Kernel Discriminative Random Fields for Land Cover Classification

Ribana Roscher, Björn Waske, Wolfgang Förstner
Department of Photogrammetry, Inst. of Geodesy and Geoinformation, University of Bonn
{*rroscher, bwaske, wfoerstn*}*@uni-bonn.de*

Abstract

Logistic Regression has become a commonly used classifier, not only due to its probabilistic output and its direct usage in multi-class cases. We use a sparse Kernel Logistic Regression approach – the Import Vector Machines – for land cover classification. We improve our segmentation results applying a Discriminative Random Field framework on the probabilistic classification output. We consider the performance regarding to the classification accuracy and the complexity and compare it to the Gaussian Maximum Likelihood classification and the Support Vector Machines.

1. Introduction

Land cover classification is perhaps the widest used application in remote sensing. Most studies use well known statistical methods like the Gaussian Maximum Likelihood classification (MLC). However, these "early" methods are often limited in the context of enhanced Earth-Observation systems and increased availability of diverse remote sensing datasets. Besides that, more stringent performance requirements like accuracy, speed (e.g. near-real time applications) and cost demand more sophisticated classification concepts (e. g. [15, 5]).

As a result, the user can choose between several widely accepted algorithms such as decision trees, neural networks and Support Vector Machines (e. g. [15, 5]) Particularly Support Vector Machines (SVM, [20]) have emerged over the past decade and have been successfully introduced in context of remote sensing. However, probabilistic discriminative models (e. g. [12, 19]) like Logistic Regression (e. g. [7, 22, 10]) are also used since the probabilities themselves are often of interest.

Another development in remote sensing image analysis is that of segment-based or object-based approaches, where adjacent pixels with similar properties are aggregated into image segments. After image segmentation, additional information such as the segments' mean value and texture as well as neighborhood relationships can be derived and included into the classification process. Shackelford and Davis [17] for example combined information from pixel- and segment-levels to separate classes that are spectrally similar at pixel level. However, the definition of adequate segmentation level might be critical ([21]) and Song et al. [18] for example demonstrated how an inadequate segmentation affects the classification accuracy.

Using a Discriminative Random Field [12] is an alternative approach to model the spatial interactions between pixels. We apply it in a probabilistic discriminative framework following the concept of Conditional Random Fields proposed by Lafferty et al. [13]. We model the probabilistic output with the Import Vector Machine (IVM, [23]) – a sparse Kernel Logistic Regression approach. Since both Kernel Logistic Regression (e. g. [14, 9, 3, 16]) and the IVMs already achieve high accuracies for machine learning datasets [23], such as speaker identification [8] and cancer diagnosis [11], they see also interesting in context of remote sensing imagery.

In the first chapter we give an overview about the Logistic Regression, its extension to Kernel Logistic Regression and to the IVMs. Afterwards we introduce the Discriminative Random Field and incorporate the probabilistic output of the IVMs. In our experiments we evaluate the proposed algorithm on a set of two Landsat images, using different amount of training points. Concerning the similarity of SVMs and IVMs we compare the performance of both algorithms in terms of accuracy and complexity to each other and to the Gaussian Maximum Likelihood classification.

We clarify the questions how good the IVM performs in comparison to the mentioned algorithms and if the usage of a Discriminative Random Field can improve the classification accuracy.

2. Logistic Regression

We assume to have a training set $(\boldsymbol{x}_n, y_n)$, $n = 1, \ldots, N$ of N labeled samples with vectors $\boldsymbol{x}_n$ of observations and class labels $y_n \in \mathcal{C} = \{1, \ldots, K\}$. The observations are collected in a matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$.

In the two-class case the posterior probability p_n of a feature vector $\boldsymbol{x}_n$ is assumed to follow the Logistic Regression model

$$p(y_n|\mathbf{x}_n; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_n)} \tag{1}$$

with the extended feature vector $\mathbf{x}_n^\mathsf{T} = [1, \boldsymbol{x}_n^\mathsf{T}] \in \mathbb{R}^M$ and the extended parameters $\mathbf{w}^\mathsf{T} = [w_{k0}, \boldsymbol{\omega}^\mathsf{T}] \in \mathbb{R}^M$ containing the bias w_{k0} and the weight vector $\boldsymbol{\omega}$.

