Combining Generative and Discriminative Methods for Pixel Classification with Multi-Conditional Learning

B. Michael Kelm
Interdisciplinary Center for Scientific Computing
University of Heidelberg
michael.kelm@iwr.uni-heidelberg.de

Chris Pal   Andrew McCallum
Department of Computer Science
University of Massachusetts Amherst
{pal, mccallum}@cs.umass.edu

Abstract

It is possible to broadly characterize two approaches to probabilistic modeling in terms of generative and discriminative methods. Provided with sufficient training data the discriminative approach is expected to yield superior accuracy as compared to the analogous generative model since no modeling power is expended on the marginal distribution of the features. Conversely, if the model is accurate the generative approach can perform better with less data. In general it is less vulnerable to overfitting and allows one to more easily specify meaningful priors on the model parameters. We investigate multi-conditional learning – a method combining the merits of both approaches. Through specifying a joint distribution over classes and features we derive a family of models with analogous parameters. Parameter estimates are found by optimizing an objective function consisting of a weighted combination of conditional log-likelihoods. Systematic experiments in the context of foreground/background pixel classification with the Microsoft-Berkeley segmentation database using mixtures of factor analyzers illustrate tradeoffs between classifier complexity, the amount of training data and generalization accuracy. We show experimentally that this approach can lead to models with better generalization performance than purely generative or discriminative approaches.

1 Introduction

There are a wide and growing variety of tasks in Computer Vision for which Machine Learning methods based on probability are being successfully applied. For classification tasks it is particularly common to make the distinction between the generative approach and the discriminative approach to probabilistic modeling. In contrast, in the approach we present here neither a purely generative nor a completely discriminative approach is used. Rather, first a joint model with latent variables is constructed which is then optimized with respect to multiple conditional likelihoods. We call this approach *multi-conditional learning*. In other work, multi-conditional learning was introduced [6] in the context of random field models for documents.

Here, we apply multi-conditional learning to a mixture of factor analyzers (MFA) model [3] which is a powerful latent variable model that allows one to perform simultaneous dimensionality reduction and clustering. As opposed to [6] where an undirected graphical model is constructed we obtain the joint model in a generative fashion here. The multi-conditional MFA is then applied to the color image segmentation problem described in [1]. The latter work has compared generative and discriminative methods for parameter estimation within an underlying (spatially coupled) Gaussian Mixture Markov Random Field model.

Recently, attention has been given to comparisons between discriminative and generative methods for *modeling* in the context of object recognition problems [10]. In [5] the Conditional Expectation Maximization CEM algorithm is proposed for parameter estimation in generative models with hidden variables based on optimizing the marginal conditional likelihood of classes. This work also illustrates the distinction between the optimization of a generative model under standard joint likelihood and for a particular conditional likelihood obtained from the underlying joint model. The work of [5] can thus be characterized as a type of discriminative learning in generative models.

Many classical statistical models can be viewed from these generative and discriminative perspectives. For example, while the name is misleading, classical linear discriminant analysis (LDA) arises from the posterior probability of a generative model consisting of a class prior $P(c)$ and class-conditional Gaussian distributions $P(x \,|\, c)$ with common covariance matrix. The parameters of the generative model are determined from the traditional maximum (joint) likelihood estimate. In contrast, a model with the same parametric form also arises in linear logistic regression which represents a discriminative $P(c \,|\, x)$ model directly. However, the parameters of the logistic regression model are determined by optimizing the conditional likelihood and without constructing an explicit model of $P(x)$. In [8] classifiers with these relationships are characterized as *Generative-Discriminative pairs*.

2 Multi-Conditional Learning

In classification problems one is interested in predicting a class c given observed features x. From decision theory it is known that the most complete characterization of the solution is given by the conditional class probabilities $\mathrm{P}(c \,|\, x, \theta)$. All models considered in the following share the same parametric form for the conditional class probabilities and only differ in the way the parameters θ are estimated.

