BAYESIAN CLASSIFICATION AND ACTIVE LEARNING USING l_p-PRIORS. APPLICATION TO IMAGE SEGMENTATION

Pablo Ruiz[1*], *Nicolás Pérez de la Blanca*[1], *Rafael Molina*[1] *and Aggelos K. Katsaggelos*[2]

[1]Dept. Ciencias de la Computación e I.A., Universidad de Granada, Spain.
[2] Dpt. of Electrical Engineering and Computer Science, Northwestern University, USA.
[*]e-mail:mataran@decsai.ugr.es

ABSTRACT

In this paper we utilize Bayesian modeling and inference to learn a softmax classification model which performs Supervised Classification and Active Learning. For $p < 1$, l_p-priors are used to impose sparsity on the adaptive parameters. Using variational inference, all model parameters are estimated and the posterior probabilities of the classes given the samples are calculated. A relationship between the prior model used and the independent Gaussian prior model is provided. The posterior probabilities are used to classify new samples and to define two Active Learning methods to improve classifier performance: Minimum Probability and Maximum Entropy. In the experimental section the proposed Bayesian framework is applied to Image Segmentation problems on both synthetic and real datasets, showing higher accuracy than state-of-the-art approaches.

1. INTRODUCTION

The goal of Supervised Classification is to learn a model which automatically assigns samples to a set of predefined categories. Different approximations have been proposed in literature. For example, Support Vector Machines (SVMs) [1, 2] find the boundary decision which maximizes the distance between support vectors, Bayesian approaches such as Relevance vector machine [3] or Gaussian Process Classification [4] attempt to learn the underlying probabilistic model.

The use of Bayesian modeling and inference provides huge benefits: prior distributions are used to introduce information on the adaptive parameters, and hyperparameters are learned from data using a consistent framework. Priors based in l_p-quasinorms, $p \leq 1$, enforce sparsity on the adaptive parameters. The use of sparse priors has already been reported for softmax classification problems, see [5] for the use of the l_1 prior, and [6, 7] for the use of quadratic prior. However, the use of l_p-quasinorms, $p < 1$, is of particular importance when only very few features are relevant to the target output of a large number of features. Current approaches utilizing l_p-regularization treat the logistic regression from a likelihood-based perspective, and employ a cross-validation procedure to estimate the required regularization parameters (see [8] for details). Here we propose a Bayesian modeling and inference approach to sparse softmax classification using l_p-priors with $p < 1$. For a given $\mathbf{x}$, the output vector $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}), \ldots, y_K(\mathbf{x})]^{\mathrm{T}}$ consists of the 1-of-K binary representation of its classification. We have

$$\mathrm{p}(\mathbf{y}(\mathbf{x})|\mathbf{W}, \mathbf{x}) = \prod_{k=1}^{K} \left(\frac{\exp(\mathbf{w}_k^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}))}{\sum_{i=1}^{K} \exp(\mathbf{w}_i^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}))} \right)^{y_k(\mathbf{x})} \quad (1)$$

where the function $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{H}$, maps the observed $\mathbf{x} \in \mathcal{X}$ into a higher dimensional feature space $\mathcal{H}$ of dimension M whose first component is 1 and $\mathbf{W}$ is a matrix whose column vectors are the so called adaptive vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$. The goal in softmax classification is to learn the adaptive matrix $\mathbf{W}$ from a set of samples $\mathbf{x}_i, i = 1, \ldots, N$ with known classification $\mathbf{y}(\mathbf{x}_i), i = 1, \ldots, N$.

Getting the ground-truth label of each sample is in general a costly task. Active Learning (AL) techniques provide an iterative alternative to minimize such cost (see [9] for a complete survey). These techniques train an initial classifier using a small dataset, then, based on an optimality criterion, iteratively select samples (without knowing their labels). These samples are then classified by an oracle and used to improve the initial classifier.

AL techniques depend on the model the classifier learns, and therefore each classifier has its own AL techniques. For SVM, relevant approaches are: the sampling approach discussed in [9], the binary- and multiclass-level uncertainty [10], and the entropy-query-by-bagging [11]. In [12] a Bayesian framework is proposed and differential entropy is used to select new samples. In [13] A Gaussian process is used to estimate the posterior distribution of the labels, and three AL methods are proposed: maximum variance (equivalent to differential entropy in [12]), minimum distance to decision boundary, and a combination of both minimum normalized distance.

