ of pl_δ for any choice of initial matrix $\overline{\mathbf{W}}^0 s$ under certain condition which was specified in [1], for example $\alpha_i = (\|\overline{\mathbf{\Sigma}}\|(\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|))^{-1}$. Then its convergence rate is less than $1 - \delta((\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|)\|\mathbf{\Gamma}^{-1}\|cond\ \overline{\mathbf{\Sigma}})^{-1}$, where $cond\mathbf{A} \equiv \|\mathbf{A}\|\|\mathbf{A}^{-1}\|(\geq 1)$ is the condition number of a matrix $\mathbf{A}$ and $\|\mathbf{A}\|$ is the spectral norm of $\mathbf{A}$. See [1] for details and the proof.

PLRM-2: Starting with an $K \times (m+1)$ matrix $\overline{\mathbf{W}}^0$, generate a sequence of matrices $\{\overline{\mathbf{W}}^i\}_{i=1,2,...}$ by the iterative formula,

$$\overline{\mathbf{W}}^{i+1} = \overline{\mathbf{W}}^i - \alpha_i\Delta\overline{\mathbf{W}}^i,\ i = 0, 1, 2, ..., \infty, \tag{17}$$

where $\Delta\overline{\mathbf{W}}^i \equiv \mathcal{G}_{plrm2}(\overline{\mathbf{W}}^i)$ is the unique solution of the linear matrix equation,

$$\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\overline{\mathbf{W}}^i(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\Delta\overline{\mathbf{W}}^i\ \overline{\mathbf{\Sigma}}$$

$$= (\mathbf{P}(\overline{\mathbf{W}}^i) - \mathbf{Y})\overline{\mathbf{\Phi}}^t + \delta\mathbf{\Gamma}\overline{\mathbf{W}}^i\overline{\mathbf{\Sigma}}. \tag{18}$$

Theorem 5: The sequence generated by PLRM-2 converges to the unique minimizer $\overline{\mathbf{W}}^*$ of pl_δ for any choice of initial matrix $\overline{\mathbf{W}}^0$, if α_i is chosen so that

$$\nu \leq \frac{pl_\delta(\overline{\mathbf{W}}^i + \alpha_i \Delta\overline{\mathbf{W}}^i) - pl_\delta(\overline{\mathbf{W}}^i)}{\alpha_i \; trace \; (\Delta\overline{\mathbf{W}}^i)^t \nabla pl_\delta(\overline{\mathbf{W}}^i)} \leq 1 - \nu, \tag{19}$$

for a scalar ν such that $0 < \nu < \frac{1}{2}$, with $\alpha_i = 1$ whenever it satisfies Ineq.(42). Further, there exist a number $\bar{i}$ such that the stepsizes $\alpha_i = 1$ is possible whenever $i > \bar{i}$, and the convergence is superlinear.

Theorem 6: The sequence generated by PLRM-2 converges quadratically to the unique minimizer $\overline{\mathbf{W}}^*$ of pl_δ if the initial matrix $\overline{\mathbf{W}}^0$ satisfies

$$\delta^{-1} cond\overline{\mathbf{\Sigma}} \; \|\mathbf{\Gamma}^{-1}\|(\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|)\|\Delta\overline{\mathbf{W}}^0\|_F \leq \frac{1}{2}, \tag{20}$$

and if $\alpha_i \equiv 1$ is chosen.
Eventually PLRM-2 has the quadratic convergence property if the stepsize is controlled in such a way as in Theorem 5, because $\|\Delta\overline{\mathbf{W}}^i\|$ tends to zero, satisfying Ineq.(42) in a final stage of the iteration.

The linearer equation (22) can be conveniently solved by the following algorithm.
CG Method: Starting with an arbitrary initial approximation $\Delta\overline{\mathbf{W}}_0$, which is usually taken to be zero matrix, we generate a sequence $\Delta\overline{\mathbf{W}}_k$ of matrices which converges to the solution of Eq. (22) by the following iterative formula.