The *generative* approach attempts to capture the manner in which observed features x are generated from given classes c by specifying a prior distribution over classes and a class-conditional distribution over the features. It therefore defines the joint distribution $\mathrm{P}(x, c) = \mathrm{P}(x \,|\, c)\,\mathrm{P}(c)$. The posterior is obtained from Bayes‘ formula as

$$\mathrm{P}(c \,|\, x, \kappa, \lambda) \;=\; \frac{\mathrm{P}(x \,|\, c, \kappa)\,\mathrm{P}(c \,|\, \lambda)}{\sum_c \mathrm{P}(x \,|\, c, \kappa)\,\mathrm{P}(c \,|\, \lambda)} \qquad (1)$$

with the two parameter vectors κ and λ. For parameter estimation the maximum likelihood principle leads to the (negative) joint loglikelihood (JL)

$$\mathcal{L}_{c,x}(\theta; \mathcal{D}) \;=\; \sum_{i=1}^{N} \log \mathrm{P}(c_i, x_i \,|\, \theta), \qquad (2)$$

where $\mathcal{D}$ denotes an iid training data set and $\theta = (\kappa, \lambda)$.

The *discriminative* approach on the other hand, directly captures $\mathrm{P}(c \,|\, x, \theta)$ but does not demand a model for the features x. In fact, any distribution $\mathrm{P}(x \,|\, \nu)$ could be assumed thus defining the joint distribution as

$$\mathrm{P}(x, c \,|\, \theta, \nu) \;=\; \mathrm{P}(c \,|\, x, \theta)\,\mathrm{P}(x \,|\, \nu) \qquad (3)$$

Since only $\mathrm{P}(c \,|\, x, \theta)$ is needed for classification the parameters ν do not have to be determined [7]. The resulting loglikelihood is therefore the conditional loglikelihood (CL)

$$\mathcal{L}_{c|x}(\theta; \mathcal{D}) \;=\; \sum_{i=1}^{N} \log \mathrm{P}(c_i \,|\, x_i, \theta) \qquad (4)$$

Parameter estimation by maximizing $\mathcal{L}_{c|x}$ w.r.t. θ is commonly referred to as *discriminative training*.

Instead of modeling $\mathrm{P}(c \,|\, x, \theta)$ directly the posterior derived in a generative way (1) can also be used for discriminative training (4). Given the same number of parameters θ discriminative training is expected to yield a more powerful model since the conditional class distribution $\mathrm{P}(c \,|\, x)$ is usually simpler than the joint distribution over c and x. However, this model is also more susceptible to overfitting, in particular if only little training data is available. To prevent overfitting prior knowledge has to be used. As prior knowledge usually results in biased estimates it is important to define correct priors. This is often easier for generative models when the model parameters are associated with some meaning and prior knowledge about these parameters is indeed available.

In this paper we propose and examine the use of *multi-conditional* models which are derived from a joint distribution by using the *multi-conditional loglikelihood* (MCL)

$$\begin{aligned}\mathcal{L}^{\alpha}_{c|x,x|c}(\theta; \mathcal{D}) \;=\; \sum_{i=1}^{N} \big[& \log \mathrm{P}(c_i \,|\, x_i, \theta) \\ & + \alpha \log \mathrm{P}(x_i \,|\, c_i, \theta) \big] \end{aligned} \qquad (5)$$

MCL is defined with a temperature parameter α. For $\alpha = 0$ MCL just turns into CL whereas for $\alpha = 1$ a pseudo-likelihood is obtained. It is well-known that pseudo-likelihood is asymptotically consistent, *i.e.* in the infinite data limit it yields the same parameter estimates as JL. By choosing α between 0 and 1 one can smoothly vary between JL and CL criteria and thus one defines a whole *family of models* (as opposed to model “pairs” [8]). MCL combines advantages from both generative and discriminative approaches since

- the second term in Eqn. (5) defines a consistent regularizer for the parameters θ since it is derived from a joint distribution.
- in our work here the model is constructed in a generative way allowing one to incorporate correct prior knowledge such as certain invariances or prior distributions over parameters.
- for small α MCL concentrates on the discriminative part of the distribution to improve classification results over joint likelihood training.

