ISCAS 2000 - IEEE International Symposium on Circuits and Systems, May 28-31, 2000, Geneva, Switzerland

MAXIMUM ENTROPY AND MAXIMUM LIKELIHOOD CRITERIA FOR FEATURE SELECTION FROM MULTIVARIATE DATA

Sankar Basu, Charles A. Micchelli, Peder Olsen

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
sankar,cam,pederao@watson.ibm.com

ABSTRACT

We discuss several numerical methods for optimum feature selection for multivariate data based on maximum entropy and maximum likelihood criteria. Our point of view is to consider observed data $x^1, x^2, \ldots, x^N$ in $\mathbf{R}^d$ to be samples from some unknown pdf P. We project this data onto d directions, subsequently estimate the pdf of the *univariate* data, then find the maximum entropy (or likelihood) of all multivariate pdfs in $\mathbf{R}^d$ with marginals in these directions prescribed by the estimated univariate pdfs and finally *maximize* the entropy (or likelihood) further over the choice of these directions. This strategy for optimal feature selection depends on the method used to estimate *univariate* data.

1. INTRODUCTION

There exists a large literature on estimating probability densities associated with univariate data. However, corresponding methods for estimating multivariate probability densities prove to be problematic due to various reasons [1]. This is especially true in the context of very high dimensions. Crucial among these difficulties is the fact that the data appears to be increasingly sparse with the increase in the number of dimensions. We approach the problem by considering projections of the high dimensional data on lower dimensional subspaces, say a single dimensional subspace, subsequently by estimating the univariate probability densities via known univariate techniques, and then by reconstructing the density in the original higher dimensional space from the collection of univariate densities so obtained. In some sense the approach is reminiscent of function reconstruction from projections (e.g., in computerized tomography). The reconstructed density is by no means unique unless further restrictions on the estimated density are imposed. The variety of choices of candidate univariate densities as well as the choices of subspaces on which to project the data including their number further add to this nonuniqueness. One can then consider pdf's that maximize certain optimility criteron as a solution to this problem. For the purpose of the present paper, we consider those pdf's that either maximize the entropy functional, or alternatively, the likelihood associated with the data.

2. MAXIMUM ENTROPY FROM ITS MARGINALS

In this section we compute the maximum of the entropy of a probability density function (pdf) on $\mathbf{R}^d$ with prescribed marginals in d linearly independent (column) vectors, $y_1, y_2, \ldots, y_d \in \mathbf{R}^d$. Let Y be the $d \times d$ matrix whose columns are these vectors. Thus, we are given univariate pdf's, $p_1, p_2, \ldots, p_d$ and consider all multivariate pdf's $P : \mathbf{R}^d \to \mathbf{R}$ such that for all $f \in C(\mathbf{R})$ (the space of continuous functions on $\mathbf{R}$) we have that

$$\int_{\mathbf{R}^d} f(y_i^T x) P(x) dx = \int_{\mathbf{R}} f(t) p_i(t) dt. \tag{1}$$

We denote the class of all such functions by $\mathcal{C}(p)$. Recall that the entropy of P is given by

$$H(P) = -\int_{\mathbf{R}^d} P(x) \log P(x) dx \tag{2}$$

see [3].

Theorem 1 *Given any pdf's $p_1, p_2, \ldots, p_d$ on $\mathbf{R}$*

$$\max\{H(P) \ : \ P \in \mathcal{C}(p)\} = -\frac{1}{2} \log \det G + \sum_{i=1}^{d} H(p_i) \ . \tag{3}$$

where $G = Y^T Y$ is the gram matrix

$$G(Y) = [y_i^T y_j]_{i,j=1,\ldots,d}. \tag{4}$$

Moreover,

$$P^*(x) = \sqrt{\det G(Y)} \prod_{i=1}^{d} p_i(y_i^T x) \ , \ x \in \mathbf{R}^d \tag{5}$$

is the unique pdf in $\mathcal{C}(p)$ at which the maximum in (3) is achieved.

In the next section we shall use this observation for maximum entropy based methods for feature selection of data.

