Maximum Entropy and Gaussian Models for Image Object Recognition

Daniel Keysers, Franz Josef Och, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, D-52056 Aachen, Germany
{keysers,och,ney}@informatik.rwth-aachen.de

Abstract. The principle of maximum entropy is a powerful framework that can be used to estimate class posterior probabilities for pattern recognition tasks. In this paper, we show how this principle is related to the discriminative training of Gaussian mixture densities using the maximum mutual information criterion. This leads to a relaxation of the constraints on the covariance matrices to be positive (semi-) definite. Thus, we arrive at a conceptually simple model that allows to estimate a large number of free parameters reliably. We compare the proposed method with other state-of-the-art approaches in experiments with the well known US Postal Service handwritten digits recognition task.

1 Introduction

The maximum entropy framework is based on principles applied in the natural sciences. It has been applied to the estimation of probability distributions [6] and to classification tasks such as natural language processing [1] and text classification [8].

The contributions of this paper are

- to show the relation between maximum entropy and Gaussian models,
- to present a framework that allows to estimate a large number of parameters reliably, e.g. the entries of full class specific covariance matrices, and
- to show the applicability of the maximum entropy framework to image object recognition.

2 Gaussian Models for Classification

To classify an observation $x \in \mathbb{R}^D$, we use the Bayesian decision rule

$$\begin{aligned} x \quad \longmapsto \quad r(x) &= \operatorname*{argmax}_k \{p(k|x)\} \\ &= \operatorname*{argmax}_k \{p(k) \cdot p(x|k)\} \,. \end{aligned}$$

Here, $p(k|x)$ is the class posterior probability of class $k \in \{1, \ldots, K\}$ given the observation x, $p(k)$ is the a priori probability, $p(x|k)$ is the class conditional probability for the observation x given class k and $r(x)$ is the decision of the classifier. This decision rule is known to be optimal with respect to the number of decision errors, if the correct distributions are known. This is generally not the case in practical situations, which means that we need to choose appropriate models for the distributions. In the training phase, the parameters of the distribution are estimated from a set of training data $\{(x_n, k_n)\}$, $n = 1, \ldots, N$, $k_n \in 1, \ldots, K$. If we denote by Λ the set of free parameters of the distribution, the maximum likelihood approach consists in choosing the parameters $\hat{\Lambda}$ maximizing the log-likelihood on the training data:

$$\hat{\Lambda} = \operatorname*{argmax}_{\Lambda} \sum_n \log p_\Lambda(x_n|k_n) \tag{1}$$

Alternatively, we can maximize the log-probability of the class posteriors,

$$\hat{\Lambda} = \operatorname*{argmax}_{\Lambda} \sum_n \log p_\Lambda(k_n|x_n) \ , \tag{2}$$

which is also called discriminative training, since the information of out-of-class data is used. This criterion is often referred to as mutual information criterion in speech recognition, information theory and image object recognition [3, 9].

We will regard Gaussian models for the class conditional distributions:

$$\begin{aligned} p(x|k) &= \mathcal{N}(x|\mu_k, \Sigma_k) \\ &= \det(2\pi\Sigma_k)^{-\frac{1}{2}} \cdot \exp\left[-\tfrac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right] \end{aligned} \tag{3}$$

The free parameters of these models are the class means μ_k and the class specific covariance matrices Σ_k. The conventional method for estimating these parameters is to maximize the log-likelihood (1) on the training data, which yields the empirical mean and the empirical covariance matrix as solutions. Problems with this approach arise if the feature dimensionality is large with respect to the number of training samples. This is common e.g. in appearance based image object recognition tasks, where each pixel value is considered a feature. The problems are that the large number of $K \cdot D \cdot (D+1)/2$ parameters of the covariance matrices often cannot be estimated reliably using the usually small amount of training data available. Common methods for coping with this problem are to constrain the covariance matrices, e.g. to use diagonal covariance matrices, or to use pooling, i.e. to estimate only one covariance matrix Σ instead of K matrices.