$$\begin{aligned}\mathbf{R}_0 = (\mathbf{P}(\overline{\mathbf{W}}^i) - \mathbf{Y})\overline{\mathbf{\Phi}}^t + \delta\mathbf{\Gamma}\overline{\mathbf{W}}^i\overline{\mathbf{\Sigma}} \\ - \sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\overline{\mathbf{W}}_0(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\Delta\overline{\mathbf{W}}_0\,\overline{\mathbf{\Sigma}}\end{aligned} \tag{21}$$

$$\mathbf{Q}_0 = \mathbf{R}_0$$

$$\mathbf{Q}_0 = \sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_0(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}^t\mathbf{R}_0\,\overline{\mathbf{\Sigma}}$$

$$\alpha_k = \frac{trace\mathbf{Q}_k^t\mathbf{R}_k}{trace(\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{Q}_k(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\mathbf{Q}_k\,\overline{\mathbf{\Sigma}})\mathbf{Q}_k^t}$$

$$\Delta\overline{\mathbf{W}}_{k+1} = \Delta\overline{\mathbf{W}}_k + \alpha_k\mathbf{Q}_k, \tag{22}$$

$$\mathbf{R}_{k+1} = (\mathbf{P}(\overline{\mathbf{W}}^i) - \mathbf{Y})\overline{\mathbf{\Phi}}^t + \delta\mathbf{\Gamma}\overline{\mathbf{W}}^i\overline{\mathbf{\Sigma}} \tag{23}$$
$$-\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\overline{\mathbf{W}}_{k+1}(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\Delta\overline{\mathbf{W}}_{k+1}\ \overline{\mathbf{\Sigma}}$$

$$\beta_k = \frac{trace(\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{Q}_k(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\mathbf{Q}_k\overline{\mathbf{\Sigma}})\mathbf{R}_{k+1}^t}{trace(\sum_{j=1}^{N}([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{Q}_k(\overline{\phi}(\mathbf{x}_j)\overline{\phi}^t(\mathbf{x}_j)) + \delta\mathbf{\Gamma}\mathbf{Q}_k\ \overline{\mathbf{\Sigma}})\mathbf{Q}_k^t}$$

$$\mathbf{Q}_{k+1} = \mathbf{R}_{k+1} + \beta_k\mathbf{Q}_k$$

7 Dual Penalized Logistic Regression Likelihood

We showed in [1] that the penalized logistic regression model also yields a certain duality which leads intrinsically to the kernel methods.

Eq.(13) implies that the minimizer $\overline{\mathbf{W}}^*$ of $pl_\delta(\overline{\mathbf{W}})$ is of the form

$$\overline{\mathbf{W}}^* = \mathbf{V}^*\overline{\mathbf{\Phi}}^t\overline{\mathbf{\Sigma}}^{-1}, \tag{24}$$

where $\mathbf{V}^* \equiv \delta^{-1}\mathbf{\Gamma}^{-1}(\mathbf{Y} - \mathbf{P}(\overline{\mathbf{W}}^*))$. Therefore, by introducing the *dual parameter* matrix $\mathbf{V}$ of size $K \times N$ in such a way as

$$\overline{\mathbf{W}} = \mathbf{V}\overline{\mathbf{\Phi}}^t\overline{\mathbf{\Sigma}}^{-1}, \tag{25}$$

we only have to minimize the negative log penalized logistic regression likelihood

$$\tilde{pl}_\delta(\mathbf{V}) \equiv pl_\delta(\overline{\mathbf{W}}) \equiv pl_\delta(\mathbf{V}\overline{\mathbf{\Phi}}^t\overline{\mathbf{\Sigma}}^{-1}) \tag{26}$$

with respect to $\mathbf{V}$ instead of matrix $\overline{\mathbf{W}}$. This transformation of the parameters naturally leads to the kernel methods. Substituting Eq.(25) into Eq.(10), we have

$$\mathbf{f}(\mathbf{x}) \equiv \overline{\mathbf{W}}\ \overline{\phi}(\mathbf{x}) = \mathbf{V}\overline{\mathbf{\Phi}}^t\ \overline{\mathbf{\Sigma}}^{-1}\overline{\phi}(\mathbf{x}) = \mathbf{V}\kappa_\omega(\mathbf{x}). \tag{27}$$

