Bayesian Logistic Regression for Classification of Tabular Data

Abstract

We extend the Relevance Vector Machine (RVM) framework to handle cases of table-structured data such as image blocks and image descriptors. This is achieved by coupling the regularization coefficients of rows and columns of features. We present two variants of this new gridRVM framework, based on the way in which the regularization coefficients of the rows and columns are combined. Appropriate variational optimization algorithms are derived for inference within this framework. The consequent reduction in the number of parameters from the product of the table's dimensions to the sum of its dimensions allows for better performance in the face of small training sets, resulting in improved resistance to overfitting problems, as well as providing better interpretation of results. These properties are demonstrated on a well-known synthetic data-set as well as on a modern and challenging visual identification benchmark.

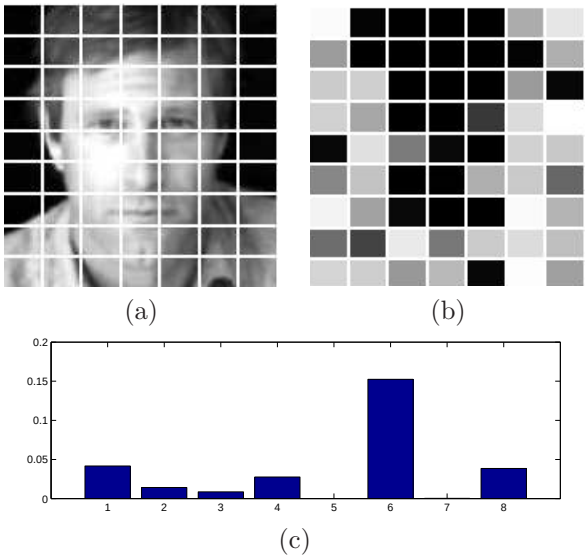

Figure 1. The illustration of "grid" approach on the LFW data. Each image is split into 63 blocks (a) and for each block 8 descriptors are computed. GridRVM assigns individual regularization coefficients for each block and each descriptor. The relevance of blocks (the darker the more informative) is shown in (b) and the relevance of descriptors (inverse regularization coefficient) is shown in (c)

1. Introduction

Generalized linear models have been a popular approach to classification problems for decades. Special attention is often paid to obtain sparse decision rules, i.e. classifiers for which most of the assigned weights equal zero. Within a Bayesian framework the detection of relevant features can be done automatically by assigning an individual regularization coefficient to each weight. This process is called automatic relevance determination (ARD). The Relevance Vector Machine (RVM) is an important example of the successful application of ARD to logistic regression (Tipping, 2001).

In this paper we generalize the RVM framework to the case of tabular data, i.e. cases where an object is described by a matrix of features. Tabular data arises in many domains (see section 2). Of course, it is always possible to convert tables to feature vectors and run standard classification algorithms. We will, however, show that this may sometimes lead to overfitting. Here, we suggest assigning individual regularization coefficients to each column and row of the table. The regularization coefficient of the feature in position ij in the table is then the result of the composition of the coefficients for the i^{th} row and the j^{th} column. We consider two variants of such compositions: product and summation of regularization coefficients, thus deriving p- and s-gridRVM models. Variational inference is used to obtain iterative equations for learning in these models. We demonstrate results on synthetic and real-world problems and show that the gridRVM approach prevents overfitting in case of small datasets. In particular we address the problem of same/not-same face classification in the Labeled Faces in the Wild (LFW) image set (Huang et al., 2007). We convert each image to a tabular presentation by computing a set of

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

descriptors on distinct image blocks. GridRVM is then applied to find the most relevant blocks and descriptors (see fig. 1).

The rest of the paper is organized as follows. Motivation and related work are given in sections 2 and 3 respectively. Section 4 presents the definitions of the p- and s-gridRVM models and establishes the notation used thereafter. Iterative learning algorithms based on variational inference are described in section 5. We conclude with experiments on illustrative and real-world problems in section 6.

2. Motivation

In classical machine learning theory, a training set consists of a number of objects (precedents), each represented as a vector of features. It is assumed that there is no hierarchy in the space of features. This is not, however, always an optimal representation. In some cases, a tabular presentation is more convenient. Objects are then described by a number of features that form a table rather than a single vector.

A natural example of such case arises in a region/descriptor-based framework for image analysis. Within this framework, an image is split into several regions (blocks) and a set of descriptors is then computed for each region. Then, we may associate each feature with the pair region/descriptor and form a tabular view of a single image. Note that often the number of features extracted from single image exceeds the number of images in the whole training set, resulting in increased risk of overfitting.