3 Maximum Entropy Modeling

The principle of maximum entropy has origins in statistical thermodynamics, is related to information theory and has been applied to pattern recognition tasks such as language modeling and text classification. Applied to classification, the basic idea is the following: We are given information about a probability distribution by samples from that distribution (training data). Now, we choose the distribution such that it fulfills all the constraints given by that information, but otherwise has the highest possible entropy. (This inherently serves as regularization to avoid overfitting.) It can be shown that this approach leads to so-called log-linear models for the distribution to be estimated.

Consider a set of so-called feature functions $\{f_i\}, i = 1, \ldots, I$ that are supposed to compute 'useful' information for classification:

$$f_i \quad : \quad \mathrm{I\!R}^D \times \{1, \ldots, K\} \longrightarrow \mathrm{I\!R} \quad : \quad (x, k) \longmapsto f_i(x, k)$$

From the information in the training set, we can compute the numbers

$$F_i := \sum_n f_i(x_n, k_n) \; .$$

Now, the maximum entropy principle consists in maximizing

$$\max_{p(k|x)} \left\{ -\sum_n \sum_k p(k|x_n) \log p(k|x_n) \right\}$$

over all possible distributions with the requirements:

– normalization constraint for each observation x:

$$\sum_k p(k|x) = 1$$

– feature constraint for each feature i:

$$\sum_n \sum_k p(k|x_n) f_i(x_n, k) = F_i$$

It can be shown that the resulting distribution has the following log-linear or exponential functional form:

$$p_\Lambda(k|x) = \frac{\exp\left[\sum_i \lambda_i f_i(x, k)\right]}{\sum_{k'} \exp\left[\sum_i \lambda_i f_i(x, k')\right]}, \quad \Lambda = \{\lambda_i\}. \tag{4}$$

Interestingly, it can also be shown that the stated optimization problem is convex and has a unique global maximum. Furthermore, this unique solution is also the solution to the following dual problem: Maximize the log probability (2) on the training data using the model (4). In this formulation of the problem, it is easier to see that there exists exactly one maximum, because (2) is a sum of convex functions and therefore also convex. A second desirable property of the discussed model is that effective algorithms are known that compute the global maximum of the log probability (2) given a training set. These algorithms fall into two categories: On the one hand, we have an algorithm known as generalized iterative scaling [4] and related algorithms that can be proven to converge to the global maximum. On the other hand, due to the convex nature of the criterion (2), we can also use general optimization strategies as e.g. conjugate gradient methods [10, pp. 420ff.]. The crucial problem in maximum entropy modeling is the choice of the appropriate feature functions $\{f_i\}$.

4 Maximum Entropy and Discriminative Training for Gaussian Models

Consider first-order feature functions for maximum entropy classification

$$
\begin{aligned}
f_{k,i}(x, k') &= \delta(k, k')\, x_i \ , \\
f_k(x, k') &= \delta(k, k') \ ,
\end{aligned}
$$

where $\delta(k, k') := 1$ if $k = k'$, and 0 otherwise denotes the Kronecker delta function. In the context of image recognition, we may call the functions $f_{k,i}$ appearance based image features, as they represent the image pixel values. The duplication of the features for each class is necessary to distinguish the hypothesized classes. The functions f_k allow for a log-linear offset in the posterior probabilities. Now, using the properties of the Kronecker delta, the structure of the posterior probabilities becomes

$$
\begin{aligned}
p_\Lambda(k|x) &= \frac{\exp\left[\alpha_k + \sum \lambda_{k,i} x_i\right]}{\sum_{k'} \exp\left[\alpha_{k'} + \sum \lambda_{k',i} x_i\right]} \\
&= \frac{\exp\left[\alpha_k + \lambda_k^T x\right]}{\sum_{k'} \exp\left[\alpha_{k'} + \lambda_{k'}^T x\right]} \qquad \Lambda = \{\lambda_{k,i}, \alpha_k\} \ ,
\end{aligned} \tag{5}
$$

where α_k denotes the coefficient for the feature function f_k.