The goal of this paper is twofold. Firstly, using a prior based on l_p-quasinorms, we formulate the softmax classification problem from a Bayesian viewpoint. All required algorithmic parameters are also included in the proposed Bayesian model, and are estimated along with the unknowns. Due to the intractability of the posterior distributions, we employ Variational Bayesian analysis to provide an approximation to the posterior distribution of the unknowns. A relationship between the prior model used and the independent Gaussian prior model is also provided. Secondly, we tackle AL by utilizing the posterior distribution of the classes.

The paper is organized as follows. In Section 2 we use Bayesian modeling to define probability distributions on the unknowns. Variational inference is used to develop a training algorithm and a classification rule in Section 3. A study on the relationship between the proposed classification model and the use of Gaussian independent prior models is presented in Section 4. AL techniques are proposed in Section 5. In Section 6, the proposed methods are applied to Image Segmentation on a synthetic example and a real dataset. Conclusions are presented in Section 7.

This work has been supported in part by the Comisión Nacional de Ciencia y Tecnología under contract TIN2010-15137, CEI BioTic at the University of Granada, and the Department of Energy grant DE-NA0000457.

2. BAYESIAN MODEL

To perform Bayesian inference we assume that we already have the $K-$dimensional classification vectors $\mathbf{y}_i = \mathbf{y}(\mathbf{x}_i)$ associated to the feature samples $\boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1, \ldots, N$. Then we can write

$$\mathrm{p}(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} \mathrm{p}(\mathbf{y}_i|\mathbf{W}) \tag{2}$$

where $\mathbf{Y}$ is a $N \times K$ matrix with i^{th} row $\mathbf{y}_i^{\mathrm{T}}$ whose components are $y_{ik}, k = 1, \ldots, K$, $\mathrm{p}(\mathbf{y}_i|\mathbf{W})$ has been defined in Eq. (1) and the set $\mathbf{X}$ containing all the used samples, has been omitted for simplicity.

To estimate $\mathbf{W}$ we use, for each of its columns, the prior distribution $\mathrm{p}(\mathbf{w}_k|\alpha_k)$ based on l_p-quasinorms

$$\mathrm{p}(\mathbf{w}_k|\alpha_k) \propto \alpha_k^{M/p} \exp\left[-\alpha_k \sum_{i=1}^{M} |w_{ki}|^p\right], \tag{3}$$

where $\alpha_k > 0$ and $0 < p \leq 1$, $\mathbf{w}_k = (w_{k1}, \ldots, w_{kM})^{\mathrm{T}}, k = 1, \ldots, K$. This type of prior has been shown to enforce sparsity in estimation problems like logistic regression (see [14] and [15] for a regularization point of view) and in areas like image restoration and compressive sensing (see, for instance [16]).

Then, given $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, we have

$$\mathrm{p}(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathrm{p}(\mathbf{w}_k|\alpha_k)\,. \tag{4}$$

Finally, we assume that each $\alpha_k, k = 1, \ldots, K$ has as hyperprior, $\mathrm{p}(\alpha_k)$, the Gamma distribution, $\mathrm{p}(\alpha_k) = \Gamma(\alpha_k|a^o_{\alpha_k}, b^o_{\alpha_k})$, where $b^o_{\alpha_k} > 0$ and $a^o_{\alpha_k} > 0$, and have the following global model

$$\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y}) = \mathrm{p}(\boldsymbol{\alpha})\mathrm{p}(\mathbf{W}|\boldsymbol{\alpha})\mathrm{p}(\mathbf{Y}|\mathbf{W}). \tag{5}$$

3. VARIATIONAL BAYESIAN INFERENCE

The Bayesian paradigm dictates that inference on $(\boldsymbol{\alpha}, \mathbf{W})$ should be based on $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})$. However, $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})$ cannot be found in closed form. Therefore, we apply variational methods to approximate this distribution by a distribution $\mathrm{q}(\boldsymbol{\alpha}, \mathbf{W})$. The variational criterion used to find $\mathrm{q}(\boldsymbol{\alpha}, \mathbf{W})$ is the minimization of the Kullback-Leibler (KL) divergence, given by

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{\alpha}, \mathbf{W})\|\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y})) &= \text{const} \\ &+ \int\int \mathrm{q}(\boldsymbol{\alpha}, \mathbf{W}) \log\left(\frac{\mathrm{q}(\boldsymbol{\alpha}, \mathbf{W})}{\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})}\right) d\boldsymbol{\alpha} d\mathbf{W}.\end{aligned} \tag{6}$$