Another example is related to the use of radial basis functions (RBF) in classification algorithms. Traditionally, RBF depends only on the distance $\rho(\vec{x}, \vec{y}_i)$ between the object $\vec{x}$ and some predefined point $\vec{y}_i$ in the space of features $\mathbb{R}^d$

$$\phi_i(\vec{x}) = f(\rho(\vec{x}, \vec{y}_i)), \quad i \in \{1, \dots, m\}.$$

Each object is described by a vector of m RBF values. Gaussian RBFs $\phi_i(\vec{x}) = \exp(-\gamma\|\vec{x} - \vec{y}_i\|^2)$ are a popular "rule-of-thumb" choice in many classification algorithms, e.g. in logistic regression. The obvious drawback of Gaussian RBF is their low discriminative ability in the presence of numerous noisy features. To deal with noisy features one may consider basis functions consisting of a single feature

$$\phi_{ij}(\vec{x}) = f(|x_j - y_{ij}|).$$

Although representing $\vec{x}$ as a vector of $(\phi_{11}(\vec{x}), \dots, \phi_{m,d}(\vec{x}))$ is still possible, it may be more natural to form a table of m columns and d rows.

The tabular presentation of data provides new options in analyzing feature sets. In particular, we may search for relevant columns and rows instead of searching for relevant features by setting one regularization coefficient for each column and row, hence reducing the number of hyperparameters from $m \times d$ to $m+d$. With the reduced number of adjustable parameters we may expect the final classifier to have better generalization properties.

3. Related work

The idea of treating image features as tables is not new and has been considered by a number of authors. Many papers on tabular data consider the problem of dimensionality reduction (either supervised or unsupervised). In (Yang et al., 2004) 2-dimensional PCA is proposed where each data point is treated as a matrix. In (Xu et al., 2004) the authors proposed an image reconstruction criterion for obtaining the original image matrices using two low dimensional coupled subspaces, which encode the row and column subspaces of the image. They suggested an iterative method, CSA (Coupled Subspaces Analysis) to optimize this criterion. They also prove that PCA and 2D-PCA are special cases of CSA. The generalization of LDA to tabular data has been proposed in (Ye et al., 2004) and (Li & Yuan, 2005). More recently, (Yang et al., 2009) have proposed projecting images along both row and column directions, in an effort to maximize the so called Laplacian Bidirectional Maximum Margin Criterion (LBMMC). A variant of the Zero-norm SVM feature selection algorithm for tabular data was presented in (Wolf et al., 2007).

In the context of sparse methods several non-Bayesian techniques have been proposed, for example (Boser et al., 1992; Tibshirani, 1996). Automatic relevance determination was first proposed in (MacKay, 1992) which provides a Bayesian framework for determining irrelevant parameters in machine learning models. The application of ARD to generalized linear models and in particular to logistic regression was proposed in (Tipping, 2001) as the Relevance vector machine model (RVM).

Since fully Bayesian inference is intractable even for regression problems, different authors have used some approximation of the general Bayesian scheme. These include evidence maximization (MacKay, 1992; Tipping, 2001), marginalization w.r.t. the hyperparameters (Williams, 1995), and variational infer-

ence (Bishop & Tipping, 2000).

For classification, further approximations are necessary to perform inference. Various approximations of the likelihood function with a Gaussian were suggested. In (Tipping, 2001) the authors used Laplace approximation. Local variational methods were proposed in (Jaakkola & Jordan, 2000). The closely related expectation propagation (Minka, 2001) technique for approximate Bayesian inference in generalized linear models was suggested in (Qi et al., 2004). Although ARD methods have been applied successfully for the search of relevant features, objects, and basis functions in many domains, over- and underfitting of RVM was reported (Qi et al., 2004) in some cases.

4. GridRVM models

Consider a two-class classification problem with tabular data. Let $(X, \vec{t}) = \{\vec{x}_n, t_n\}_{n=1}^N$ be the training set where $t_n \in \{-1, 1\}$ are class labels and each object $\vec{x}_n$ is represented as a table of generalized features $(\phi_{ij}(\vec{x}_n))_{i,j=1}^{M_1,M_2}$. Note that we will also use one-index notation $(\phi_k(\vec{x}_n))_{k=1}^M$, $M = M_1 M_2$ when we need to treat the description of the object as a vector. Define the following probabilistic model (p-gridRVM):

$$p(\vec{t}, \vec{w}, \vec{\alpha}, \vec{\beta}|X) = p(\vec{t}|X, \vec{w})p(\vec{w}|\vec{\alpha}, \vec{\beta})p(\vec{\alpha})p(\vec{\beta}).$$

Here

$$p(\vec{t}|X, \vec{w}) = \prod_{n=1}^{N} \sigma\left(t_n \vec{w}^T \vec{\phi}(\vec{x}_n)\right), \tag{1}$$

$$p(\vec{w}|\vec{\alpha}, \vec{\beta}) = \frac{\prod_{i,j=1}^{M_1,M_2} \sqrt{\alpha_i \beta_j}}{\sqrt{2\pi}^{M_1 M_2}} \exp\left(-\frac{1}{2} \sum_{i,j=1}^{M_1,M_2} \alpha_i \beta_j w_{ij}^2\right), \tag{2}$$

