Unsupervised Learning and Generalization[*]

Lars Kai Hansen and Jan Larsen
CONNECT, Section for Digital Signal Processing
Department of Mathematical Modelling, B349
Technical University of Denmark
DK-2800 Lyngby, Denmark
Phones: +45 4525+ ext. 3889,3923
Fax: +45 45880117
emails: lkhansen,jlarsen@ei.dtu.dk

ABSTRACT

The concept of generalization is defined for a general class of unsupervised learning machines. The generalization error is a straightforward extension of the corresponding concept for supervised learning, and may be estimated empirically using a test set or by statistical means – in close analogy with supervised learning. The empirical and analytical estimates are compared for Principal Component Analysis and for K-means clustering based density estimation.

1. Introduction

The goal of unsupervised learning is to identify and explore regularities and dependencies in data (for an introduction see e.g., [4]). Principal Component Analysis (PCA) [5] and Clustering [3] are two prominent examples of unsupervised learning schemes that are widely used in applications. Like supervised learning schemes, unsupervised learning proceeds from a finite sample of training data. This means that the learned concepts are stochastic variables depending on the particular (random) training set. This opens the question of robustness and generalization: how robust are the learned concepts to fluctuation and noise in the training set, and how well will they perform on a new (test) datum? Generalization is a key topic in the theory of supervised learning, and significant progress has been reported. The most universally applicable algebraic results were recently published by Murata *et al.* [10], describing the asymptotic generalization ability of supervised algorithms that are continuously parameterized.

The aim of this paper is to extend the theory of Murata *et al.* to unsupervised learning and show how it may be used to optimize the generalization performance of PCA and clustering.

2. Generalization

While supervised learning concerns the identification of *functional dependencies*, the objective of unsupervised learning is to capture statistical dependencies, i.e., the structure of the underlying data distributions. In both cases we are interested in robust modeling, i.e., that the knowledge obtained is generic and, as far as possible, independent of the particular training set provided. It is a common observation that good generalization is obtained when the model capacity is well optimized[1]. If the model capacity is too limited the model will not be able capture the full complexity of the distributions, while a high capacity model will support many different solutions to the learning problem and is likely to focus on non-generic details of the particular training set (overfitting). The proposed scheme attempts to estimate the effects of the finite random training set, with the purpose of minimizing their role.

[*]Appears in Proceedings of the IEEE International Conference on Neural Networks, Washington DC, June 1996. This manuscript differs from the printed version due to minor corrections.

[1]Also referred to as the bias/variance dilemma.

Like [10] we analyze models that are smoothly parametrized and whose training can be described in terms of a cost function. If a particular data vector is denoted x and the model, denoted H, involves the parameter vector θ, the associated cost will be denoted by $\epsilon(x|\theta, H)$.

A training set is a finite sample $D = \{x_\alpha\}_{\alpha=1}^N$ of the stochastic vector x. Let $p(x)$ be the "true" distribution of x, while the empirical distribution associated with D, is given by $p_e(x) = 1/N \sum_{\alpha=1}^N \delta(x - x_\alpha)$. For a specific model and a specific set of parameters we define the training and generalization errors as follows,

$$
\begin{aligned}
E(\theta, H) &= \int dx\, p_e(x)\epsilon(x|\theta, H) = \frac{1}{N}\sum_{\alpha=1}^{N} \epsilon(x_\alpha|\theta, H), &(1)\\
G(\theta, H) &= \int dx\, p(x)\epsilon(x|\theta, H). &(2)
\end{aligned}
$$

Note that the generalization error is non-observable, i.e., it has to be estimated either from a finite *test set* drawn from $p(x)$, or estimated from the training set using statistical arguments.

3. Analytical Estimate of the Generalization Error

To estimate the generalization error we will assume that learning results in the selection of the parameters $\widehat{\theta}$ pertinent to the specific training set by solving $\partial C/\partial\theta = 0$ where C is the cost function $C(\theta) = E(\theta) + R(\theta)$ including a regularization term $R(\theta)$. Note even though learning is done by minimizing $C(\theta)$, the generalization error is still defined as in (2). This matter is, however, not discussed in [10].