Now, consider a Gaussian model (3) for $p(x|k)$ with pooled covariance matrix $\Sigma_k = \Sigma$. Using Bayes' rule, and the relation

$$
\begin{aligned}
\log \mathcal{N}(x|\mu_k, \Sigma_k) &= -\tfrac{1}{2} \log \det(2\pi \Sigma_k) - \tfrac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \\
&= -\tfrac{1}{2} \log \det(2\pi \Sigma_k) - \tfrac{1}{2} x^T \Sigma_k^{-1} x + \mu_k^T \Sigma_k^{-1} x - \tfrac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k \ ,
\end{aligned}
$$

we can rewrite the class posterior probability (note that the terms that do not depend on the class k cancel in the fraction):

$$
\begin{aligned}
p(k|x) &= \frac{p(k)\, \mathcal{N}(x|\mu_k, \Sigma)}{\sum_{k'} p(k')\, \mathcal{N}(x|\mu_{k'} \Sigma)} \\
&= \frac{\exp\left[(\log p(k) - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k) + (\mu_k^T \Sigma^{-1}) x\right]}{\sum_{k'} \exp\left[(\log p(k') - \frac{1}{2}\mu_{k'}^T \Sigma^{-1} \mu_{k'}) + (\mu_{k'}^T \Sigma^{-1}) x\right]} \\
&= \frac{\exp\left[\alpha_k + \lambda_k^T x\right]}{\sum_{k'} \exp\left[\alpha_{k'} + \lambda_{k'}^T x\right]}
\end{aligned} \tag{6}
$$

As result, we see that for unknown class priors $p(k)$ the resulting model (6) is identical to the maximum entropy model (5). We can conclude that the discriminative training criterion (2) for the Gaussian model (3) with pooled covariance matrices results in exactly the same functional form as the maximum entropy model for first-order features. This allows to use the well understood algorithms for maximum entropy estimation to estimate the parameters of a Gaussian model discriminatively.

If we repeat the same argument as above for the case of Gaussian densities without pooling of the covariance matrices, we find that we can again establish a correspondence to a maximum entropy model:

$$\begin{aligned} p(k|x) &= \frac{p(k)\ \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{k'} p(k')\ \mathcal{N}(x|\mu_{k'} \Sigma_k)} \\ &= \frac{\exp\left[\alpha_k + \lambda_k^T x + x^T S_k x\right]}{\sum_{k'} \exp\left[\alpha_{k'} + \lambda_{k'}^T x + x^T S_{k'} x\right]} \end{aligned}$$

Here, the square matrix S_k corresponds to the negative of the inverse of the covariance matrix Σ_k. These parameters can be estimated using a maximum entropy model with the second-order feature functions

$$\begin{aligned} f_{k,i,j}(x, k') &= \delta(k, k')\ x_i x_j\ , \quad i \geq j\ , \\ f_{k,i}(x, k') &= \delta(k, k')\ x_i\ , \\ f_k(x, k') &= \delta(k, k')\ . \end{aligned}$$

One interesting consequence of using the corresponding maximum entropy model and estimation is that we implicitly relax the constraints on the covariance matrices to be positive (semi-) definite. Therefore, the resulting model is not exactly equivalent to a Gaussian model.

This result is in contrast to the approach taken in [5], where the authors derive discriminative models for Gaussian densities based on priors of the parameters and the minimum relative entropy principle. Their solution results in discriminatively trained weights for the training data and therefore preserves the mentioned constraints.

5 Experiments and Results

We performed experiments on the well known US Postal Service handwritten digit recognition task (USPS). It contains normalized greyscale images of handwritten digits taken from US zip codes of size 16×16 pixels. The corpus is divided into a training set of 7,291 images and a test set of 2,007 images. Reported recognition error rates for this database are summarized in Table 1.

In most of the experiments performed we obtained better results using 'feature normalization'. This means that we enforced for each observation during training and testing that the sum of all feature values is equal to one by scaling the feature values appropriately. Thus, we obtain new feature functions $\{\tilde{f}_i\}$:

$$\forall x, k, i: \ \tilde{f}_i(x, k) = \Big(\sum_{i'} f_{i'}(x, k)\Big)^{-1} \cdot f_i(x, k)$$

In the following, we only report result obtained using feature normalization. The parameters were trained using generalized iterative scaling [4].