$$p(\vec{\alpha}) = \prod_{i=1}^{M_1} \mathcal{G}(\alpha_i|a_0, b_0), \tag{3}$$

$$p(\vec{\beta}) = \prod_{j=1}^{M_2} \mathcal{G}(\beta_j|c_0, d_0), \tag{4}$$

where $\sigma(y) = 1/(1 + \exp(-y))$ is a logistic function, $\mathcal{G}(\alpha_i|a_0, b_0)$ stands for a gamma distribution over α_i with parameters a_0, b_0 and all $\alpha_i, \beta_j \geq 0$. Note that the number of regularization coefficients $\vec{\alpha}$ and $\vec{\beta}$ is M_1+M_2 while the number of weights is $M_1 M_2$. Within this model we assign independent regularization coefficients to each row and column of the tabular presentation. The regularization coefficient for the weight w_{ij} is the result of a combination of α_i and β_j. In p-gridRVM we take the product of the two. Alternatively, we may consider the sum, i.e.

$$p(\vec{w}|\vec{\alpha}, \vec{\beta}) = \frac{\prod_{i,j=1}^{M_1,M_2} \sqrt{\alpha_i + \beta_j}}{\sqrt{2\pi}^{M_1 M_2}} \times \exp\left(-\frac{1}{2} \sum_{i,j=1}^{M_1,M_2} (\alpha_i + \beta_j) w_{ij}^2\right), \tag{5}$$

We refer to this model as s-gridRVM.

In both cases we consider the joint influence of the row and column of each table entry on the associated feature weight. However the models have one important distinction. In the case of s-gridRVM, large values of α_i mean that all features from the i^{th} row have values at least as large as the regularization coefficient. This, since $\alpha_i + \beta_j > \alpha_i$ for all admissible β_j. The same of course holds for large values of β_j. In p-gridRVM the situation is different. Large values of, say, α_i do not necessarily imply large values of the regularization coefficient for a particular weight w_{ij} since the coefficient β_j may have a small value. Thus we may expect a different behavior from these models.

5. Variational learning

Variational methods (Jordan et al., 1998) are popular technique for inference in Bayesian models. These methods allow to move from hardly computable model evidence to its lower bound, which is much simpler for estimation. In this section we first briefly describe basic ideas of the variational approach and then show its application for learning in the p- and s-gridRVM models.

5.1. Global variational inference

Suppose we are given a probabilistic model with variables $(\vec{t}, \vec{\theta})$, where $\vec{t}$ is observable and $\vec{\theta}$ is not. We would like to estimate the model evidence

$$p(\vec{t}) = \int p(\vec{t}, \vec{\theta}) d\vec{\theta},$$

which we assume cannot be found analytically. Variational inference introduces here some distribution over the unobservable variables $q(\vec{\theta})$. Using this distribution the model evidence can be decomposed as follows

$$\log p(\vec{t}) = \mathcal{L}(q) + KL(q||p(\vec{\theta}|\vec{t})),$$

where

$$\mathcal{L} = \int q(\vec{\theta}) \log \frac{p(\vec{\theta}|\vec{t})}{q(\vec{\theta})} d\vec{\theta} \tag{6}$$

and $KL(q||p)$ is a Kullback-Leibler divergence between two distributions. Since $KL(q||p) \geq 0$, $\mathcal{L}$ is a lower bound on the log-evidence. Besides, $\log p(\vec{t})$ does not depend on $q(\vec{\theta})$ and hence maximization of the lower bound $\mathcal{L}$ w.r.t. $q(\vec{\theta})$ is equivalent to minimization of the KL divergence between $q(\vec{\theta})$ and posterior distribution $p(\vec{\theta}|\vec{t})$.

Now consider the case when a distribution $q(\vec{\theta})$ has a factorized form

$$q(\vec{\theta}) = \prod_i q_i(\vec{\theta}_i).$$

Here $\{\vec{\theta}_i\}$ is some decomposition of a full set of variables so that $\vec{\theta} = \sqcup_i \vec{\theta}_i$. In (Jordan et al., 1998) it's shown that maximization of (6) can be done iteratively by the following recalculation formula:

$$q_i(\vec{\theta}_i) = \frac{1}{Z} \exp\left(\int \log p(\vec{t}, \vec{\theta}) \prod_{j \neq i} q_j(\vec{\theta}_j) d\vec{\theta}_j \right), \quad (7)$$

where Z is a normalization constant ensuring that $q_i(\vec{\theta}_i)$ is a distribution. In this recalculation process the lower bound (6) monotonically increases.

5.2. Local variational inference

Global variational methods are supposed to move from the hardly computable model evidence to its lower bound. However, in many practical models (including p- and s-gridRVM) this lower bound is still analytically intractable. The local variational approach (Jaakkola & Jordan, 2000) introduces a further bound on $p(\vec{\theta}|\vec{t})$:

$$p(\vec{\theta}|\vec{t}) \geq F(\vec{\theta}, \vec{t}, \vec{\xi}).$$

This bound is tight for some particular value of $\vec{\xi}$ and so it is local. Substituting this bound into (6) gives the following result:

$$\log p(\vec{t}) \geq \mathcal{L} \geq \mathcal{L}_{local} = \int q(\vec{\theta}) \log \frac{F(\vec{\theta}, \vec{t}, \vec{\xi})}{q(\vec{\theta})} d\vec{\theta}.$$

The last expression can be optimized w.r.t. $q(\vec{\theta})$ and $\vec{\xi}$ for some sensible choice of local variational bound.