The matrix $\mathbf{P}(\overline{\mathbf{W}})$ is unchanged with this transformation, but can also be denoted by $\tilde{\mathbf{P}}(\mathbf{V}) \equiv \mathbf{P}(\overline{\mathbf{W}}) \equiv \mathbf{P}(\mathbf{V}\overline{\mathbf{\Phi}}^t\overline{\mathbf{\Sigma}}^{-1})$, whose j-th column vector $\mathbf{p}(\mathbf{x}_j) \equiv \hat{\mathbf{p}}(\mathbf{f}(\mathbf{x}_j)) = \hat{\mathbf{p}}(\mathbf{V}\kappa_\omega(\mathbf{x}_j))$ can be computed by using the map $\kappa_\omega(\mathbf{x})$ only. Since the parameter transformation is linear, the function $\tilde{pl}_\delta(\mathbf{V})$ is also a convex function with rerspct to the *dual parameter* $\mathbf{V}$, in terms of which it is represented by

$$\tilde{pl}_\delta(\mathbf{V}) \equiv pl_\delta(\mathbf{V}\overline{\mathbf{\Phi}}^t\overline{\mathbf{\Sigma}}^{-1}) = -\sum_{j=1}^{N} log\ \hat{p}_{c_j}(\mathbf{V}\kappa_\omega(\mathbf{x}_j)) + \frac{\delta}{2}trace\mathbf{\Gamma}\mathbf{V}\mathcal{K}_\omega^d\mathbf{V}^t. \tag{28}$$

Remark: The transformed negative log penalized likelihood $\tilde{pl}_\delta(\mathbf{V})$ involves only the matrix $\mathcal{K}_\omega^d$, its column vectors $\{\kappa_\omega(\mathbf{x}_j)\}$ and $\mathbf{\Gamma}$. Also the predictor probability vector $\{\mathbf{p}(\mathbf{x}_j)\}$ involves only the kernel function, $\kappa_\omega(\mathbf{x})$. They do not depend explicitly on $\overline{\mathbf{\Phi}}$, $\overline{\phi}(\mathbf{x})$ nor $\overline{\mathbf{\Sigma}}$.

Lemma 7: The derivatives of $\tilde{pl}_\delta$ with respect to the dual parameter $\mathbf{V}$ are given as follows.

$$\nabla\tilde{pl}_\delta(\mathbf{V}) \equiv (\tilde{\mathbf{P}}(\mathbf{V}) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V})\mathcal{K}_\omega^d, \tag{29}$$

$$\nabla^2\tilde{pl}_\delta(\mathbf{V}) \equiv \sum_{j=1}^{N}\{(\kappa_\omega(\mathbf{x}_j)\kappa_\omega^t(\mathbf{x}_j)) \otimes ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)\mathbf{p}^t(\mathbf{x}_j))\} + \delta\mathcal{K}_\omega^d \otimes \mathbf{\Gamma} \tag{30}$$

The second derivatives are uniformly bounded. See [1] for more details.

8 Dual Penalized Logistic Regression Machines

We constructed in [1] the *dual penalized logistic regression machine* (dPLRM) for computing a minimizer $\mathbf{V}^{**}$ of the function $\tilde{pl}_\delta(\mathbf{V})$.

If the matrix $\mathcal{K}_\omega^d$ is nonsingular, $\mathbf{V}^{**} = \mathbf{V}^*$ is the minimizer which is the unique solution of the matrix equation,

$$\mathbf{D}(\mathbf{V}) \equiv \tilde{\mathbf{P}}(\mathbf{V}) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V} = \mathbf{O}_{K,N}. \tag{31}$$

An algorithm[1] was given for this case.
dPLRM-0: Starting with an $K\times N$ matrix $\mathbf{V}^0$, generate a sequence of matrices $\{\mathbf{V}^i\}_{i=1,2,...}$ by the iterative formula for i=0,1,2,..., ∞

$$\mathbf{V}^{i+1} = \mathbf{V}^i - \alpha_i(\mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V}^i). \tag{32}$$

Theorem 8: The sequence generated by dPLRM-0 converges to the unique minimizer $\mathbf{V}^*$ of $\tilde{pl}_\delta(\mathbf{V})$ for any choice of initial matrix $\mathbf{V}^0$ under certain condition, which was specified in [1], for example $\alpha_i = (\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|)^{-1}$. Then its convergence rate is less than $1 - \delta((\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|)\|\mathbf{\Gamma}^{-1}\|)^{-1}$. See [1] for the proof.

Remark : The vector field defined by $\mathbf{D}(\mathbf{V})$ is not a gradient vector field. Since the dual machine dPLRM-0 which employs this vector field is such a simple process as to require only the evaluations of $\tilde{\mathbf{P}}(\mathbf{V}) - \mathbf{Y}$ and $\mathbf{\Gamma}\mathbf{V}$ and no matrix inversion, the present author cannnot help speculating if this dual machine is a better approximation to phisiological reality of learing process than the existing machines.