The objective function $\mathcal{Q}_0(\mathbf{w})$ of the standard logistic regression model is given by the negative log-likelihood function

$$\mathcal{Q}_0(\mathbf{w}) = -\frac{1}{N}\sum_n \left[t_n \log p_n + (1 - t_n)\log\left(1 - p_n\right)\right]. \tag{2}$$

The binary target vector $t \in \{0, 1\}$ of length N codes the labels with $t_n = 0$ for $\mathcal{C}_1$ and $t_n = 1$ for $\mathcal{C}_2$.

The Newton-Raphson iteration scheme for the minimization of (2) is given by

$$\mathbf{w}^{(i)} = \mathbf{w}^{(i-1)} - \boldsymbol{H}^{-1}\nabla\boldsymbol{E} \tag{3}$$

with the gradient $\nabla\boldsymbol{E} = \boldsymbol{X}(\boldsymbol{p} - \boldsymbol{t})$ and the Hessian $\boldsymbol{H} = \boldsymbol{X}^\mathsf{T}\boldsymbol{R}\boldsymbol{X}$. The $(N \times N)$-dimensional diagonal matrix $\boldsymbol{R}$ has the elements $r_{nn} = p_n(1 - p_n)$, which can be obtained from (1).

We can reformulate the Newton-Raphson iteration scheme in (3) and obtain the iterated reweighted least squares solution

$$\mathbf{w}^{(i)} = \left(\frac{1}{N}\boldsymbol{X}^\mathsf{T}\boldsymbol{R}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{R}\boldsymbol{z}, \tag{4}$$

$$\boldsymbol{z} = \frac{1}{N}\left(\boldsymbol{X}\mathbf{w}^{(i-1)} + \boldsymbol{R}^{-1}(\boldsymbol{p} - \boldsymbol{t})\right) \tag{5}$$

to obtain adjusted parameters in each iteration i. We also introduce a regularization parameter λ to prevent overfitting, especially in the case of separable or nearly separable data.

3. Kernel Logistic Regression and Import Vector Machines

To use the linear classifier for solving a non-linear problem we introduce kernels to map the original observations from the input space into a higher-dimensional kernel space. This approach has already been successfully applied to several applications like object recognition or speech recognition (e. g. [16, 8, 11]).

3.1 Kernel Logistic Regression

We introduce kernels and transform the features $\boldsymbol{X} \in \mathbb{R}^M$ to a higher dimensional feature space $\mathcal{F}$ making use of the kernel trick [1]. The kernel function K is given by

$$\boldsymbol{K}(\mathbf{x}_n, \mathbf{x}_m) = \boldsymbol{\phi}^\mathsf{T}(\mathbf{x}_n)\,\boldsymbol{\phi}(\mathbf{x}_m). \tag{6}$$

Following the Representer Theorem the parameters $\boldsymbol{W}$ lie within the span of the feature vectors $\boldsymbol{\Phi}$:

$$\boldsymbol{W} = \sum_n \alpha_n \boldsymbol{\phi}_n = \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\alpha}. \tag{7}$$

The vector $\boldsymbol{\alpha}$ contains the parameters which define the linear decision boundaries in kernel space.

With (6) and (7), (4) and (5) become

$$\boldsymbol{\alpha}^{(i)} = \left(\frac{1}{N}\boldsymbol{K}^\mathsf{T}\boldsymbol{R}\boldsymbol{K} + \lambda\boldsymbol{K}\right)^{-1}\boldsymbol{K}^\mathsf{T}\boldsymbol{R}\boldsymbol{z}, \tag{8}$$

$$\boldsymbol{z} = \frac{1}{N}\left(\boldsymbol{K}\boldsymbol{\alpha}^{(i-1)} + \boldsymbol{R}^{-1}(\boldsymbol{p} - \boldsymbol{t})\right). \tag{9}$$

3.2 Import Vector Machines

The problem with Kernel Logistic Regression is that all training samples are included to train the classifier, which is computationally expensive in datasets with many training samples. Similar to the widely used SVMs only a few feature vectors are necessary to define the decision boundaries. These feature vectors are called import vectors. Using only these vectors we obtain a sparse solution of the Kernel Logistic Regression – the Import Vector Machines [23].