5.3. p-gridRVM

In a classification problem we wish to calculate

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) = \int p(t_{new}|\vec{x}_{new}, \vec{w}) \times p(\vec{w}, \vec{\alpha}, \vec{\beta}|\vec{t}, X) d\vec{w} d\vec{\alpha} d\vec{\beta} \quad (8)$$

for any new object $\vec{x}_{new}$. For the model p-gridRVM (1)-(4) as well as for the model s-gridRVM (1),(5),(3),(4) this integration is intractable and hence some approximation scheme is needed. Here we use the variational approach, which has been successfully applied for the conventional RVM model in (Bishop & Tipping, 2000), and try to find a variational approximation $q(\vec{w}, \vec{\alpha}, \vec{\beta})$ to the true posterior $p(\vec{w}, \vec{\alpha}, \vec{\beta}|\vec{t}, X)$ in the following family of factorized distributions:

$$q(\vec{w}, \vec{\alpha}, \vec{\beta}) = q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}). \quad (9)$$

Then (8) can be reduced to integration over the factorized distribution q:

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) \simeq \int p(t_{new}|\vec{x}_{new}, \vec{w}) \times q(\vec{w}, \vec{\alpha}, \vec{\beta}) d\vec{w} d\vec{\alpha} d\vec{\beta}. \quad (10)$$

This integration is still intractable. However, if we are interested only in a point estimate for t_{new}, a useful approximation is

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) \simeq p(t_{new}|\vec{x}_{new}, \mathbb{E}_{\vec{w}} \vec{w}),$$

where the symbol $\mathbb{E}_{\vec{w}}$ stands for expectation w.r.t. the factorized distribution $q_{\vec{w}}(\vec{w})$.

Use of the global variational approach for the p-gridRVM model leads to estimation of the following lower bound of the model log-evidence:

$$\log p(\vec{t}|X) \geq \mathcal{L} = \int \log \frac{p(\vec{t}|X, \vec{w}) p(\vec{w}|\vec{\alpha}, \vec{\beta}) p(\vec{\alpha}) p(\vec{\beta})}{q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta})} \times q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) d\vec{w} d\vec{\alpha} d\vec{\beta}. \quad (11)$$

This integral cannot be taken analytically. Following the variational method for the conventional RVM model (Bishop & Tipping, 2000) we use here the local variational approach and introduce the Jaakkola-Jordan inequality (Jaakkola & Jordan, 2000) for the likelihood function:

$$p(\vec{t}|X, \vec{w}) \geq F(\vec{t}, X, \vec{w}, \vec{\xi}) = \prod_{n=1}^{N} \sigma(\xi_n) \exp\left(\frac{z_n - \xi_n}{2} - \lambda(\xi_n)(z_n^2 - \xi_n^2) \right), \quad (12)$$

where $\sigma(y) = 1/(1 + \exp(-y))$ — sigmoid function, $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$, $z_n = t_n \vec{w}^T \vec{\phi}(\vec{x}_n)$. This bound is tight for $\xi_n = z_n$. Then substituting the inequality (12) into the lower bound (11) we obtain:

$$\mathcal{L} \geq \mathcal{L}_{local} = \int \log \frac{F(\vec{t}, X, \vec{w}, \vec{\xi}) p(\vec{w}|\vec{\alpha}, \vec{\beta}) p(\vec{\alpha}) p(\vec{\beta})}{q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta})} \times q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) d\vec{w} d\vec{\alpha} d\vec{\beta}. \quad (13)$$

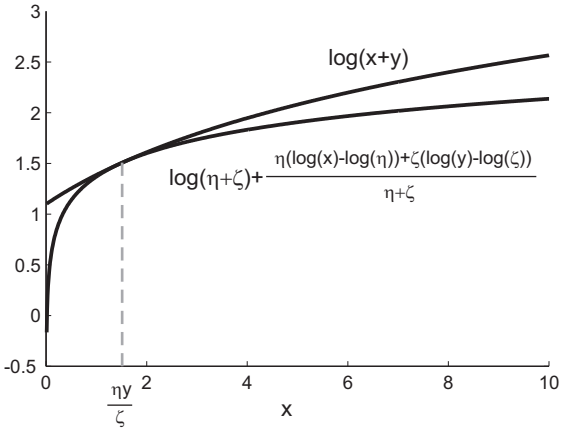

Figure 2. One-dimensional projection of the bound (23) for parameters $y = 3, \eta = 2, \zeta = 4$.