Unfortunately, due to the form of the prior and the observation models defined in (4) and (2) respectively, the integral above cannot be calculated. To solve this problem we proceed to bound below the distribution $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ by a function which renders the calculation of $\mathrm{KL}(\mathrm{q}(\boldsymbol{\alpha}, \mathbf{W}) \parallel \mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}|\mathbf{Y}))$ possible when $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ is replaced by such a function. A lower bound on $\mathrm{p}(\mathbf{w}_k|\alpha_k), k = 1, \ldots, K$ is found by using the following inequality (see [17], and [18] based on [19])

$$a^{\frac{p}{2}} \leq \frac{p}{2}\,\frac{a + \frac{2-p}{p} b}{b^{1-p/2}}\,, \tag{7}$$

for $a \geq 0$, $b > 0$, and $0 \leq p \leq 2$, which applied to the energy of the prior produces

$$\alpha_k \sum_{i=1}^{M} |w_{ki}|^p \leq \frac{1}{2}\alpha_k p \sum_{i=1}^{M} \frac{w_{ki}^2 + \frac{2-p}{p}\theta_{ki}}{\theta_{ki}^{1-p/2}}\,, \tag{8}$$

where $\theta_i > 0$. Consequently, for the prior in Eq. (3) we have

$$\begin{aligned}\mathrm{p}(\mathbf{w}_k|\alpha_k) &\geq \mathbf{M}(\alpha_k, \mathbf{w}_k, \boldsymbol{\theta}_k) = \\ &= \text{const} \times \alpha_k^{M/p} \exp\left(-\frac{1}{2}\alpha_k p \sum_{i=1}^{M} \frac{w_{ki}^2 + \frac{2-p}{p}\theta_{ki}}{\theta_{ki}^{1-p/2}}\right),\end{aligned} \tag{9}$$

where $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kM})^{\mathrm{T}}$, and we can write

$$\mathrm{p}(\mathbf{W}|\boldsymbol{\alpha}) \geq \prod_{k=1}^{K} \mathbf{M}(\alpha_k, \mathbf{w}_k, \boldsymbol{\theta}_k) = \mathbf{M}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta}). \tag{10}$$

where $\boldsymbol{\Theta}$ is a matrix with column vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. In order to obtain a lower bound on $\mathrm{p}(\mathbf{Y}|\mathbf{W})$ we follow [6] and notice that for any $\mathbf{u} \in \mathbb{R}^K$ and $\beta \in \mathbb{R}$ we have

$$\begin{aligned}\ln \sum_{k=1}^{K} e^{u_k} &\leq \beta + \sum_{k=1}^{K} \frac{u_k - \beta - \xi_k}{2} \\ &+ \sum_{k=1}^{K} (\lambda(\xi_k)((u_k - \beta)^2 - \xi_k^2) + \ln(1 + e^{\xi_k}))\end{aligned} \tag{11}$$

for all $\xi_k \in \mathbb{R}_0^+$ with $\lambda(\xi_k) = \frac{1}{2\xi_k}\left(\frac{1}{1+e^{-\xi_k}} - \frac{1}{2}\right)$. Applying (11) to Eq. (2) we obtain

$$\begin{aligned}\ln \mathrm{p}(\mathbf{Y}|\mathbf{W}) &= \sum_{i=1}^{N} \ln \mathrm{p}(\mathbf{y}_i|\mathbf{W}) \geq \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K} (\frac{\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i) - \beta_i - \xi_{ik}}{2} + \ln(1 + e^{\xi_{ik}})) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K} \lambda(\xi_{ik})((\mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i) - \beta_i)^2 - \xi_{ik}^2) \\ &- \sum_{i=1}^{N} \beta_i = \ln \mathbf{H}(\mathbf{W}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y}),\end{aligned} \tag{12}$$

where $\boldsymbol{\Xi}$ is a matrix with row vectors $\boldsymbol{\xi}_i^{\mathrm{T}}, i = 1 \ldots N$, each of these vectors has the form $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^{\mathrm{T}}$.

Notice that in [6] the same parameter β is used for all the samples.

Using the lower bounds in (10) and (12), the joint distribution is bounded below by

$$\begin{aligned}\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y}) \geq &\mathrm{p}(\boldsymbol{\alpha})\mathbf{M}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta})\mathbf{H}(\mathbf{W}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y}) \\ = &\mathbf{F}(\boldsymbol{\alpha}, \mathbf{W}, \boldsymbol{\Theta}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y})\,.\end{aligned} \tag{13}$$

We replace $\mathrm{p}(\boldsymbol{\alpha}, \mathbf{W}, \mathbf{Y})$ by this lower bound in (6) and use the factorization $\mathrm{q}(\boldsymbol{\alpha}, \mathbf{W}) = \mathrm{q}(\boldsymbol{\alpha})\mathrm{q}(\mathbf{W})$.