Hence $\widehat{\theta} = \theta(D, H)$, is a stochastic variable. Let $m = \dim(\theta)$ be the dimensionality of the parametrization the model. Under fairly mild conditions it is possible to show that in the limit $m/N \to 0$, the distribution of $\widehat{\theta}$ becomes asymptotically Gaussian, $\widehat{\theta} \sim \mathcal{N}(\theta^*, \Sigma^*)$ where the optimal value $\theta^* = \arg\min_\theta(G + R)$, while the covariance matrix is given by,

$$\Sigma^* = \frac{1}{N} J^{-1} Q J^{-1} \tag{3}$$

where the matrices, J, Q, are defined by,

$$Q_{ij} = \int dx\, p(x) \frac{\partial\epsilon(x|\theta)}{\partial\theta_i}\frac{\partial\epsilon(x|\theta)}{\partial\theta_j} + \frac{\partial R(\theta)}{\partial\theta_i}\frac{\partial R(\theta)}{\partial\theta_j}, \quad J_{ij} = \int dx\, p(x)\frac{\partial^2\epsilon(x|\theta)}{\partial\theta_i\partial\theta_j} + \frac{\partial^2 R(\theta)}{\partial\theta_i\partial\theta_j}. \tag{4}$$

Since $\Sigma^* \propto 1/N$ we can estimate the generalization error by expanding in the small fluctuations around θ^* induced by the training set,

$$\langle G(\widehat{\theta})\rangle_D \approx G(\theta^*) + R(\theta^*) - \langle R(\widehat{\theta})\rangle_D + \frac{1}{2}\text{Trace}\left[J\langle\Delta\theta\Delta\theta^\top\rangle_D\right]. \tag{5}$$

$\langle\cdot\rangle_D$ signifies the average over all possible training sets of size N. We have denoted the fluctuation of the optimal parameters as $\Delta\theta = \widehat{\theta} - \theta^*$. Inserting the expression for the covariance of these fluctuations as given by (3) we find

$$\langle G(\widehat{\theta})\rangle_D \approx G(\theta^*) + R(\theta^*) - \langle R(\widehat{\theta})\rangle_D + \frac{1}{2N}\text{Trace}\left[QJ^{-1}\right]. \tag{6}$$

The *order one* matrices Q, J may in turn be estimated from the empirical distribution $p_e(x)$,

$$\widehat{Q}_{ij} = \frac{1}{N}\sum_{\alpha=1}^{N}\frac{\partial\epsilon(x_\alpha|\theta)}{\partial\theta_i}\frac{\partial\epsilon(x_\alpha|\theta)}{\partial\theta_j} + \frac{\partial R(\theta)}{\partial\theta_i}\frac{\partial R(\theta)}{\partial\theta_j}, \quad \widehat{J}_{ij} = \frac{1}{N}\sum_{\alpha=1}^{N}\frac{\partial^2\epsilon(x_\alpha|\theta)}{\partial\theta_i\partial\theta_j} + \frac{\partial^2 R(\theta)}{\partial\theta_i\partial\theta_j}. \tag{7}$$

However, this still leaves us with the unknown "noise level" $G(\theta^*)$. Fortunately this quantity may be estimated from the averaged training error. By expansions similar to those entering the estimation of the generalization we find,

$$\langle E(\widehat{\theta})\rangle_D \approx G(\theta^*) + R(\theta^*) - \langle R(\widehat{\theta})\rangle_D - \frac{1}{2N}\text{Trace}\left[QJ^{-1}\right]. \tag{8}$$

Which by elimination of the noise level leads us to the final relation,

$$\langle G(\widehat{\theta})\rangle_D \approx \langle E(\widehat{\theta})\rangle_D + \frac{\text{Trace}\left[QJ^{-1}\right]}{N}. \tag{9}$$

The relation (9) provides a link between the averaged training and test errors analogous to the results of Murata *et al.* [10], and similar to Akaike's Final Prediction Error [1], thus extending these key results to unsupervised learning schemes. For further discussion see also [7].

4. Examples

To illustrate how the generalization estimate (9) is used in more specific contexts, we analyze two important unsupervised learning machines, in particular we show how one may select the optimal number of principal components in PCA and, secondly, how to select the optimal number of clusters in a radial basis function network based on the K-means clustering algorithm. These schemes are among the most popular for unsupervised learning in practical applications. Note that we do not employ any regularization in th examples.

4.1. Principal Component Analysis

In PCA the objective is to provide a simplified data description by projection of the data vector onto the eigendirections corresponding to the largest eigenvalues of the covariance matrix [5]. This scheme is well-suited for high-dimensional, highly correlated data, as, e.g., found in explorative analysis of brain scan volumes [8]. Simple neural network architectures have been suggested that are able to recursively estimate subsets of the principal components, see e.g., [4, 11].

However, the selection of the optimal number of PCs is a largely unsolved problem, although many statistical tests and heuristics have been proposed [5]. Here we suggest to use the estimated generalization error to select the number, in close analogy with the approach of [12] for optimization of feed-forward nets in a supervised learning context.