Finally, all objectives are easily calculated given the joint loglikelihood $\mathcal{L}_{c,x}$ and the two marginal loglikelihoods $\mathcal{L}_x$ and $\mathcal{L}_c$, as

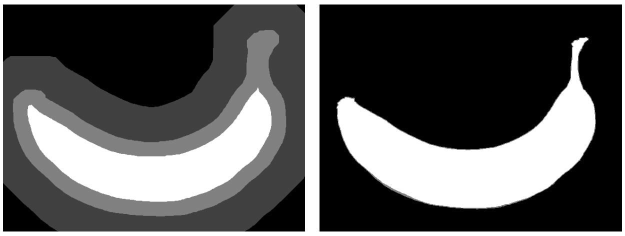
$$\mathcal{L}_{c|x} \;=\; \mathcal{L}_{c,x} - \mathcal{L}_x \qquad (6)$$

$$\mathcal{L}^{\alpha}_{c|x,x|c} \;=\; (1+\alpha)\mathcal{L}_{c,x} - \mathcal{L}_x - \alpha\mathcal{L}_c. \qquad (7)$$

3 MFA Models for Pixel Classification

In [1] a segmentation database was constructed[1] for which 30 color images were combined with 20 images from

[1] http://www.research.microsoft.com/vision/cambridge/segmentation/



(a) lasso labeling (b) ground truth

Figure 1. A users lasso labeling of the boundary of a banana and the corresponding ground truth fore- and background labeling.

the Berkeley segmentation database[2] An example from the database is shown in Figure 1. The users' labels are specified by a tri-map obtained with a lasso or pen tool as shown in Figure 1(a). Given background (dark gray) and foreground (white) pixels the task is to classify every pixel in the inference region (light gray) as fore- or background pixel. For each image ground truth labels are available (*c.f.* Figure 1(b)) and are used to determine the classification test accuracies in the inference region.

For every color pixel a 9-dimensional feature vector is constructed by concatenating three color values (CIE Lab) and six texture values. The latter are obtained from the three times two eigenvalues of the structure tensor [4] in each of the Lab planes.

Multiple colors and textures can appear in the fore- and background regions. Therefore, a mixture model is required to describe the feature distribution [1]. Furthermore, since the 9-dimensional feature vector certainly carries redundant information dimensionality reduction should be applied. Both can be achieved simultaneously with a mixture of factor analyzers (MFA) model [3], a latent variable model. Hence, the MFA model served as a basis for the derived family of models in the following.

The generative model for pixel classification is depicted in Figure 2 and defined by the following distributions

$$\begin{aligned} \mathrm{P}(c \,|\, \pi) &= \exp[\pi^T c] && (8) \\ \mathrm{P}(s \,|\, c, \Omega) &= \exp[c^T \Omega\, s] && (9) \\ \mathrm{P}(x \,|\, s, \mu_s, \Lambda_s, \Psi) &= \mathcal{N}_d(\mu_s, \Lambda_s \Lambda_s^T + \Psi) && (10) \end{aligned}$$

where s and c are multinomial variables, π a vector and Ω a table of log-probabilities. Ω is constrained such that an equal number K of subclasses s are associated with each of the C classes. The normalization constraints on π and Ω have been enforced by a suitable parameterization (softmax). $\mathcal{N}_d(\mu_s, \Sigma_s)$ is a d-dimensional Gaussian with constrained covariance matrix Σ_s. Each subclass has its own mean μ_s and a $d \times p$ factor loading matrix Λ_s ($p \ll d$). The additive diagonal covariance matrix Ψ is common to all subclasses. Contrasting [3] we do not introduce additional Gaussian latent variables z. The dimensionality reduction is expressed in the non-square loading matrices Λ_s and manifests itself as the factorized covariance matrix in Eqn. (10).