Then the posterior distribution $\mathrm{q}(\mathbf{w}_k), k = 1, \ldots, K$ is the multivariate normal distribution $\mathcal{N}(<\mathbf{w}_k>, \Sigma_{\mathbf{w}_k})$ where

$$\Sigma_{\mathbf{w}_k}^{-1} = \Lambda_k + 2\sum_{i=1}^{N} \lambda(\xi_{ik})\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x}_i), \tag{14}$$

$$<\mathbf{w}_k> = \Sigma_{\mathbf{w}_k} \sum_{i=1}^{N} ((y_{ik} - \frac{1}{2})\boldsymbol{\phi}(\mathbf{x}_i) + 2\beta_i\lambda(\xi_{ik})\boldsymbol{\phi}(\mathbf{x}_i))$$

with $\Lambda_k = \mathrm{diag}\left(<\alpha_k> p\theta_{ki}^{p/2-1}\right), i = 1, \ldots, M$.

Furthermore we have

$$\theta_{ki} = <w_{ki}^2> = (\Sigma_{\mathbf{w}_k})_{ii} + (<w_{ki}>)^2\,. \tag{15}$$

Furthermore $\mathrm{q}(\alpha_k) = \Gamma(\alpha_k | a^o_{\alpha_k} + \frac{M}{p}, b^o_{\alpha_k} + \sum_{i=1}^M \theta_{ki}^{p/2})$ with mean

$$<\alpha_k> = \frac{1}{p} \frac{a^o_{\alpha_k}\, p + M}{b^o_{\alpha_k} + \sum_{i=1}^M (\theta_{ki})^{p/2}}. \quad (16)$$

Finally we have

$$\xi_{ik} = \sqrt{\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x}_i)\Sigma_{\mathbf{w}_k}\boldsymbol{\phi}(\mathbf{x}_i) + (<\mathbf{w}_k>^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i) - \beta_i)^2}, \quad (17)$$

and

$$\beta_i = \frac{K - 2 + 4\sum_{k=1}^K \lambda(\xi_{ik}) <\mathbf{w}_k>^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i)}{4\sum_{k=1}^K \lambda(\xi_{ik})}. \quad (18)$$

Notice that the uncertainty of the estimate of $\mathbf{w}_k$ is incorporated into the estimation procedure of the other unknowns by the use of the covariance matrix $\Sigma_{\mathbf{w}_k}$ in (15), (16) and (17).

The above inference leads to a learning procedure which is summarized in Algorithm 1. At convergence this algorithm estimates all the parameters, including the distribution of the adaptive vectors $\mathbf{w}_k$. The point estimates of the adaptive vectors are $\langle \mathbf{w}_k \rangle$ in Eq. (14). Given a new sample $\mathbf{x}^*$, we utilize as predictive distribution of the classes

$$\mathrm{p}(\mathcal{C}_k|\mathbf{x}^*) = \frac{\exp(\langle \mathbf{w}_k \rangle^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^*))}{\sum_{i=1}^K \exp(\langle \mathbf{w}_i \rangle^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^*))} \quad (19)$$

and assign $\mathbf{x}^*$ to the class with maximum probability.

Algorithm 1 Learning Procedure

Require: $\boldsymbol{\alpha}^0 = (1, \ldots, 1)^{\mathrm{T}}$, $\theta^0_{ki} = 1$, $\xi^0_{ik} = 1$ and $\beta_i = 1$.

1: **repeat**
2: Calculate $\mathrm{q}(\mathbf{W})^{n+1}$ using Eq. (14).
3: Calculate $\mathrm{q}(\boldsymbol{\alpha})^{n+1}$ using Eq. (16).
4: Parameters θ^{n+1}_{ki}, ξ^{n+1}_{ik}, and β^{n+1}_i are updated using Eq. (15), Eq. (17) and Eq. (18) respectively.
5: **until** convergence

4. RELATION TO INDEPENDENT GAUSSIAN PRIOR MODEL

Let us study here the relationship between the proposed classification model and the use of Gaussian independent prior models on the components of $\mathbf{w}_k, k = 1, \ldots, K$. Let us assume that

$$\mathrm{p_G}(\mathbf{w}_k|\boldsymbol{\upsilon}_k) \propto \prod_{i=1}^M \upsilon_{ki}^{1/2} \exp\left[-\frac{1}{2}\upsilon_{ki} w_{ki}^2\right], \quad (20)$$

$$\mathrm{p}(\boldsymbol{\upsilon}_k) = \prod_{i=1}^M \mathrm{p}(\upsilon_{ki}) = \prod_{i=1}^M \Gamma(\upsilon_{ki}|a^o_{\alpha_k}, b^o_{\alpha_k}), \quad (21)$$

where $\boldsymbol{\upsilon}_k = (\upsilon_{k1}, \ldots, \upsilon_{kM})^{\mathrm{T}}, k = 1, \ldots, K$ and the parameters $a^o_{\alpha_k}, b^o_{\alpha_k}$ are the ones defined for the l_p-quasinorms.