Table 2 shows the main results obtained in comparison to other approaches along with the number of free parameters of the respective models. The error

Table 1. Summary of results for the USPS corpus (error rates, [%]). *: training set extended with 2,400 machine-printed digits

method		ER[%]
human performance	[SIMARD et al. 1993] [14]	2.5
relevance vector machine	[TIPPING et al. 2000] [15]	5.1
neural net (LeNet1)	[LECUN et al. 1990] [13]	4.2
support vectors	[SCHÖLKOPF 1997] [11]	4.0
invariant support vectors	[SCHÖLKOPF et al. 1998] [12]	3.0
neural net + boosting	[DRUCKER et al. 1993] [13]	*2.6
tangent distance	[SIMARD et al. 1993] [14]	*2.5
nearest neighbor classifier	[7]	5.6
mixture densities	[2] baseline	7.2
	+ LDA + virtual data	3.4
kernel densities	[7] baseline	5.5
	+ tangent vectors + virtual data	2.4

rates show that we can already gain recognition accuracy by using the maximum entropy framework to only estimate the pooled covariance matrix of a Gaussian model, while fixing the mean vectors to their maximum likelihood values. Taking into account the class information in training using the maximum entropy framework increases the recognition accuracy for first-order features from 18.6% to 8.2% error rate using less parameters.

Furthermore, it can be observed that the maximum entropy models perform better for second-order features than for first-order features. This is in contrast to the experience gained with maximum likelihood estimation of Gaussian densities, where best results were obtained using pooled diagonal covariance matrices [2]. Note for example that the maximum likelihood estimation of class specific diagonal covariance matrices already imposes problems for the USPS data, because in some of the classes some of the dimensions have zero variance in the training data. This can be overcome e.g. by using interpolation with the identity matrix, but the maximum entropy framework offers an effective way to overcome these problems.

Using the equivalent of a full class specific covariance matrix, i.e. second-order features, the error rate of a 'pseudo Gaussian' model with 5.7% error rate

Table 2. Overview of the results obtained on the USPS corpus using maximum entropy modeling in comparison to other models (error rates, [%]). ML: maximum likelihood, MMI: maximum mutual information, *: with pooled diagonal covariance matrix.

model	training criterion	# parameters	ER[%]
Gaussian model*	ML	2 816	18.6
	Σ: MMI, μ_k: ML	2 816	14.2
maximum entropy, first-order features	MMI	2 570	8.2
second-order features	MMI	331 530	5.7
nearest neighbor classifier		1 866 496	5.6

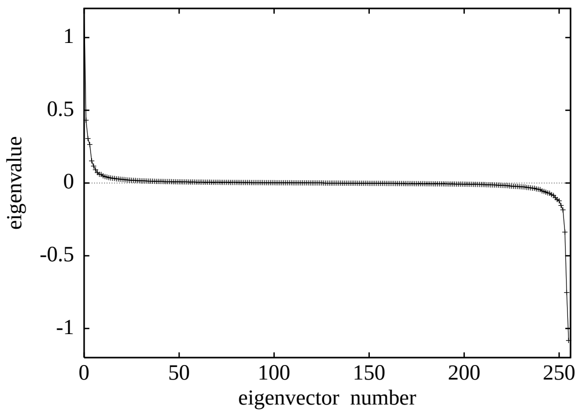

Fig. 1. Eigenvalue distribution for the 'covariance matrix' of the class '5', estimated using the maximum entropy approach.

approaches that of a nearest neighbor classifier, which has more than five times as many parameters.

Fig. 1 shows the eigenvalues of the 'covariance matrix' of this 'pseudo Gaussian' model for the class '5' ordered by size. It can be observed that about half of the eigenvalues are positive, while the other half is negative. The distribution of the negative eigenvalues seems to match the distribution of the positive eigenvalues. We can conclude that besides the typical important eigenvectors with large positive eigenvalues there are also important eigenvectors with large negative eigenvalues in this discriminative context. This means that the relaxation of the constraint on the covariance matrix to be positive (semi-) definite leads to discriminative models that are not Gaussian any more.