Maximization of the criterion function (13) w.r.t. distributions $q_{\vec{w}}(\vec{w}), q_{\vec{\alpha}}(\vec{\alpha}), q_{\vec{\beta}}(\vec{\beta})$ and variational parameters $\vec{\xi}$ leads to the following result:

$$q_{\vec{w}}(\vec{w}) = \mathcal{N}(\vec{w}|\vec{\mu}, \Sigma), \tag{14}$$

$$q_{\vec{\alpha}}(\vec{\alpha}) = \prod_{i=1}^{M_1} \mathcal{G}(\alpha_i|a_i, b_i), \tag{15}$$

$$q_{\vec{\beta}}(\vec{\beta}) = \prod_{j=1}^{M_2} \mathcal{G}(\beta_j|c_j, d_j), \tag{16}$$

where

$$\Sigma = \left(\mathrm{diag}(\mathbb{E}_{\vec{\alpha}}\alpha_i\mathbb{E}_{\vec{\beta}}\beta_j) + 2\Phi^T\Lambda\Phi\right)^{-1},\ \Lambda = \mathrm{diag}(\lambda(\xi_n)),$$

$$\vec{\mu} = \frac{1}{2}\Sigma\Phi^T\vec{t},$$

$$a_i = a_0 + \frac{M_2}{2}, \quad b_i = b_0 + \frac{1}{2}\sum_{j=1}^{M_2}\mathbb{E}_{\vec{\beta}}\beta_j\mathbb{E}_{\vec{w}}w_{ij}^2,$$

$$c_j = c_0 + \frac{M_1}{2}, \quad d_j = d_0 + \frac{1}{2}\sum_{i=1}^{M_1}\mathbb{E}_{\vec{\alpha}}\alpha_i\mathbb{E}_{\vec{w}}w_{ij}^2,$$

$$\xi_n^2 = \vec{\phi}^T(\vec{x}_n)\mathbb{E}_{\vec{w}}\vec{w}\vec{w}^T\vec{\phi}(\vec{x}_n)\ (\vec{\xi}^2 = \mathrm{diag}(\Phi S\Phi^T)).$$

The necessary statistics are calculated as follows:

$$\mathbb{E}_{\vec{w}}\vec{w} = \vec{\mu}, \tag{17}$$

$$\mathbb{E}_{\vec{w}}w_{ij}^2 = S_{ij,ij} + \mu_{ij}^2, \tag{18}$$

$$\mathbb{E}_{\vec{\alpha}}\alpha_i = \frac{a_i}{b_i}, \tag{19}$$

$$\mathbb{E}_{\vec{\alpha}}\log\alpha_i = \Psi(a_i) - \log b_i, \tag{20}$$

$$\mathbb{E}_{\vec{\beta}}\beta_j = \frac{c_i}{d_i}, \tag{21}$$

$$\mathbb{E}_{\vec{\beta}}\log\beta_j = \Psi(c_j) - \log d_j, \tag{22}$$

where $\Psi(a) = \frac{d}{da}\log\Gamma(a)$ — digamma function.

5.4. s-gridRVM

Similar to the previous case we propose to apply the variational approach for the s-gridRVM model. In this way we try to find a variational approximation q to the true posterior $p(\vec{w}, \vec{\alpha}, \vec{\beta}|\vec{t}, X)$ in the family of factorized distributions (9) by optimizing the lower bound (13). However, in the case of the s-gridRVM model the criterion function (13) becomes intractable and we need a further lower bound in the sense of the local variational methods. For this reason let us consider the function $f(x, y) = \log(x + y)$. This function is strictly concave. Now let us make the change of variables $x_1 = \log(x), y_1 = \log(y)$ and consider the function $f_1(x_1, y_1) = f(\exp(x_1), \exp(y_1)) = \log(\exp(x_1) + \exp(y_1))$. The function f_1 is convex and hence satisfies the following inequality:

$$f_1(x_1, y_1) \geq \frac{\partial f_1}{\partial x_1}(\eta)(x_1 - \eta) + \frac{\partial f_1}{\partial y_1}(\zeta)(y_1 - \zeta) + f_1(\eta, \zeta)$$

for arbitrary η and ζ. This inequality is just a relation between the function and its tangent line and becomes equality when $x_1 = \eta, y_1 = \zeta$. Moving back to initial variables x, y, we obtain the following variational bound:

$$\log(x + y) \geq \log(\eta + \zeta) + \frac{\eta(\log(x) - \log(\eta)) + \zeta(\log(y) - \log(\zeta))}{\eta + \zeta}, \tag{23}$$

which is tight when $x/y = \eta/\zeta$. One-dimensional projection of this bound is illustrated in Figure 2. Inequality (23) leads to the following bound on $\log p(\vec{w}|\vec{\alpha}, \vec{\beta})$:

$$\log p(\vec{w}|\vec{\alpha}, \vec{\beta}) = \frac{1}{2}\sum_{i,j=1}^{M_1,M_2}[\log(\alpha_i+\beta_j) - (\alpha_i+\beta_j)w_{ij}^2] -$$