Utilizing the same observation bound in (12), we obtain

$$\begin{aligned}\mathrm{p_G}(\boldsymbol{\Upsilon}, \mathbf{W}, \mathbf{Y}) &= \mathrm{p}(\mathbf{Y}|\mathbf{W}) \prod_{k=1}^K \mathrm{p}(\boldsymbol{\upsilon}_k)\mathrm{p_G}(\mathbf{w}_k|\boldsymbol{\upsilon}_k) \\ &\geq \mathrm{H}(\mathbf{W}, \boldsymbol{\Xi}, \boldsymbol{\beta}, \mathbf{Y}) \prod_{k=1}^K \mathrm{p}(\boldsymbol{\upsilon}_k)\mathrm{p_G}(\mathbf{w}_k|\boldsymbol{\upsilon}_k). \quad (22)\end{aligned}$$

where $\boldsymbol{\Upsilon}$ is a matrix with row vectors $\boldsymbol{\upsilon}_k^{\mathrm{T}}, k = 1 \ldots K$, each of these vectors has the form $\boldsymbol{\upsilon}_k = (\upsilon_{k1}, \ldots, \upsilon_{kM})^{\mathrm{T}}$

Utilizing $\mathrm{q_G}(\mathbf{W}) = \prod_{k=1}^K \mathrm{q_G}(\mathbf{w}_k)$, the variational posterior distribution $\mathrm{q_G}(\mathbf{w}_k)$ is $\mathcal{N}(<\mathbf{w}_k>_{\mathrm{G}}, \Sigma_{\mathbf{w}_k,\mathrm{G}})$ with parameters

$$(\Sigma_{\mathbf{w}_k,\mathrm{G}})^{-1} = \Lambda_{k,\mathrm{G}} + 2\sum_{i=1}^N \lambda(\xi_{ik})\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x}_i), \quad (23)$$

$$<\mathbf{w}_k>_{\mathrm{G}} = \Sigma_{\mathbf{w}_k,\mathrm{G}} \sum_{i=1}^N ((y_{ik} - \frac{1}{2})\boldsymbol{\phi}(\mathbf{x}_i) + 2\beta_i\lambda(\xi_{ik})\boldsymbol{\phi}(\mathbf{x}_i),$$

$$\Lambda_{k,\mathrm{G}} = \mathrm{diag}\,(<\upsilon_{ki}>)\,. \quad (24)$$

The mean of the posterior distribution approximation of υ_{ki} is

$$<\upsilon_{ki}> = \frac{a^o_{\alpha_k} + \frac{1}{2}}{b^o_{\alpha_k} + \frac{<w^2_{ki}>}{2}}\,. \quad (25)$$

Let us assume that $a^o_{\alpha_k} = b^o_{\alpha_k} = 0$ and rewrite (14) making explicit its dependency on p. Utilizing (16) we have

$$\Lambda_{k,p} = \mathrm{diag}\left(\frac{a^o_{\alpha_k}\, p + M}{b^o_{\alpha_k} + \sum_{i=1}^M \theta_{ki}^{p/2}} \theta_{ki}^{p/2-1}\right) \quad (26)$$

Taking the limit $p \to 0$ and using (15), we obtain

$$\lim_{p\to 0} \Lambda_{kp} = \mathrm{diag}\left(\theta_{ki}^{-1}\right) = \mathrm{diag}\left(<w^2_{ki}>^{-1}\right)\,. \quad (27)$$

Let us now examine the Gaussian model. When $a^o_{\alpha_k} = b^o_{\alpha_k} = 0$, we have from (24) and (25)

$$\Lambda_{k,\mathrm{G}} = \mathrm{diag}\,(<\upsilon_{ki}>) = \mathrm{diag}\left(<w^2_{ki}>^{-1}\right)\,. \quad (28)$$

Consequently, when the starting distributions of the variational algorithms are the same we have $\lim_{p\to 0} \Lambda_{k,p} = \Lambda_{k,\mathrm{G}}$. Therefore, in the limiting case $p \to 0$, the posterior distributions associated with the l_p-prior and the independent Gaussian priors for each component of $\mathbf{w}_k$ coincide.