To proceed we need to specify PCA in terms of a cost function. In particular we assume that the data vector x (of dimension L) can be modelled as a Gaussian distributed multivariate variable whose main variation is confined to a subspace of dimension K_0, this component being degraded by additive, independent isotropic noise, $x = s + n$ where the "signal" $s \sim \mathcal{N}(x_0, \Sigma_s)$, while the "noise" is distributed $n \sim \mathcal{N}(0, \Sigma_n)$. We assume that Σ_s is singular, i.e., of rank $K_0 < L$, while $\Sigma_n = \sigma^2 I_L$, where I_L is a $L \times L$ unit matrix and σ^2 is a noise level.

Using well-known properties of Gaussian random variables we find $x \sim \mathcal{N}(x_0, \Sigma_s + \Sigma_n)$. Hence, we can use straightforward maximum likelihood estimation ($R(\theta) \equiv 0$) to learn the parameters $\theta \equiv (x_0, \Sigma_s, \Sigma_n)$. Maximum likelihood estimation – or equivalently minimum of the negative log-likelihood – is precisely of the form we have discussed with $\epsilon(x|\theta) = -\log p(x|\theta)$ where $p(x|\theta)$ is the p.d.f. of the data given the parameter vector. With this form of the cost we can simplify the design relation (9) considerably. Note that by Fisher's argument $Q = J$, hence, Trace $\left[QJ^{-1}\right] = \dim(\theta)$ and

$$\langle G(\widehat{\theta})\rangle_D \approx \langle E(\widehat{\theta})\rangle_D + \frac{\dim(\theta)}{N}. \tag{10}$$

where the dimensionality of the parametrization of course depends on the number, say $K \in [1; L]$, of PCs retained in the PCA context. Since we also estimate the mean value vector x_0 and the noise variance σ^2 the total number of estimated parameters can be shown to yield $\dim(\theta) = L + 1 + K(2L - K + 1)/2$.

Assuming the examples to be drawn independently, we obtain the maximum likelihood estimate as:

$$\widehat{x}_0 = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha, \quad \widehat{\Sigma} = \frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha - \widehat{x}_0)(x_\alpha - \widehat{x}_0)^\top. \tag{11}$$

Assume $\widehat{\Sigma} = S\Lambda S^\top$ where S is the matrix of eigenvectors and $\Lambda = \text{diag}([\lambda_1, \cdots, \lambda_L])$ is the diagonal matrix of eigenvalues λ_i ranked in decreasing order. By fixing the dimensionality of the signal subspace, K, we identify the covariance matrix of the subspace spanned by the K largest PCs by $\widehat{\Sigma}_K = S \cdot \mathrm{diag}([\lambda_1, \cdots, \lambda_K, 0, \cdots, 0]) \cdot S^\top$. The noise level is subsequently estimated as $\widehat{\sigma}^2 = 1/(L-K)\mathrm{Trace}[\widehat{\Sigma} - \widehat{\Sigma}_K]$, hence $\widehat{\Sigma}_n = \widehat{\sigma}^2 I_L$ and $\widehat{\Sigma}_s = S \cdot \mathrm{diag}([\lambda_1 - \widehat{\sigma}^2 \cdots, \lambda_K - \widehat{\sigma}^2, 0 \cdots, 0]) \cdot S^\top$. For each value of K we then estimate the generalization error and commend the value that provides the minimal error.

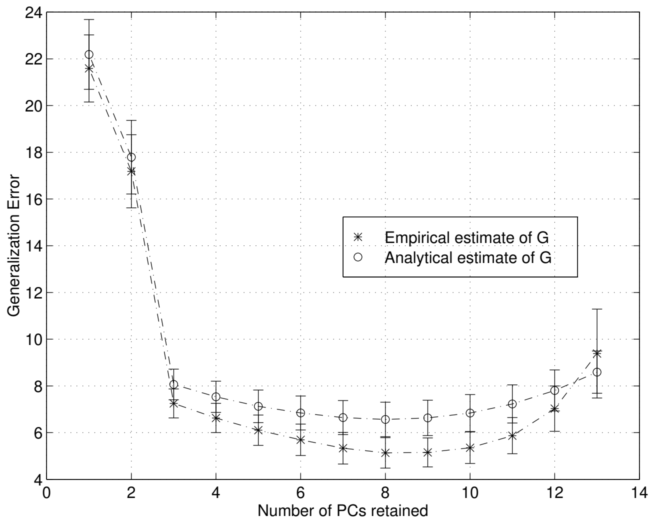

Fig. 1: Numerical experiment illustrating the selection of signal space dimension (number of Principal Components retained in PCA). Artificial multivariate ($L = 20$) Gaussian data was created with a covariance matrix that was composed from two components: one component is singular (rank $K_0 = 3$) and the second component being isotropic (white) noise. The signal to noise ratio was set to 10. PCAs were carried out retaining an increasing number of PCs. The noise subspace was modelled by a covariance structure proportional to the unit matrix ($L \times L$). The analytical estimate of the generalization error is compared to an empirical estimate based on a test sets of size $N_{\mathrm{test}} = 120$. Error bars are estimated by the standard deviation within 10 replications of the experiment.