From the joint distribution defined by Eqns. (8)-(10), models based on JL, CL and MCL have been derived. Due to its good convergence properties an expectation-gradient algorithm [9] has been used to find parameter estimates. Therefore, expressions for the loglikelihoods $\mathcal{L}_c$, $\mathcal{L}_x$ and $\mathcal{L}_{c,x}$ and their gradients have been derived:

$$\begin{aligned} \mathcal{L}_c &= \sum_{i=1}^{N} \log \mathrm{P}(c_i) && (11) \\ \mathcal{L}_x &= \sum_{i=1}^{N} \log \left[\sum_c \mathrm{P}(c) \sum_s \mathrm{P}(s \,|\, c)\, \mathrm{P}(x_i \,|\, s) \right] && (12) \\ \mathcal{L}_{c,x} &= \sum_{i=1}^{N} \log \sum_s \mathrm{P}(s)\, \mathrm{P}(c_i, x_i \,|\, s) \\ &= \sum_{i=1}^{N} \left[\langle \log \mathrm{P}(c_i, x_i, s) \rangle + \mathrm{H}(s \,|\, c_i, x_i) \right] && (13) \end{aligned}$$

with the conditional expectation $\langle \cdot \rangle \equiv \mathrm{E}_{\mathrm{P}(s \,|\, c_i, x_i)}[\cdot]$ and the entropy $\mathrm{H}(\cdot)$.

The gradients for the marginal loglikelihoods (11)-(13) are easily determined given the gradient of the multinomial loglikelihood function $g(\pi) = \log \exp(\pi^T c)$

$$\frac{\partial}{\partial \pi} g = c \qquad (14)$$

and the gradient of the Gaussian loglikelihood function $f(\mu_s, \Lambda_s, \Psi) = \log \mathcal{N}_d(\mu_s, \Sigma_s = \Lambda_s \Lambda_s^T + \Psi)$

$$\begin{aligned} \frac{\partial}{\partial \mu_s} f &= \Sigma_s^{-1}(x - \mu_s) && (15) \\ \frac{\partial}{\partial \Lambda_s} f &= \Sigma_s^{-1}(S - \Sigma_s)\Sigma_s^{-1}\Lambda_s && (16) \\ \frac{\partial}{\partial \Psi} f &= \frac{1}{2} \operatorname{diag}\left(\Sigma_s^{-1}(S - \Sigma_s)\Sigma_s^{-1} \right) && (17) \end{aligned}$$

with $S = (x - \mu_s)(x - \mu_s)^T$. The parametric constraints on π and Ω are easily incorporated by multiplying the gradient with the Jacobian of the softmax function.

Hence, the three likelihood objectives JL, CL and MCL can be minimized using any suitable gradient-based optimization routine.

4 Results

A mixture of factor analyzers was fit to random subsets of the lasso encoded foreground and background pixels with

[2] http://www.cs.berkeley.edu/projects/vision/grouping/segbench/

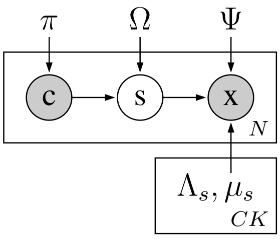


Figure 2. Graphical model of the generative MFA model used for pixel classification.

N between 500 and 5000 examples. A mixture of $K \in \{2, 3, 5, 10\}$ factor analyzers was used for each class and each factor analyzer had $p = 3$ latent dimensions.

For the optimization a BFGS quasi-Newton method was used. The same initial values, obtained with k-means, have been used for all likelihood criteria to obtain comparable results. Among five repetitions the run with the best likelihood criterion has been chosen and evaluated on the training data and on test data, the pixels in the inference region (*c.f.* the light gray border in Figure 1(a)).

The average test accuracies over the 50 images with $K = 5$ are shown in Figure 3(a). Despite considerable variance general tendencies are clearly visible. Whenever possible, statistical significance has been assessed with a Wilcoxon signed rank test. Training accuracies for the same models are shown in Figure 3(b).