If $\mathcal{K}_\omega^d$ is a singular matrix, then we have the following algorithm, whose convergence is generally slower than dPLRM-0.
dPLRM-1: Starting with an $K\times N$ matrix $\mathbf{V}^0$, generate a sequence of matrices $\{\mathbf{V}^i\}_{i=1,2,...}$ by the iterative formula for i=0,1,2,..., ∞,

$$\mathbf{V}^{i+1} = \mathbf{V}^i - \alpha_i(\mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V}^i)\mathcal{K}_\omega^d. \tag{33}$$

We also give a dPLRM which has rapid converge property.

dPLRM-2: Starting with an $K \times N$ matrix $\mathbf{V}^0$, generate a sequence of matrices $\{\mathbf{V}^i\}_{i=1,2,...}$ by the iterative formula,

$$\mathbf{V}^{i+1} = \mathbf{V}^i - \alpha_i \Delta \mathbf{V}^i, \ i = 0, 1, 2, ..., \infty, \tag{34}$$

where $\Delta \mathbf{V}^i$ is the solution of the linear matrix equation,

$$\sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t) \Delta \mathbf{V}^i (\kappa_\omega(\mathbf{x}_j) \kappa_\omega^t(\mathbf{x}_j)) + \delta \mathbf{\Gamma} \Delta \mathbf{V}^i \mathcal{K}_\omega^d$$
$$= (\mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta \mathbf{\Gamma} \mathbf{V}^i) \mathcal{K}_\omega^d, \tag{35}$$

which, if $\mathcal{K}_\omega^d$ is nonsingular, is equivalent to the linear equation,

$$\sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t) \Delta \mathbf{V}^i (\kappa_\omega(\mathbf{x}_j) \mathbf{e}_j^t) + \delta \mathbf{\Gamma} \Delta \mathbf{V}^i$$
$$= \mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta \mathbf{\Gamma} \mathbf{V}^i. \tag{36}$$

where $\mathbf{e}_j$ is the j-th unit vector.

Theorem 9: The sequence generated by PLRM-2 converges to the unique minimizer $\mathbf{V}^*$ of $\tilde{pl}_\delta$ for any choice of initial matrix $\mathbf{V}^0$, if α_i is chosen so that

$$\nu \leq \frac{\tilde{pl}_\delta(\mathbf{V}^i + \alpha_i \Delta \mathbf{V}^i) - \tilde{pl}_\delta(\mathbf{V}^i)}{\alpha_i \ trace \ (\Delta \mathbf{V}^i)^t \nabla \tilde{pl}_\delta(\mathbf{V}^i)} \leq 1 - \nu, \tag{37}$$

for a scalar ν such that $0 < \nu < \frac{1}{2}$, with $\alpha_i = 1$ whenever it satisfies Ineq.(37). Further, there exist a number $\bar{i}$ such that the stepsizes $\alpha_i = 1$ is possible whenever $i > \bar{i}$, and the convergence is superlinear.

Theorem 10: The sequence generated by dPLRM-2 converges quadratically to the unique minimizer $\mathbf{V}^*$ of $\tilde{pl}_\delta$if the initial matrix $\mathbf{V}^0$ satisfies

$$\delta^{-1} \ \|\mathbf{\Gamma}^{-1}\|(\|\mathcal{K}_\omega^d\| + \delta\|\mathbf{\Gamma}\|)\|\Delta \mathbf{V}^0\|_F \leq \frac{1}{2}, \tag{38}$$

and if $\alpha_i \equiv 1$ is chosen.

Eventually dPLRM-2 has the quadratic convergence property if the stepsize is controlled in such a way as in Theorem 21, because $\|\Delta \mathbf{V}^i\|$ tends to zero, satisfying Ineq.(38) in a final stage of the iteration.

The linearer equation (35) can be conveniently solved by the following algorithm.