Following Zhu and Hastie [23] we only choose a subset $\mathcal{S}$ out of the training set $\mathcal{T}$ with $S = |\mathcal{S}|$ samples and yield for (8) and (9)

$$\boldsymbol{\alpha}^{(i)} = \left(\frac{1}{N}\boldsymbol{K}_S^\mathsf{T}\boldsymbol{R}\boldsymbol{K}_S + \lambda\boldsymbol{K}_R\right)^{-1}\boldsymbol{K}_S^\mathsf{T}\boldsymbol{R}\tilde{\boldsymbol{z}} \tag{10}$$

$$\tilde{\boldsymbol{z}} = \frac{1}{N}\left(\boldsymbol{K}_S\boldsymbol{\alpha}^{(i-1)} + \boldsymbol{R}^{-1}(\boldsymbol{p} - \boldsymbol{t})\right) \tag{11}$$

with an $(N \times S)$-dimensional kernel matrix $\boldsymbol{K}_S$, an $(S \times S)$-dimensional regularization matrix $\boldsymbol{K}_R$ and the probabilities $\boldsymbol{p} = \frac{1}{1+\exp(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{K}_S)}$.

Zhu and Hastie [23] describes the detailed algorithm composed of the simultaneous selection of the subset $\mathcal{S}$ and the regularization parameter λ, the optimization procedure and the convergence criterion.

3.3 Extension to the Multi-class Case

We can generalize the two class model to the multi-class model. Then the objective function is

$$\mathcal{Q} = -\frac{1}{S}\sum_s \boldsymbol{t}_s^{\mathsf{T}} \log \boldsymbol{p}_s + \frac{\lambda}{2}\sum_k \boldsymbol{\alpha}_k^{\mathsf{T}} \boldsymbol{K}_R \boldsymbol{\alpha}_k \qquad (12)$$

with the probabilities $\boldsymbol{P} = [p_1, \ldots, p_S]$ obtained by

$$p_{sk} = \frac{\exp(\boldsymbol{\alpha}_k^{\mathsf{T}} \boldsymbol{k}_{S,s})}{\sum_j \exp(\boldsymbol{\alpha}_j^{\mathsf{T}} \boldsymbol{k}_{S,s})}. \qquad (13)$$

The kernel matrix $\boldsymbol{K}_s$ consists of the columns $\boldsymbol{k}_{S,s}$. The binary target vector $\boldsymbol{t}_n$ of length K uses the 1-of-K coding scheme so that all components but t_{nk} are 0 if the feature $\boldsymbol{\phi}_n$ is from class $\mathcal{C}_k$. The complete parameter matrix is $\boldsymbol{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K]$ collecting the individual parameter vectors.

In the Newton-Raphson procedure in (10) and (11) we have to use one $\boldsymbol{R}_k$ and one $\tilde{\boldsymbol{z}}$ for each class. In consequence of the over-determined system we need to apply the pseudo-inverse instead of the normal inverse.

4 Discriminative Random Field

Following Kumar et al. [12] we combine a logistic classifier – in our case the IVMs – with a smoothing over the image label field and obtain a Discriminative Random Field. We use the following simplified model

$$P(\boldsymbol{y}|\boldsymbol{X}) = \frac{1}{Z}\exp\left[\sum_{n\in\mathcal{I}} \log P(y_n|\mathbf{x}_n) + \beta \sum_{n\in\mathcal{I}}\sum_{m\in\mathcal{N}_n} \delta(y_n, y_m)\right], \qquad (14)$$

where $\mathbf{x}_n$ is the observed feature vector from the nth site, $\mathcal{I}$ being the set of all sites and δ being the Kronecker delta function. For each site there is one label $y_n \in \mathcal{C} = \{1, \ldots, K\}$. The normalization constant is Z and the interaction parameter is β. The first term in (14) models the association of the site n with $\mathcal{C}_n$ defined by the probabilistic output of the IVM. The second term describes the interaction potential as a Potts model over a 2D lattice penalizing every dissimilar pair of labels and therefore heterogeneous regions. We choose a simplified model with a data-independent term, which can be seen as a special case of Conditional Random Field. The set of neighbors of y_n is given by $\mathcal{N}_n$.