5. INCREMENTAL AND ACTIVE LEARNING

Let us now assume that we want to add a new observation $\mathbf{x}_{N+1}$ to the training set, whose corresponding $\mathbf{y}(\mathbf{x}_{N+1})$ will be provided by an oracle. To select $\mathbf{x}_{N+1}$ we propose two active learning methods which are based on the posterior probabilities of the classes.

In the first method, called *Minimum Probability Criteria*, we select the next sample to be used to improve the classifier as

$$\mathbf{x}_{N+1} = \arg\min_{\mathbf{x}^*}(\max_k(\mathrm{p}(\mathcal{C}_k|\mathbf{x}^*))). \quad (29)$$

In the second method, named *Maximum Entropy Criteria*, we select the sample whose posterior distribution of the classes is less informative. Formally

$$\mathbf{x}_{N+1} = \arg\max_{\mathbf{x}^*} -\sum_{k=1}^K \mathrm{p}(\mathcal{C}_k|\mathbf{x}^*)\ln \mathrm{p}(\mathcal{C}_k|\mathbf{x}^*). \quad (30)$$

6. EXPERIMENTAL RESULTS

Due to space limitations, in this section we provide a limited number of experiments to analyze the performance of the proposed model for classification and AL.

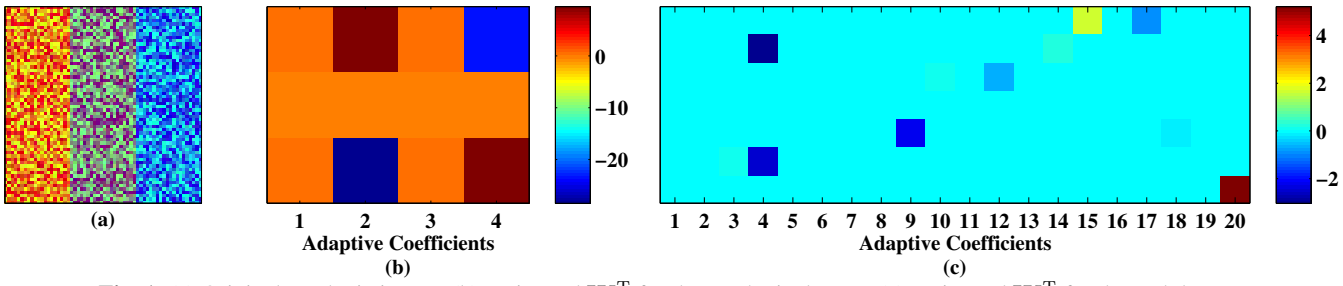

Fig. 1. (a) Original synthetic image. (b) Estimated $\mathbf{W}^{\mathrm{T}}$ for the synthetic dataset. (c) Estimated $\mathbf{W}^{\mathrm{T}}$ for the real dataset.

6.1. Supervised Classification results

Figure 1(a) shows a synthetically generated 60×60 image. The goal is to segment the three vertical rectangles in the image. Each rectangle represents one class in our segmentation problem. The pixels in each class are drawn from Gaussian distributions with mean vectors $\mu_1 = (0.9, 0.5, 0.1)^{\mathrm{T}}$, $\mu_2 = (0.5, 0.5, 0.5)^{\mathrm{T}}$ and $\mu_3 = (0.1, 0.5, 0.9)^{\mathrm{T}}$, respectively. The three components of each pixel are normalized RGB values, each component is corrupted with noise of standard deviations 0.05, 0.5 and 0.05 respectively. Notice that the G band does not provide information to the classifier.

The experiment is repeated 10 times with 10 different training sets, each with 12 samples (4 from each class). As accuracy measure, the Cohen's Kappa statistic (κ-index) is calculated on a test set of 1500 samples (500 from each class).

The values $p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ were tested. For values $p = 0.1, \ldots, 0.6$ the obtained κ-index was 1, and 0.99 for the other values. Therefore, for $p = 0.1, \ldots, 0.6$, the proposed method segments the synthetic image correctly on the test set. Fig. 1(b) shows the coefficients of the estimated adaptive matrix $\mathbf{W}^{\mathrm{T}}$ for $p = 0.1$. Non zero entries represent components relevant to classification. The proposed model does not use the G band and assigns zero to the corresponding adaptive coefficients.

We compare the proposed method with an SVM classifier. To perform a fair comparison we use a Gaussian kernel whose parameter is manually tunned to obtain the best performance. The SVM cost parameter is estimated using cross-validation. The obtained mean κ-index was 0.95, and therefore, the SVM classifier does not segment correctly the whole test sets from the synthetic image.