CG Method: Starting with an arbitrary initial approximation $\Delta\mathbf{V}_0$, which is usually taken to be zero matrix, we generate a sequence $\Delta\mathbf{V}_k$ of matrices which converges to a solution of Eq. (22) by the following iterative formula.

$$\begin{aligned}\mathbf{R}_0 = \mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V}^i \\ - \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\mathbf{V}_0\kappa_j\mathbf{e}_j^t + \delta\mathbf{\Gamma}\Delta\mathbf{V}_0\end{aligned} \tag{39}$$

$$\mathbf{Q}_0 = \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_0\mathbf{e}_j\kappa_j^t + \delta\mathbf{\Gamma}^t\mathbf{R}_0$$

$$\alpha_k = \frac{\left\| \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_k\mathbf{e}_j\kappa_j^t + \delta\mathbf{\Gamma}^t\mathbf{R}_k \right\|_F^2}{\left\| \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{Q}_k\kappa_j\mathbf{e}_j^t + \delta\mathbf{\Gamma}^t\mathbf{Q}_k \right\|_F^2}$$

$$\Delta\mathbf{V}_{k+1} = \Delta\mathbf{V}_k + \alpha_k\mathbf{Q}_k, \tag{40}$$

$$\begin{aligned}\mathbf{R}_{k+1} = \mathbf{P}(\mathbf{V}^i) - \mathbf{Y} + \delta\mathbf{\Gamma}\mathbf{V}^i \\ - \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\Delta\mathbf{V}_{k+1}\kappa_j\mathbf{e}_j^t + \delta\mathbf{\Gamma}\Delta\mathbf{V}_{k+1}\end{aligned} \tag{41}$$

$$\beta_k = \frac{\left\| \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_{k+1}\mathbf{e}_j\kappa_j^t + \delta\mathbf{\Gamma}^t\mathbf{R}_{k+1} \right\|_F^2}{\left\| \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_k\mathbf{e}_j\kappa_j^t + \delta\mathbf{\Gamma}^t\mathbf{R}_k \right\|_F^2}$$

$$\mathbf{Q}_{k+1} = \sum_{j=1}^{N} ([\mathbf{p}(\mathbf{x}_j)] - \mathbf{p}(\mathbf{x}_j)(\mathbf{p}(\mathbf{x}_j))^t)\mathbf{R}_{k+1}\mathbf{e}_j\kappa_j^t + \delta\mathbf{\Gamma}^t\mathbf{R}_{k+1} + \beta_k\mathbf{Q}_k$$

where $\kappa_j \equiv \kappa_\omega(\mathbf{x}_j)$ and $\| \ \|_F$ is the Frobenius norm of a matrix.

Remark: We can work out the process of getting the predictor only with the quantities related to the symbols $\mathbf{V}, \tilde{\mathbf{P}}(\mathbf{V})$, $\mathcal{K}_\omega^d$, $\kappa_\omega(\mathbf{x})$ and $\mathbf{\Gamma}$ without resorting at all to $\overline{\mathbf{W}}, \mathbf{P}(\overline{\mathbf{W}})$, $\overline{\mathbf{\Phi}}$, $\overline{\phi}(\mathbf{x})$ nor $\overline{\mathbf{\Sigma}}$, which we can also do without for evaluation of the functions,

$$\mathbf{f}^*(\mathbf{x}) = \mathbf{V}^*\kappa_\omega(\mathbf{x}), \tag{42}$$

and the predictive probability

$$\mathbf{p}^*(\mathbf{x}) \equiv \hat{\mathbf{p}}(\mathbf{f}^*(\mathbf{x})) = \hat{\mathbf{p}}(\mathbf{V}^* \kappa_\omega(\mathbf{x})), \quad (43)$$

for the prediction given $\mathbf{x}$. This implies that we could perform the above mentioned learning and prediction process only with the kernel function $\mathcal{K}(\mathbf{x}, \mathbf{y})$ without resort to the original map $\phi(\mathbf{x})$ itself. It does not even matter whether the kernel function is constructed from such a map. The situation is largely parallel to that of SVM. See Scholkopf et al.(1999). The methods described in this and the previous sections will be called *dual* penalized logistic regression methods and the method described in Sections 4 and 5 are called *primal* methods. Likewise, the algorithms given in Section 5 are called *primal PLR machines* as against dual PLR machines.

The full reference is not included here due to space limitation. For the literatures cited in this paper and other related works, see the extensive refereces of[1].

References

[1] Kunio Tanabe, *Penalized Logistic Regression Machines: New methods for statistical prediction 1*, ISM Cooperative Research Report 143, Estimation and Smoothing Methods in Nonparametric Statistical Models, 163-194, March 2001.

[2] Kunio Tanabe, *Penalized Logistic Regression Machines: New methods for statistical prediction 2*, Proceedings of 2001 Workshop on Information-Based Induction Science(IBIS2001), 71-76, August 2001