3. FEATURE SELECTION USING MAXIMUM ENTROPY

Our goal in this section is to use Theorem 1 as a means to isolate desirable features of multivariate data. Thus, for univariate data $x = (x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$, we assume $p(\cdot; x)$ to be a reasonably good estimator for the true pdf on $\mathbf{R}$. We shall consider several standard choices of this family later. For a given $y \in \mathbf{R}^d$ of norm one, we form the vector of univariate data $X^T y \in \mathbf{R}^N$ where X is the $d \times N$ matrix whose columns are the multivariate data $x^1, x^2, \ldots, x^N$ and use $p(\cdot; X^T y)$ as an estimate for the marginal pdf of P in the direction y. Given d directions

0-7803-5482-6/99/$10.00 ©2000 IEEE

$y_1, y_2, \ldots, y_d$ following Theorem 1, our maximum entropy estimator for the unknown pdf P is

$$P^*(x) = \sqrt{\det G(Y)} \prod_{j=1}^{d} p(y_j^T x; X^T y_j) \ , \ x \in \mathbf{R}^d \ , \qquad (6)$$

and

$$\mathcal{J}(Y) := H(P^*) = -\frac{1}{2} \log \det G(Y) + \sum_{j=1}^{d} \mathcal{H}(y_j) \ , \qquad (7)$$

where we define

$$\mathcal{H}(y) = H(p(\cdot; X^T y)) \ , \ y \in \mathbf{R}^d \ . \qquad (8)$$

We desire to make the entropy (7) as large as possible by varying our choice of vectors $y_1, y_2, \ldots, y_d$. This is not fruitful without additional constraints on the vectors $y_1, y_2, \ldots, y_d$. Indeed, we can have independent vectors which are nearly dependent so that their grammian is nearly singular. Hence, the first term in (7) will be large. Our preference is to choose an *orthonormal basis* so that the corresponding grammian is the identity matrix. We postpone the important issue of choosing the optimal orthonormal basis until later. Instead, we turn to an example. We have in mind as our univariate estimate for the data vector $x = (x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$ the gaussian density

$$p(t; x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t-\mu)^2}{2\sigma}} \ , \ t \in \mathbf{R}, \qquad (9)$$

with mean

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j \qquad (10)$$

and variance

$$\sigma = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu)^2 \ . \qquad (11)$$

Thus, we have that $H(p(\cdot; x)) = \frac{1}{2} \log(2\pi e \sigma)$ and the function we shall optimize is

$$\mathcal{J}(Y) = -\frac{1}{2} \log \det G(Y) + \frac{d}{2} \log(2\pi e) + \frac{1}{2} \sum_{j=1}^{d} \log y_j^T V y_j \ , \qquad (12)$$

where V is the $d \times d$ matrix $V = UU^T$ and U is the $d \times N$ matrix, whose columns are the vectors

$$u_j = \frac{1}{\sqrt{N}} (x^j - \bar{x}), \ j = 1, 2, \ldots, N \qquad (13)$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x^i, \qquad (14)$$

We now discuss the maximum and minimum of $\mathcal{J}(Y)$ over all $d \times d$ orthogonal matrices Y. First we consider the *minimum* of this function. We suppose that $v_1, v_2, \ldots, v_d$ are orthonormal eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ with $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ for V. We let

$$D = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_d]$$

and O be the orthogonal $d \times d$ matrix whose columns are $v_1, v_2, \ldots, v_d$, i.e., $V = O^T D O$. We then have that

$$\min\{\mathcal{J}(Y) : Y^T Y = I\}$$

$$= \mathcal{J}(O) = \frac{1}{2} d \log(2\pi e) + \frac{1}{2} \sum_{j=1}^{d} \log \lambda_j \ .$$

Note that the vectors at which the minimum is attained are *orthonormal*. The connection of this observation to maximum likelihood estimation (MLE) will be further elaborated upon and will be made available in [2].