To illustrate the procedure we arrange an experiment based on $L = 20$ dimensional data vectors. The signal subspace dimensionality is set to $K_0 = 3$. In the experiment the training set comprises $N = 42$ data vectors and for comparison we compute the empirical test error on an independent test set of size $N_{\mathrm{test}} = 120$. In figure 1 the result of the evaluation is shown as function of K. As seen, not only do the analytical theory predict correctly the optimal value of K, but in fact, it is also able to provide a reasonable estimate of the numerical values of the negative log-likelihood. It is worth noting that the optimal generalization is obtained by modeling the "signal" space as well as parts of the noise subspace. This is expected since our model can capture both the structure signal space ($K_0 = 3$) and the noise subspace. The optimal number of PCs for the given example is about $K = 9$.

4.2. K-means Clustering

Our second demonstration concerns estimation of cluster centers and widths for use in radial basis function based density estimation. There are a large number of clustering based neural net algorithms in the literature see e.g., [2, 6, 9, 13] and optimization of the architecture, as exemplified by selection of the optimal number of basis functions (clusters) is a largely unsolved problem [6]. Many heuristics have been suggested [2, 13], e.g., formulated in terms of *prior* complexity penalties. [14] have proposed the AIC criterion for cluster selection. The model we consider here is a Gaussian mixture

$$p(x) = \sum_{k=1}^{K} \beta_k p_k(x). \tag{12}$$

where β_k is the probability of cluster k and $\sum_k \beta_k = 1$. and $p_k(x) = p(x|x \in \text{cluster } k)$ is the conditional p.d.f. (basis function). As before, the data vector is L dimensional, and we assume the basis function to be isotropic and Gaussian:

$$p_k(x) = \left(2\pi\sigma_k^2\right)^{-L/2} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right). \tag{13}$$

The cluster centers are denoted μ_k, while the widths are denoted σ_k^2. We further constrain the widths to be identical $\sigma_k^2 = \sigma^2$. Maximum likelihood estimation is used to estimate the parameters on the finite training set. This allows us to employ the theory developed, with the negative log-likelihood cost function, $\epsilon(x|\theta) = -\log p(x|\theta)$, as in the previous example. The parametrization comprises: $\theta = (\mu_1, \cdots, \mu_K, \sigma^2, \beta_1, \cdots \beta_{K-1})$. In this case $\dim(\theta) = K(L+1)$ real parameters[2].

Generalization theory is used in this example to select the number of clusters K (radial basis functions). To determine the cluster centers we employ a simple *K-means* procedure[3], while the radial basis function widths are determined by maximum likelihood assuming simple isotropic Gaussian basis functions. The relative probabilities of the clusters, β_k, are determined simply as the relative frequencies observed on the training set. K-means is, in fact, not an exact maximum likelihood procedure for center selection. However, it is a computationally attractive and close approximation for well defined clusters.

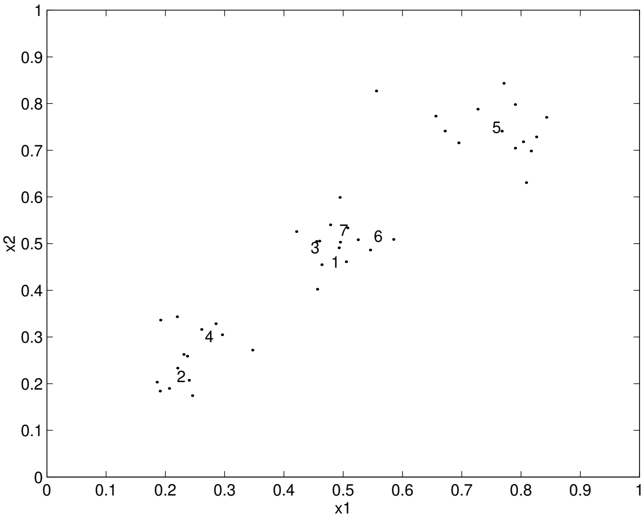

Fig. 2: Numerical demonstration of clustering by K-means. The data are generated by three ($K_0 = 3$) cluster centers with additive isotropic Gaussian distributed noise. The resulting cluster centers for $K = 7$ are indicated by numerals.