Figure 4 shows the average test accuracies for different numbers of subcomponents used for the mixture of factor analyzers. Since each component is modeled by a factor analyzer the number of parameters increases correspondingly and so do the resulting model complexities. Results comparing $K \in \{2, 3, 5, 10\}$ are provided.

5 Discussion

As mentioned before the representational power of a probabilistic model with a certain number of parameters differs significantly between joint likelihood and conditional likelihood models. Multi-conditional models seek a tradeoff between these two extremes. A thorough analysis of the benefits of multi-conditional training must therefore consider the triple tradeoff between model complexity, amount of training data and test accuracy, fundamentally inherent to all supervised machine learning problems [2].

Along this line Figure 3(a) shows the test accuracy over the model complexity (as varied by α) for different sizes of the training data set. Comparing the spread of the test accuracies of CL and JL for different training sample sizes confirms the increased susceptibility to overfitting of the conditional model, in particular when contrasted with the corresponding training accuracies (*c.f.* Figure 3(b)). While for large training samples ($N = 5000$) the conditional model outperforms the test accuracy of the joint model ($p = .0088$) its test performance degrades more for smaller samples although training accuracies raise.

Increasing α in the multi-conditional model results in performance close to the performance of the joint model but not equal (Figure 3). For $\alpha = 1$ (pseudo-likelihood) the MCL model is in general better than JL ($p < .0001$) which might be surprising since pseudo-likelihood is known as a tractable approximation to JL which often exhibits inferior generalization performance. In the context of MCL however a “meaningful” pseudo-likelihood is constructed by partitioning the random variables in the two groups “class labels” c and “features” x and not with the primary goal of obtaining tractability.

Decreasing α on the other hand results in performance more similar to the conditional model (Figure 3). The best performance, however, is obtained for some α between 0 and 1. Figure 3(a) suggest that a small α around .01 constantly yields good results in the pixel classification problem. With $N = 5000$, for instance, the performance of the MCL model with $\alpha = .01$ is significantly better than the JL model ($p < .0001$) and has a tendency for being better than the CL model ($p = .0411$).

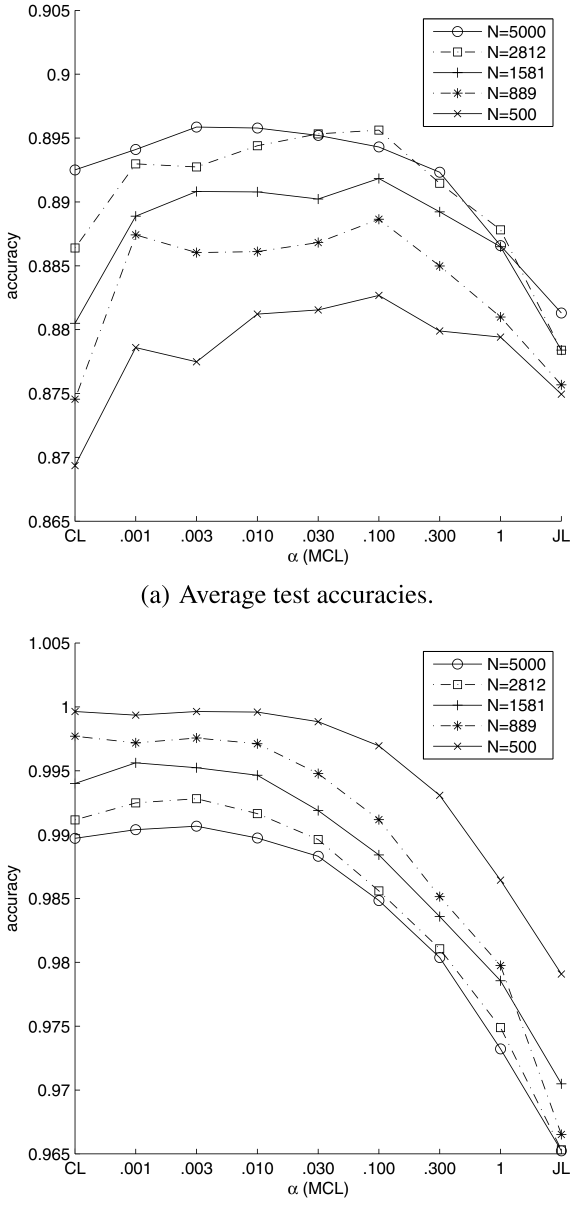
Figure 4 shows that across a variety of different base models the MCL approach can improve the performance consistently. In fact, for all mixture of factor analyzers with different numbers of subcomponents an α between .001 and .01 seems to be optimal. This also suggests that the choice of α within this interval is less critical than one could fear.