Let us now consider the problem of *maximizing* the function $\mathcal{J}$. As mentioned before, in this case, we restrict Y to be *orthogonal* and obtain

$$\max\{\mathcal{J}(Y) : \mathrm{Y \ \ is \ \ orthogonal}\} = \mathcal{J}(R) \qquad (15)$$

where the columns $r_1, r_2, \ldots r_d$ of the matrix R satisfy for all $j = 1, 2, \ldots, d$

$$r_j^T V r_j = \frac{1}{d} \sum_{i=1}^{d} \lambda_i. \qquad (16)$$

The existence of an orthogonal matrix R with columns $r_1, r_2, \ldots, r_d$ such that (16) is satisfied is ensured: Suppose H is a $d \times d$ *Hadamard matrix*. We set $R = d^{-\frac{1}{2}} HO$ and so that R and its columns $r_1, r_2, \cdots, r_d$ satisfy for $i = 1, 2, \ldots, d$ the equations

$$w_i^T V w_i = \frac{1}{d} \sum_{j=1}^{d} \lambda_j. \qquad (17)$$

4. FEATURE SELECTION USING MAXIMUM ENTROPY AND MIXTURES OF GAUSSIANS

Often one finds modeling univariate data by a single gaussian to be inadequate. We, therfore, consider a mixture model of m gaussians

$$p(t|\mu, \pi, \sigma) = \sum_{j=1}^{m} \frac{\pi_j}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{(t-\mu_j)^2}{2\sigma_j}\right), \quad x \in \mathbf{R}, \qquad (18)$$

where the mixture weights satisfy the constraints

$$\sum_{j=1}^{m} \pi_j = 1 \ \ \pi_i > 0; \ \ i = 1, 2, \ldots, m. \qquad (19)$$

Our first requirement is to compute the entropy function of p given a choice of means $\mu_1, \mu_2, \ldots, \mu_m \in \mathbf{R}$, variances $\sigma_1, \sigma_2, \ldots, \sigma_m > 0$ and mixture weights $\pi_1, \pi_2, \ldots, \pi_m$. We suggest use of a Monte-Carlo approach based on data generated from the pdf given in (18).

We turn to the problem of selecting the means and variances of the pdf in (19) from a given but arbitrary data vector $(x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$. As before, we take an MLE approach. However, as is well known for $m \ge 2$ the likelihood function will not have a maximum, see for example [9]. Instead, we cannot use it as an absolute criterion to select the means and variances. Therefore, we turn to the EM algorithm to *iteratively* set the means and variances. Here we fix a number of iterations in the EM update formulas appearing below *independent* of the data vector x. The first step is the initialization of the means and variances. This is done by *clustering* the

data vector into m groups of (approximately) equal size starting with its first coordinate and proceeding sequentially through all its components. On each of the m clusters we use the formulas (10) and (11) for the mean and the variances. We then update them by the EM formulas, (see, for instance, [4, 7]), which produces a sequence of MLE's whose likelihoods are nondecreasing. We now denote by $p(\cdot; x)$ our univariate estimate for our data vector $x \in \mathbf{R}^N$ by mixture models of gaussians.

The next step is to iteratively update the orthogonal matrix Y. Our iteration is initialized per our previous discussion of the single gaussian model. Consequently, all our experiments are performed in d-dimensions where d is chosen to admit a Hadamard matrix. We then modify the eigenvalues of the matrix $V = UU^T$ corresponding to the column vectors $w_1, w_2, \ldots, w_N$ given (17) per our discussions of the single gaussian model. Thus, for any vector $y \in \mathbf{R}^d$ and data $x^1, x^2, \ldots, x^N \in \mathbf{R}^d$ organized into a $d \times N$ matrix X whose columns are $x^1, x^2, \ldots, x^N \in \mathbf{R}^d$ we can evaluate the function

$$\mathcal{H}(y) = H(p(\cdot; X^T y)).$$

Our next task is to iteratively update our orthonormal feature vectors $y_1, y_2, \ldots, y_d$ to enhance our entropy functional (8) which specializes to the form

$$\mathcal{J}(Y) = \sum_{j=1}^{d} \mathcal{H}(y_j).$$

To update the vectors we used two dimensional *planar rotations*. Thus, we choose two consecutive vectors y_j, y_{j+1} and form new vectors

$$\begin{aligned} y_\ell(\theta) &= y_\ell, \quad \ell \neq j, j+1, j = 1, 2, \ldots, d, && (20)\\ y_j(\theta) &= y_j \cos\theta - y_{j+1} \sin\theta && (21)\\ y_{j+1}(\theta) &= y_j \sin\theta + y_j \cos\theta && (22)\\ & && (23) \end{aligned}$$