6 Conclusion

We presented the connection between the following classification models: (a) discriminative training using the maximum mutual information criterion of Gaussian models for the class conditional probability and (b) models for the class posterior probability based on the principle of maximum entropy. We showed that these models lead to identical functional forms for the correct choice of feature functions for the maximum entropy model. One of the main differences is that the maximum entropy model implicitly relaxes the constraint on the covariance matrices to be positive (semi-) definite. This leads to a conceptually simpler model with well understood estimation algorithms. A further advantage of the maximum entropy approach is that it is easily possible to include new feature functions into the classifier.

We evaluated the approach for image object recognition using the US Postal Service handwritten digits recognition task, obtaining significant improvements with respect to maximum likelihood based training. The best result of 5.7% error rate using second-order features is competitive with other results reported on this dataset, although approaches with significantly better performance exist. (Note that the latter are highly tuned to the specific task at hand while the maximum entropy approach is of very general nature.) The accuracy of the resulting model shows that the maximum entropy approach allows robust estimation of the equivalent of full covariance matrices even on this small training set, which may be a problem for approaches based on maximum likelihood.

References

1. A.L. Berger, S.A. Della Pietra, V.J. Della Pietra: A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–72, March 1996.
2. J. Dahmen, D. Keysers, H. Ney, M.O. Güld: Statistical Image Object Recognition using Mixture Densities. *J. Mathematical Imaging and Vision*, 14(3):285–296, May 2001.
3. J. Dahmen, R. Schlüter, H. Ney: Discriminative Training of Gaussian Mixture Densities for Image Object Recognition. In *21. DAGM Symposium Mustererkennung*, Bonn, Germany, pp. 205–212, September 1999.
4. J.N. Darroch, D. Ratcliff: Generalized Iterative Scaling for Log-Linear Models. *Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
5. T. Jaakkola, M. Meila, T. Jebara: Maximum Entropy Discrimination. In *Advances in Neural Information Processing Systems 12*, MIT Press, Cambridge, MA, pp. 470–476, 2000.
6. E.T. Jaynes: On the Rationale of Maximum Entropy Models. *Proc. of the IEEE*, 70(9):939–952, September 1982.
7. D. Keysers, J. Dahmen, T. Theiner, H. Ney: Experiments with an Extended Tangent Distance. In *Proc. 15th IEEE Int. Conf. on Pattern Recognition*, volume 2, Barcelona, Spain, pp. 38–42, September 2000.
8. K. Nigam, J. Lafferty, A. McCallum: Using Maximum Entropy for Text Classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden, pp. 61–67, August 1999.
9. Y. Normandin: Maximum Mutual Information Estimation of Hidden Markov Models. In C.H. Lee, F.K. Soong, K.K. Paliwal (Eds.): *Automatic Speech and Speaker Recognition*, Kluwer Academic Publishers, Norwell, MA, pp. 57–81, 1996.
10. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery: *Numerical Recipes in C.* Cambridge University Press, Cambridge, second edition, 1992.
11. B. Schölkopf: *Support Vector Learning.* Oldenbourg Verlag, Munich, 1997.
12. B. Schölkopf, P. Simard, A. Smola, V. Vapnik: Prior Knowledge in Support Vector Kernels. In *Advances in Neural Information Processing Systems 10.* MIT Press, pp. 640–646, 1998.
13. P. Simard, Y. Le Cun, J. Denker, B. Victorri: Transformation Invariance in Pattern Recognition — Tangent Distance and Tangent Propagation. In G. Orr, K.R. Müller (Eds.): *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, Springer, Heidelberg, pp. 239–274, 1998.
14. P. Simard, Y. Le Cun, J. Denker: Efficient Pattern Recognition Using a New Transformation Distance. In *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann, San Mateo, CA, pp. 50–58, 1993.
15. M.E. Tipping: The Relevance Vector Machine. In *Advances in Neural Information Processing Systems 12.* MIT Press, pp. 332–388, 2000.