6 Conclusion

We have presented multi-conditional learning for simultaneous clustering, dimensionality reduction and classification. Starting with a mixture of factor analyzers we have shown that multi-conditional learning combines favorable properties of both the corresponding generative and discriminative models. In the context of the foreground/background pixel classification problem [1] we have demonstrated that a generalization performance superior to both can be achieved. Considering the random field approach in [1] and the positive results in [6] using random field models of documents we believe that multi-conditional methods derived from random field models for visual problems are a promising avenue of exploration.

7 Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval and Microsoft Research under grants from the Memex and eScience programs.



(a) Average test accuracies.

(b) Average training accuracies.

Figure 3. MFA models trained with various amounts (N) of the training data and w.r.t. joint (JL), conditional (CL) and multi-conditional (MCL, with different values of α) likelihood. Best test performance is obtained for MCL with α around $.01$. As opposed to the test accuracies the training accuracies do not decrease toward CL clearly indicating overfitting.

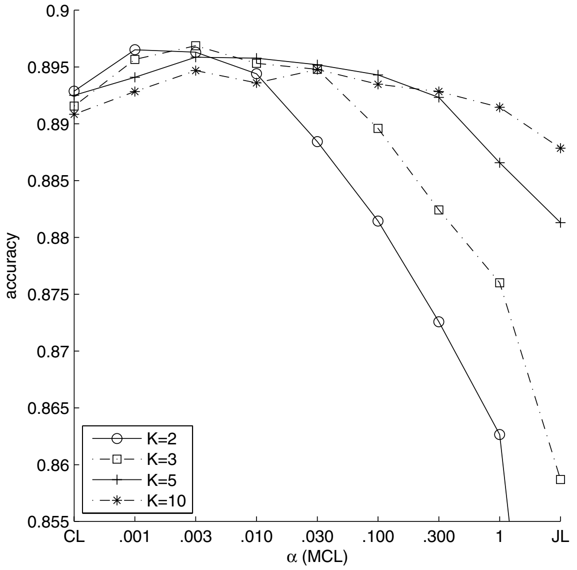


Figure 4. Average test accuracies for models with different numbers of subcomponents K. The JL performance benefits most from increasing model complexities. In all cases maximum performance is obtained with MCL for small α (between .001 and .01).

References

[1] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *Proc. ECCV*, 2004.

[2] T. G. Dietterich. Machine learning. In *Nature Encyclopedia of Cognitive Science*. London: Macmillan, 2003.

[3] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, Dept. of Computer Science, University of Toronto, 1997.

[4] B. Jähne. *Digital image processing*. Springer, Berlin, 5th edition, 2002.

[5] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the CEM algorithm. In *NIPS 11*, 1999.

[6] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative / Discriminative training for clustering and classification. In *Proc. AAAI*, 2006.

[7] T. Minka. Discriminative models, not discriminative training. Technical report, Microsoft Research Cambridge, 2005.

[8] A. Y. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS 14*, 2002.

[9] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with EM and expectation-conjugate-gradient. In *Proc. 20th ICML*, 2003.

[10] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In *Proc. CVPR*, 2005.