$$\frac{M_1M_2}{2}\log 2\pi \geq G(\vec{w}, \vec{\alpha}, \vec{\beta}, \vec{\eta}, \vec{\zeta}) = \frac{1}{2}\sum_{i,j=1}^{M_1,M_2}\left[\log(\eta_{ij}+\zeta_{ij}) + \frac{\eta_{ij}(\log(\alpha_i) - \log(\eta_{ij}))}{\eta_{ij} + \zeta_{ij}} + \frac{\zeta_{ij}(\log(\beta_j) - \log(\zeta_{ij}))}{\eta_{ij} + \zeta_{ij}}\right] -$$

$$\frac{1}{2}\sum_{i,j=1}^{M_1,M_2}(\alpha_i + \beta_j)w_{ij}^2 - \frac{M_1M_2}{2}\log 2\pi.$$

This bound is tight if $\eta_{ij} = \alpha_i$ and $\zeta_{ij} = \beta_j$. Substituting this inequality into (13) we obtain:

$$\log p(\vec{t}|X) \geq \int \log\frac{F(\vec{t}, X, \vec{w}, \vec{\xi})G(\vec{w}, \vec{\alpha}, \vec{\beta}, \vec{\eta}, \vec{\zeta})p(\vec{\alpha})p(\vec{\beta})}{q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})} \times q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})d\vec{w}d\vec{\alpha}d\vec{\beta}. \tag{24}$$

Maximization of the criterion function (24) w.r.t. distributions $q_{\vec{w}}(\vec{w}), q_{\vec{\alpha}}(\vec{\alpha}), q_{\vec{\beta}}(\vec{\beta})$ and variational parameters $\vec{\xi}, \vec{\eta}, \vec{\zeta}$ leads to the formula (14)-(16), where

$$\Sigma = \left(\mathrm{diag}(\mathbb{E}_{\vec{\alpha}}\alpha_i + \mathbb{E}_{\vec{\beta}}\beta_j) + 2\Phi^T\Lambda\Phi\right)^{-1},$$
$$\Lambda = \mathrm{diag}(\lambda(\xi_n)),$$
$$\vec{\mu} = \frac{1}{2}\Sigma\Phi^T\vec{t},$$
$$a_i = a_0 + \frac{1}{2}\sum_{j=1}^{M_2}\frac{\eta_{ij}}{\eta_{ij}+\zeta_{ij}}, \quad b_i = b_0 + \frac{1}{2}\sum_{j=1}^{M_2}\mathbb{E}_{\vec{w}}w_{ij}^2,$$
$$c_j = c_0 + \frac{1}{2}\sum_{i=1}^{M_1}\frac{\zeta_{ij}}{\eta_{ij}+\zeta_{ij}}, \quad d_j = d_0 + \frac{1}{2}\sum_{i=1}^{M_1}\mathbb{E}_{\vec{w}}w_{ij}^2$$
$$\eta_{ij} = \exp(\mathbb{E}_{\vec{\alpha}}\log\alpha_i), \ \zeta_{ij} = \exp(\mathbb{E}_{\vec{\beta}}\log\beta_j),$$
$$\xi_n^2 = \vec{\phi}^T(\vec{x}_n)\mathbb{E}_{\vec{w}}\vec{w}\vec{w}^T\vec{\phi}(\vec{x}_n) \quad (\vec{\xi}^2 = \mathrm{diag}(\Phi S\Phi^T))$$

The necessary statistics are still calculated using (17)–(22).

Table 1. Train/Test error/number of relevant weights for standard variational RVM, p- and s-gridRVM respectively. Rows $1, 4, 7$ correspond to direct classification on top of the vector of features. Rows $2, 5, 8$ correspond to the use of RBFs (25) whilst rows $3, 6, 9$ correspond to the use of basis functions (26). Note that in the presence of noisy features (the last 6 rows) the accuracy of RBF classifiers decreases. With the increase of M standard RVM also overfits the training data.

d	M	RVM	P-GRIDRVM	S-GRIDRVM
2	2	27/27.08/ 2	27/27.08/ 2	27/27.08/ 2
2	200	15.5/20.44/ 40	15.5/20.46/ 38	15.5/21.22/ 196
2	400	16.50/22.50/ 33	15.50/22.24/ 14	17.50/22.82/ 274
4	4	26.50/27.96/ 2	26.50/27.94/ 2	26/28.70/ 4
4	200	3.5/31.12/ 160	3.5/29.64/ 158	1.5/29.86/ 200
4	800	16.5/23.22/ 65	16/22.26/ 53	17/22.86/ 271
17	17	27/28.78/ 6	27/28.78/ 6	24.5/30.06/ 15
17	200	0/49.7/ 200	0/40.40/ 200	0/40.40/ 200
17	3400	6.5/29.46/ 125	14.5/22.74/ 195	15/23.40/ 271

6. Experiments

We start with an artificial problem and then consider real-world problem for which tabular presentation of data is natural.