where $\theta \in [0, 2\pi]$. We then optimize the univariate function

$$F(\theta) = \mathcal{H}(y_j(\theta)) + \mathcal{H}(y_{j+1}(\theta))$$

over $\theta \in [0, 2\pi]$ and replace $y_1, y_2, \ldots y_d$ by $y_1(\hat{\theta}), y_2(\hat{\theta}), \ldots y_d(\hat{\theta})$ where $\hat{\theta}$ is a good choice for the maximum of the function F. If $j < d-1$, we increment j by one and repeat the procedure otherwise we set $j = 1$ and repeat the procedure, cycling several times through the columns of Y. To avoid computing the derivatives of F we compute its values on an acceptably fine grid and choose its largest value.

5. FEATURE SELECTION USING MAXIMUM LIKELIHOOD AND PARZEN ESTIMATORS

In Section 4 we considered modeling our multivariable data by choosing our univariate marginals $p(\cdot; x)$, $x = (x_1, x_2, \ldots, x_N)^T \in \mathbf{R}^N$ to be an MLE mixture model of gaussians.

In this section, we study the possibility of using maximum likelihood estimation as a method for optimum feature selection. Thus, as before, we suppose that $p(\cdot; x)$ is a pdf on $\mathbf{R}$ which models univariate data $x = (x_1, x_2, \cdots, x_N)^T$.

Given d orthonormal vectors $y_1, y_2, \ldots, y_d$ and corresponding marginals $p(\cdot; X^T y)$ the maximum entropy estimator is given by Theorem 1 as

$$P^*(x) = \prod_{j=1}^{d} p(y_j^T x; X^T y_j) \ , \ x \in \mathbf{R}^d \ . \qquad (24)$$

For numerical efficiency in this section we propose to choose the orthogonal matrix Y that maximizes the loglikelihood function,

$$L(Y) = \log \prod_{i=1}^{N} P^*(x^i) \ . \qquad (25)$$

We consider the problem of maximizing $L(Y)$ for two specific choices of $p(\cdot; x)$. In the first case, we consider a Parzen estimator based on the gaussian. In other words, we choose for $x = (x_1, x_2, \ldots, x_N)^T$ and a bin width h the pdf

$$p(t; x) = \frac{1}{N} \sum_{j=1}^{N} K(\frac{t - x_j}{h}) \ , t \in \mathbf{R}.$$

where

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right), \quad t \in \mathbf{R}.$$

For a good choice of h see [10].

We also consider using a univariate mixtures of gaussians

$$p(x) = \sum_{\ell=1}^{m} \frac{\omega_\ell}{\sqrt{2\pi}\sigma_\ell} \exp(-\frac{1}{2}(x - \mu_\ell)^2).$$

Taking advantage of the special analytical structure of these two univariate densities combined with the planar rotations and EM strategy [7, 8] described before in Section 4, we develop an effective, efficient numerical strategy for the maximizing the loglikelihood $L(Y)$, see [2] for the full details.

6. REFERENCES

[1] David W. Scott, *Multivariate Density Estimation*, Wiley Interscience, 1992.

[2] S. Basu, Charles A. Micchelli, Peder Olsen Maximum entropy for feature selection, in preparation.

[3] I. M. Cover and J. A. Thomas, Elements of Information Theory, John Wiley & Sons, Inc. 1991.

[4] C. M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.

[5] C. Miller and Roger Horn, Probability density estimation using entropy maximization, Neural Computation, vol. 10, pp.1925-1938, 1998.

[6] Nagendra Kumar, Investigation of silicon auditory model and generalization of linear discriminant analysis for improved speech recognition, Ph.D dissertation, Johns Hopkins University, 1997.

[7] A. P. Dempster, N. M. Laird and D. B. Baum, Maximum likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Soc., Ser. B, vol. 39, pp. 1-38, 1977.

[8] L. E. Baum, Ted Petrie, George Soules, Norman Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, The Annals of Math. Stat., vol. 41, No. 1, 164-171, 1970,

[9] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, 1995.

[10] Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CPC, 1996.