5. Experimental Setup and Results

In our experiments we consider the performance of the IVMs regarding to the classification accuracy and the complexity and compare it to the Gaussian Maximum Likelihood classification and the SVMs, which both are standard in land cover classification.

5.1 Dataset

The study site is dominated by agriculture and characterized by typical spatial patterns and temporal variation caused by differences in the crop phenology, with cereals and sugar beets being the main crops. The data set contains two Landsat 5 TM images from April 3 and May 28, 2005 (i. e. 12 bands in total). The classification is aiming on eight land cover classes, focused on ARABLE CROPS, CEREALS, FOREST, GRASSLAND, ORCHARDS, RAPESEED, ROOT CROPS, and URBAN. A map from a detailed field survey was available for generating the training and test sample sets.

5.2 Methods

We transform the features into kernel space with a radial basis function

$$K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\frac{|\mathbf{x}_n - \mathbf{x}_m|^2}{2\sigma^2}\right) \qquad (15)$$

and determine the kernel parameter σ^2 through cross-validation.

To investigate the possible influence of the number of training samples on the performance of the classifier, training sets with different size N were generated, containing 50 and 200, respectively.

In the IVMs approach we use N samples and split off $1/5$ of them for the tuning set to determine the regularization parameter λ. We start with $\lambda = \exp(0)$ and decrease with a factor of $\exp(1)$.

To achieve an optimal segmentation result with the Discriminative Random Field we test possible $\beta > 0$ and choose the one yielding the lowest training error. We perform the optimization with loopy-belief propagation (e. g. [2]) using an own implementation.

The SVM classification is based on imageSVM [6], a freely available IDL/ENVI implementation. imageSVM is based on the LIBSVM approach by Chang and Lin [4] for the training. A Gaussian radial basis function kernel is used and the kernel parameter is determined via 3-fold cross validation.

5.3 Results and Discussion

Comparing the results achieved by three different algorithms (Table 1) it can be assessed, that the SVM and IVM perform equally and outperform the Gaussian Maximum Likelihood classifier in terms of accuracy.

	MLC		SVM			IVM			DRF	
# training points per class	acc	κ	acc	κ	# support points	acc	κ	# import points	acc	κ
50	75.19	0.68	79.50	0.74	298	79.10	0.74	21	**84.90**	**0.80**
200	80.49	0.75	81.94	0.77	902	82.30	0.77	22	**86.83**	**0.83**

Table 1. Classification results with MLC, SVM, IVM and DRF. The best result is in bold print.

Class name	MLC	SVM	IVM	DRF
ARABLE CROPS	62.62	**71.02**	65.81	63.18
CEREALS	70.24	73.27	73.78	**80.89**
FOREST	94.44	95.46	95.78	**97.57**
GRASSLAND	71.84	69.23	68.02	**74.94**
ORCHARDS	44.53	56.54	59.05	**63.96**
RAPESEED	76.35	76.25	75.96	**82.17**
ROOT CROPS	75.13	68.47	70.77	**76.37**
URBAN	80.17	84.90	84.44	**90.94**

Table 2. Accuracies with 200 training points per class. Best result in bold print.

Using a higher number of training samples for the Gaussian Maximum Likelihood classifier, the total accuracy is increased by $5.3\ \%$ compared to the classification results achieved with 50 samples per class. In contrast to this the impact of the number of training samples on the total accuracy is less dominant for the SVMs and IVMs.

The results show the positive impact of the Discriminative Random Field on the overall accuracy. The total accuracy is significantly improved by Discriminative Random Field and increased by $5.8\ \%$ and $4.5\ \%$, respectively. The positive impact of the Discriminative Random Field on the classification accuracy is underlined by a visual assessment of the classification results (Figure 1). Noise within the pixel-based results is clearly reduced and field plots appear generally more homogeneous. However, in images with complex, non-smooth structures, small regions can be eliminated and the accuracy might be decreased.