6.1. Toy example

First consider an artificial dataset[1] taken from (Friedman et al., 2001). This is a 2-class problem with 200 objects in the training set and 5000 objects in the test set. The feature space is two-dimensional and the data are generated from a specified distribution with Bayesian error rate 19%. The optimal discriminative surface is non-linear. RVM with 200 basis functions

$$\phi_j(\vec{x}) = \exp\left(-\frac{\|\vec{x}-\vec{x}_j\|^2}{2\sigma^2}\right) \tag{25}$$

and $\sigma = 0.3$ almost achieves Bayesian error level (see table 1). We considered two modifications of the problem by adding 2 and 15 noisy features respectively. In both cases we treated objects first as initial vectors (corresponding to a linear hyperplane), then as vectors of basis functions values (25), and third as tables of basis functions evaluated per each dimension

$$\phi_{ij}(\vec{x}) = \exp\left(-\frac{(x_i - x_{ji})^2}{2\sigma^2}\right). \tag{26}$$

In the latter case we may either treat the object as the vector of basis functions and run standard RVM (or another classifier) on it, or as a table and then run gridRVM. All classifiers were trained on the set of 200 objects and then run on the test set of 5000 objects. The results are shown in table 1. It can be seen that performance of RVM with basis functions (25) degrades in the presence of noisy features. Note that training error is very small (zero with 15 noisy features) thus evidently the classifiers suffer from over-fitting. Since in the case of RBFs (25) the data points are vectors, the performance of gridRVM is very similar to standard RVM. However, RVMs with basis functions (26) are more stable and maintain their accuracy. In the case of 2 noisy features, the difference between standard RVM and gridRVM is insignificant but in the presence of 15 noisy features RVM overfits the data since 3400 regularization coefficients are to be adjusted. In gridRVMs there are only $17 + 200 = 217$ regularization coefficients in this case. Both of them have almost the same accuracy being less affected by over-fitting. One may consider the use of 3400 basis

[1] http://www-stat.stanford.edu/~ hastie/ /ElemStatLearnII/figures2.pdf to view http://www-stat.stanford.edu/~tibs/ElemStatLearn/ /datasets/mixture.example.data to download

functions for 200 of training objects in 17 dimensional space to be somewhat artificial but the results of RVMs on 200 RBFs of the form (25) are disastrous. Besides the training and test errors, the number of relevant weights (the ones which exceed 0.01) is also shown. Both p- and s-gridRVM tend to be sparse with slight advantage of p-gridRVM.

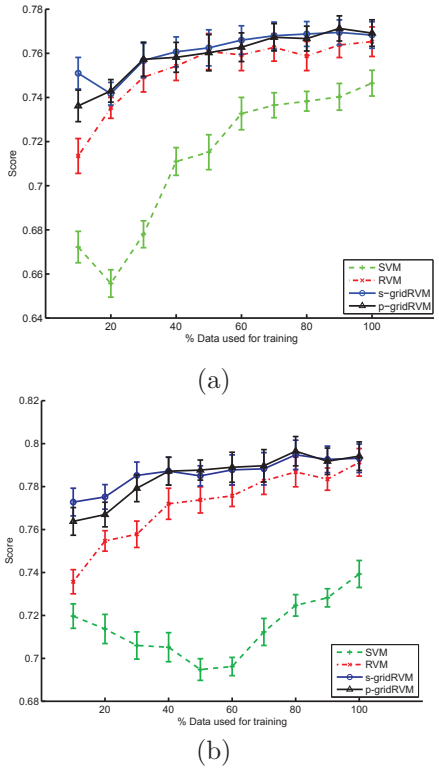

Figure 3. LFW results. Please see the text for more details

6.2. Face image pair-matching

We test our method on the Labeled Faces in the Wild (LFW) pair-matching benchmark (Huang et al., 2007). The LFW data set provides around 13,000 facial images of 5,749 individuals, each having anywhere from one to 150 images. These images were automatically harvested from news websites and thus present faces under challenging, unconstrained viewing conditions. The goal of the benchmark is to determine, given a pair of images from the collection, whether the two images match (portray the same subject) or not.

We represent the images using the following four image descriptors: Local Binary Patterns (LBP) (Ojala et al., 2001), Center Symmetric LBP (CSLBP) (Heikkilä et al., 2006), and the Three and Four Patch LBP descriptors (TPLBP and FPLBP resp.) (Wolf et al., 2008). Each face image was subdivided into 63 non-overlapping blocks of 23×18 pixels centered on the face. A separate histogram of codes was computed for each block, with 59 values for the uniform version of the LBP descriptor, sixteen values for each of the CSLBP and FPLBP descriptors, and 256 values for the TPLBP descriptor.

Each pair of images to be compared is represented by one table of similarity values. The rows of the tables correspond to types of similarities values, and the columns correspond to the 63 facial regions depicted in Figure 1(a). The types of similarity values are all possible combinations of the four image representation above, and four histogram distances and similarities.