In our second classification experiment we evaluate the proposed Bayesian classifier on the real data set "Image Segmentation", available on-line at the "UCI Machine Learning Repository" [20]. The goal is to classify a set of pixels in 7 classes: "BRICKFACE", "SKY", "FOLIAGE", "CEMENT", "WINDOW", "PATH" and "GRASS". The data set has 2310 samples (330 from each class). Each sample is a 19 component vector representing different attributes measured on a 3×3 neighborhood of the pixel of interest.

The experiment is repeated 10 times on 10 different training sets, each with 126 samples (18 from each class). The κ-index is calculated on a test set with 1050 samples (150 from each class). For $p = 1$, the obtained κ-index was 0.86. The best κ-index, 0.88, was obtained at $p = 0.02$, this implies that l_p-quasinorms with $p < 1$ can outperform the l_1-norm.

Fig. 1(c) shows the absolute value of the estimated adaptive coefficients in $\mathbf{W}^T$. Components 9, 12, 14, 15, 17, 18, 20 correspond to attributes "'horizontal edge mean", "rawred-mean", "rawgreen-mean", "excess red", "excess green", "value-mean" and "hue-mean", respectively. Attributes like "row" or "column", which correspond to pixel position in the image, have no discriminative information. In those components, the estimated values of $\mathbf{W}$ were 0 (second and third columns in the figure). The fourth component is "number of pixel where attributes were measured", this component is equal to 9 for all samples, consequently the fourth component acts as the bias for each class while the first component, which was introduced for this purpose, takes the value zero. Interestingly, and as expected, if we remove the fourth component, the estimated values of the first components are the values of the fourth components multiplied by 9. Notice that because of the prior used, the classifier prefers to make zero the first component and assign small values to the adaptive coefficients of the fourth feature.

Finally we compare again the proposed method with an SVM classifier. Its mean κ-index was 0.84. Its performance is 0.02 and 0.04 lower than the proposed classifier for $p = 1$ and $p = 0.02$, respectively. Additionally we note that our proposed method does not need parameter tunning.

6.2. Active Learning results

To evaluate the performance of the proposed AL methods, we utilize learning curves. We start by training the classifier using Algorithm 1 on a reduced subset from the training set. The estimated adaptive matrix $\mathbf{W}$ is then used to classify the test set, the κ-index is utilized as accuracy measure in the learning curves. Next, the AL methods proposed in Section 5 are used to select a new sample from the training set and the classifier updated.

The proposed AL methods in Sections 5 are noted MIN PRO (minimum probability) and MAX ENTRO (maximum entropy). They are compared to the following AL methods: margin sampling (SVM-MS) [9], entropy-query-by-bagging (SVM-EQB) [11] and multiclass-level uncertainty (SVM-MCLU) [10]. All of them use SVM as classifier. The cost parameter is estimated by cross-validation.

For the synthetic dataset, the experiment is repeated 10 times with 10 different initial training sets. The starting training set has 6 samples (2 from each class) and the whole training and test sets have 1500 samples (500 from each class). We use $p = 0.1$.

Figure 2(a) shows the mean κ-index learning curves. The proposed methods start at κ-index=0.91. Their learning rates are very fast, reaching κ-index=1 after adding only 2 samples to the inital training set. Both methods have the same behavior and perform better than randomly selecting the new samples from the training set and using the proposed classifier. The random approach does not reach κ-index=1 even after 20 samples have been added. Methods that use a SVM classifier start at κ-index=0.78, so they initially perform worse than our classification method. SVM-MCLU needs 5 to reach κ-index= 1. SVM-EQB obtains a κ-index=1 when 11 samples have been added. Furthermore SVM-MS does not achieve κ-index =1 even when 20 samples have been added.

For the real dataset we use a test set with 1050 samples (150 from each class), the whole training set also contains 1050 samples

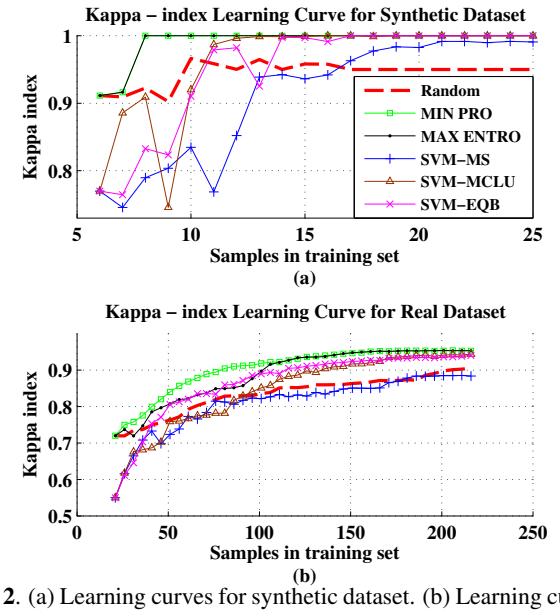

Fig. 2. (a) Learning curves for synthetic dataset. (b) Learning curves for real dataset.