Table 2 shows, that the IVM with Discriminative Random Field has in 7 of 8 classes a higher accuracy than the other algorithms. Only the accuracy of Arable crops is higher for the SVM.

The computational cost of the SVMs is $O(N^2 N_S)$, where N_S is the number of support points. The computational cost of the IVMs is $O(N^2 N_I^2)$ with N_I as the number of import points. Table 1 shows that the number of import vectors is considerably less the than number of support vectors, particularly for the large sample set. This is in accordance with Zhu and Hastie [23], which have demonstrated that the number import vectors tends not to increase with the number of used training samples, whereas the amount of support vectors usually increase with the number of training points. Seen from the computational aspect the IVMs can have a less computational amount for training and testing than the SVMs.

6. Conclusion

We apply the Import Vector Machines with a Discriminative Random Field to the image classification in remote sensing. We tested the algorithm on a land cover classification task with a Landsat image and compared it on the one hand to the Gaussian Maximum Likelihood classifier and on the other hand to the SVMs regarding to the achieved accuracy and the complexity.

The IVMs show similar results to the SVMs and both outperform the Gaussian Maximum Likelihood classification in terms of accuracy.

In comparison to the SVMs the IVM has the advantage of probabilistic outputs, often has a lower complexity and can directly applied to the multi-class case. They probabilistic output can be incorporated into a Discriminative Random Field, which increases the classification accuracy significantly.

Within future research we will focus on an automatically determination of the regularization parameter. A fixed regularization parameter further decreases the number of used import vectors and so fastens the algorithm. First results with a data-driven regularization parameter shows promising results.

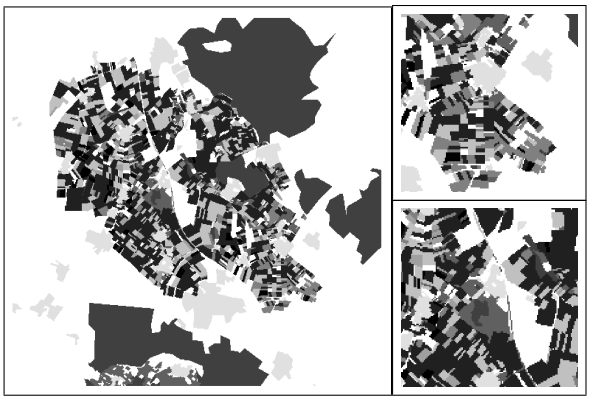
References

[1] M. Aizerman, E. Braverman, and L. Rozonoèr. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, 25(6):821–837, 1964.

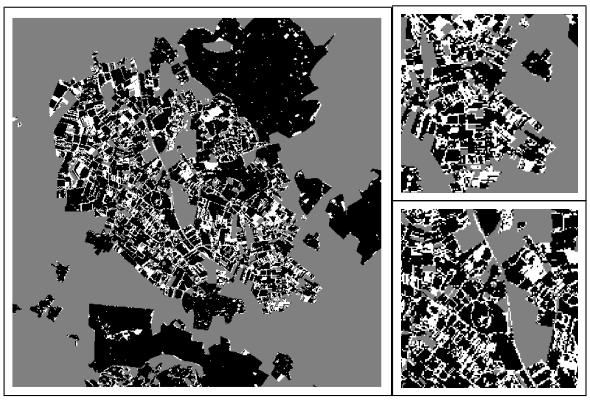
[2] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

[3] G. Cawley and N. Talbot. Efficient Model Selection for Kernel Logistic Regression. *Proc. ICPR*, 2:439–442, 2004.

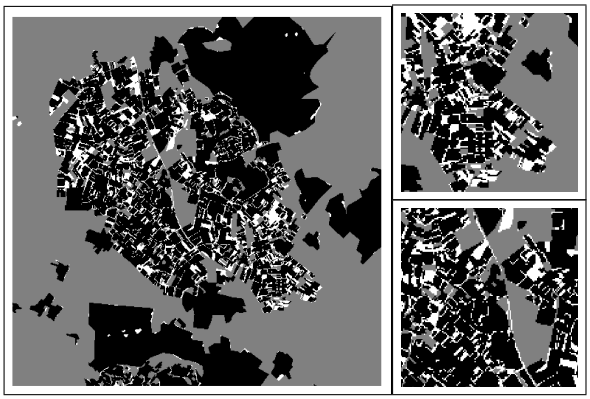
[4] C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines, 2001.