These four different histogram distances/similarities are computed block by block between the corresponding histograms of the two images. They are the L2 norm, the Hellinger Distance obtained by taking the square root of the histogram values, the so called One-Shot Similarity (OSS) measure (Wolf et al., 2008) (using code made available by the authors), and OSS applied to the square root of the histogram values. To compute OSS scores we used 1,200 images in one of the training splits as a "negative" training set.

We report our results in Figure 3 where the pair-matching performance of s-gridRVM and p-gridRVM is compared against two baseline methods. Both figures plot classification scores across the ten-folds of the LFW benchmark, along with standard error values for different amounts of training (measured as the percentage of nine splits used as a training set). Figure 3(a) presents results using an 8×63 features of L2 and Hellinger distances between the four image descriptors; in Figure 3(b) we add also the four OSS scores and four OSS scores applied to the square roots of the histogram values.

As baseline methods we take linear SVM and standard RVM. As can be seen, the gridRVM methods show a clear advantage over both baseline methods. This is particularly true when only a small amount of training data is available. Although this advantage diminishes as more training is made available, both grid methods remain superior. Note that the results improve the ones reported in (Wolf et al., 2007), where the same features were used for the whole image and the reported accuracy was 0.7847. p-gridRVM and s-gridRVM showed 0.7934 and 0.7942 of correct answers respectively. Note that since the publication of (Wolf et al., 2007), higher performance rates were reported on this benchmark. These, however, required additional training information, were obtained through a different protocol, or made possible by further processing of the images.

7. Conclusions

The experiments allow us to draw some conclusions. The first observation is that in some learning scenarios gridRVM is significantly more robust w.r.t. overfitting than standard RVM. This is particularly true for the case of small training samples with large amounts of basis functions. It is important to stress that in case of large samples both standard and gridRVMs show almost identical results, so gridRVMs are not affected by underfitting although we reduced the number of adjustable regularization coefficients. The second observation is that both gridRVMs are sparse both in terms of regularization coefficients (many of them having large values) and in terms of the weights (many of whom are close to zero). Therefore, this useful and important property of standard RVM is kept. The grid approach can be straightforwardly generalized for the case of tensors (multidimensional tables) as well. For example we could treat the blocks in Figure 1 as a two-dimensional array hence obtaining a third dimension (together with the descriptor dimension) in the objects' description. Finally one might also consider regression problems with tabular data in similar manner.

References

Bishop, C. and Tipping, M. Variational relevance vector machine. In UAI, 2000.

Boser, B. E., Guyon, I., and Vapnik, V. A training algorithm for optimal margin classifiers. In COLT, pp. 144–152, 1992.

Friedman, J., Hastie, T., and Tibshirani, R. The Elements of Statistical Learning. Springer, 2001.

Heikkilä, M., Pietikäinen, M., and Schmid, C. Description of interest regions with center-symmetric local binary patterns. In Computer Vision, Graphics and Image Processing, 5th Indian Conference, pp. 58–69, 2006.

Huang, G.B., Ramesh, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. UMASS, TR 07-49, 2007.

Jaakkola, T. and Jordan, M. Bayesian parameter estimation through variational methods,. Statistics and Computing, 10:25–37, 2000.

Jordan, M. I., Gharamani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. In M. I. Jordan eds. Learning in Graphical Models, pp. 105–162, 1998.

Li, M. and Yuan, B. 2D–LDA: A statistical linear discriminant analysis for image matrix. Pattern Recognition Letters, 26(5):527–532, 2005.

MacKay, D. Bayesian interpolation. Neural Computation, 4(3):415–447, 1992.

Minka, T. Expectation propagation for approximate bayesian inference. In UAI, 2001.

Ojala, T., Pietikäinen, M., and Mäenpää, T. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. In ICAPR, 2001.

Qi, Y., Minka, T., Picard, R., and Gharamani, Z. Predictive automatic relevance determination by expectation propagation. In ICML, 2004.

Tibshirani, R. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., 58(1):267–288, 1996.

Tipping, M. Sparse Bayesian learning and the Relevance Vector Machine. Journal of Machine Learning Research, 1:211–244, 2001.

Williams, P. Bayesian regularization and pruning using a Laplace prior. Neural Computation, 7(1):117–143, 1995.

Wolf, L., Jhuang, H., and Hazan, T. Modeling appearances with low-rank SVM. In CVPR, 2007.

Wolf, L., Hassner, T., and Taigman, Y. Descriptor based methods in the wild. In Real-Life Images workshop at ECCV, October 2008.

Xu, D., Yan, S., Zhang, L, Liu, Z., and Zhang, H. Coupled subspaces analysis. Technical Report MSR-TR-2004-106, 2004.

Yang, J., Zhang, D., Frangi, A., and Yang, J. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(1):131–137, 2004.

Yang, W., Wang, J., Ren, M., Yang, J., Zhang, L., and Liu, G. Feature extraction based on laplacian bidirectional maximum margin criterion. Pattern Recognition, 42(11):2327–2334, 2009.

Ye, J., Janardan, R., and Li, Q. Two-dimensional linear discriminant analysis. In NIPS, 2004.