(150 from each class). 10 initial training sets with 21 samples (3 from each class), are used. We use $p = 0.02$.

Figure 2(b) depicts the mean κ-index. The proposed methods start at 0.72 and reach κ-index=0.98 when the training set has 150 samples. After that the corresponding learning curves become flat. In this experiment MIN PRO outperforms MAX ENTRO, in particular notice the difference between both methods when we have less 100 samples. Both methods outperform random sampling which reaches κ-index=0.9 when 200 samples have been added.

The SVM classifiers utilize a Gaussian kernel whose parameters are manually tunned to obtain the performance. They start almost 0.15 below the proposed methods. SVM-MS does not perform well and its learning curve is similar to random sampling. SVM-MCLU and SVM-EQB performs similarly when 150 samples have been added and reach κ-index = 0.96. However SVM-EQB is better than SVM-MCLU for less than 150 samples. None of these methods outperformed the proposed ones.

7. CONCLUSIONS

In this work Bayesian modeling and inference have been used to address Supervised Classification and AL problems. The l_p-prior models utilized on the adaptive coefficients have promoted sparsity on the estimated adaptive parameters. Variational inference has been used to estimate all the model parameters and connections with independent Gaussian priors established. The predictive distribution of the classes has been calculated. This distribution has been used to define two AL methods. In the experimental section the proposed approach has been applied to Image Segmentation problems. Experimental results have shown that the use of l_p-priors allows the classifier to select discriminative features and discard non-relevance components. The proposed approach has shown higher accuracy than SVM methods in both classification and AL problems.

REFERENCES

[1] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambrigde, MA, 2002.

[2] C.M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.

[3] M. E. Tipping, "The relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[4] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, MIT Press, NY, 2006.

[5] G. C. Cawley, N. L. C. Talbot, and M. Girolami, "Sparse multinomial logistic regression via bayesian l1 regularisation," in *Neural Information Processing Systems*, 2006, pp. 209–216.

[6] G. Bouchard, "Efficient bounds for the softmax function and applications to approximate inference in hybrid models," in *NIPS 2007*, 2007.

[7] N. Ahmed and M. Campbell, "Variational bayesian learning of probabilistic discriminative models with latent softmax variables," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 7, pp. 3143–3154, July 2011.

[8] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fmri data," *NeuroImage*, vol. 51, no. 2, pp. 752–764, 2010.

[9] B. Settles, *Active Learning*, Morgan & Claypool, 2012.

[10] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, March 2011.

[11] D. Tuia, F. Ratle, F. Pacifici, M.F. Kanevski, and W.J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 –2232, July 2009.

[12] J. Paisley, X. Liao, and L. Carin, "Active learning and basis selection for kernel-based linear models: A Bayesian perspective," *IEEE Trans. on Sig. Proc.*, vol. 58, pp. 2686–2700, 2010.

[13] P. Ruiz, J. Mateos, G. Camps-Valls, R. Molina, and A.K. Katsaggelos, "Bayesian active remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2186–2196, April 2014.

[14] A. Kabán and R.J. Durrant, "Learning with $l_{q<1}$ vs l_1-norm regularization with exponentially many irrelevant features," in *Proc. of ECML PKDD*. 2008, pp. 580–596, Springer-Verlag.

[15] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S.J. Meltzer, and M. Tan, "Sparse logistic regression with L_p penalty for biomarker identification," *Statistical Applications in Genetics and Molecular Biology*, 2007.

[16] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. on Image Processing*, vol. 19, no. 2, pp. 53–63, Jan. 2010.

[17] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Parameter estimation in TV image restoration using variational distribution approximation," *IEEE Trans. on Image Processing*, vol. 17, no. 3, pp. 326–339, March 2008.

[18] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," in *NIPS 2006*.

[19] R. T. Rockafellar, *Convex Analysis (Princeton Mathematical Series)*, Princeton University Press, 1970.

[20] K. Bache and M. Lichman, "UCI machine learning repository," 2013.