(a) Ground truth



(b) Difference IVM – ground truth



(c) Difference DRF – ground truth

Figure 1. Classification results of IVM (b) and DRF (c) in comparison with the ground truth (a). In the two lower images: white pixels are false, black are true and gray are not labeled.

[5] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Trans. PAMI*, 22(1):4–37, 2000.

[6] A. Janz, S. van der Linden, B. Waske, and P. Hostert. imageSVM-A User-Oriented Tool for Advanced Classification of Hyperspectral Data Using Support Vector Machines. In *Proceedings 5th EARSeL Workshop on Imaging Spectroscopy*, volume 1, 2007.

[7] P. Karsmakers, K. Pelckmans, and J. Suykens. Multi-Class Kernel Logistic Regression: A Fixed-Size Implementation. In *Proc. IJCNN*, pages 1756–1761, 2007.

[8] M. Katz, S. Kruger, M. Schaffoner, E. Andelic, and A. Wendemuth. Speaker Identification and Verification Using Support Vector Machines and Sparse Kernel Logistic Regression. *Lecture Notes in Computer Science*, 4153:176, 2006.

[9] S. Keerthi, K. Duan, S. Shevade, and A. Poo. A Fast Dual Algorithm for Kernel Logistic Regression. *Machine Learning*, 61(1):151–165, 2005.

[10] M. A. Komarek, P. Making Logistic Regression A Core Data Mining Tool With TR-IRLS. In *Proceedings of ICDM*, page 4, 2005.

[11] J. Koo, I. Sohn, S. Kim, and J. Lee. Structured Polychotomous Machine Diagnosis of Multiple Cancer Types Using Gene Expression. *Bioinformatics*, 22(8):950, 2006.

[12] S. Kumar and M. Hebert. Discriminative Random Fields. *IJCV*, 68(2):179–201, 2006.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. ICML*, pages 282–289. Citeseer, 2001.

[14] J. Lafferty, X. Zhu, and Y. Liu. Kernel Conditional Random Fields: Representation and Clique Selection. *Proc. ICML*, page 64, 2004.

[15] J. A. Richards. Analysis of Remotely Sensed Data: The Formative Decades and the Future. *IEEE Trans. Geosci. and Remote Sens.*, 43(3):422–432, 2005.

[16] V. Roth. Probabilistic Discriminative Kernel Classifiers for Multi-class Problems. *Lecture Notes in Computer Science*, pages 246–253, 2001.

[17] A. K. Shackelford and C. H. Davis. A Combined Fuzzy Pixel-Based and Object-Based Approach for Classification of High-Resolution Multispectral Data Over Urban Areas. *IEEE Trans. Geosci. and Remote Sens.*, 41(10 Part 1):2354–2363, 2003.

[18] M. Song, D. L. Civco, and J. D. Hurd. A Competitive Pixel-Object Approach for Land Cover Vlassification. *Int. J. Remote. Sens.*, 26(22):4981–4997, 2005.

[19] Z. Tu. Probabilistic Boosting-Tree: Learning Discriminative Models for Classification, Recognition, and Clustering. In *Proc. of the ICCV*, volume 3. Citeseer, 2005.

[20] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[21] B. Waske and S. Van der Linden. Classifying Multilevel Imagery from SAR and Optical Sensors by Decision Fusion. *IEEE Trans. Geosci. and Remote Sens.*, 46(5):1457–1466, 2008.

[22] J. Zhu and T. Hastie. Classification of Gene Microarrays by Penalized Logistic Regression. *Biostatistics*, 5(3):427, 2004.

[23] J. Zhu and T. Hastie. Kernel Logistic Regression and the Import Vector Machine. *J. Comput. Grap.*, 14(1):185–205, 2005.