As numerical test bed we construct 2D ($L = 2$) data based on three clusters as depicted in figure 2. We subsequently draw at random a training set consisting of $N = 42$ points. K-means is used to find centers for $K = 1, \cdots, 7$. The resulting cluster centers are indicated in figure 2, for $K = 7$.

Like in the previous demonstration we estimate the generalization error by the analytical expressions and by means of a test set, as shown in figure 3. In line with the previous example we conclude that it is possible to estimate the form of the generalization error analytically. Using either the empirical or the analytical estimate we conclude that the proposed generalization error can indeed be used to select the optimal cluster number.

5. Conclusion

This paper introduced the concept of generalization in an unsupervised context. The objective of unsupervised learning is to identify structure in the probability distribution of a data vector. Formulating the unsupervised learning problem as a minimization task, by parameterizing the probability density of the data

[2]Note that $\beta_K = 1 - \sum_{k \neq K} \beta_k$.

[3]The K-means procedure is an iterative assignment and averaging procedure. First a fixed number of cluster centers, K, are initialized at random positions. All points in the training set are assigned to a specific cluster center by proximity in the Euclidian metric. The cluster centers are subsequently moved to the center of mass of it's member data points.

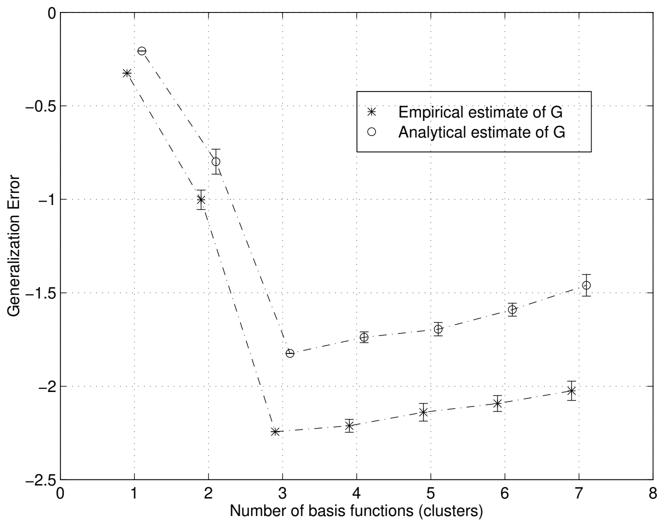

Fig. 3: Numerical experiment illustrating the selection of the number of clusters in a radial basis function network estimation the density of the 2D data vectors of figure 2. The analytical estimate of the generalization error is compared to an empirical estimate based on a test sets of size $N_{\mathrm{test}} = 120$. Error bars are estimated by the standard deviation within 10 replications of the experiment.

vector, enables us to define an associated generalization error. The generalization error is used to determine the correct dimensionality of the parameterization, e.g., the number of principal components in principal component analysis or the number of clusters in K-means clustering.

The suggested method was successfully applied to determine the number of principal components and the number of clusters.

6. Acknowledgments

LKH thanks the participants of the 1995 Telluride Neural Net Workshop for discussions of clustering that initiated the work on unsupervised generalization. We thank Ulrik Kjems, Nick Lange, Benny Lautrup, Niels Mørck, Steve Strother, and Claus Svarer for valuable discussions related to the present work.

This research is supported by Christian and Ottilia Brorsons Rejselegat, NIH Grant DA09246, and the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center. Furthermore, JL acknowledge the Radio Parts Foundation for financial support.

References

[1] H. Akaike: "Fitting Autoregressive Models for Prediction," *Ann. Inst. Stat. Mat.*, vol. 21, pp. 243–247, 1969.

[2] J. Buhmann and H. Kuhnel: "Complexity optimized Data Clustering by Competitive Neural Networks," *Neural Computation*, vol. 5, pp. 75–88, 1993 (note reprinted in later volume due to misprints).

[3] R.O. Duda and P.E. Hart: *Pattern Classification and Scene Analysis* New York: Wiley-Interscience, , 1973.

[4] J. Hertz A. Krogh & R.G. Palmer: *Introduction to the Theory of Neural Computation*, Reedwood City CA: Addison-Wesley, 1991.

[5] J.E. Jackson: *A User's Guide to Principal Components*, Wiley Series on Probability and Statistics, New York: John Wiley and Sons, 1991.

[6] A.K. Jain & J. Mao: "Neural Networks and Statistical Pattern Recognition," *1994 IEEE International Conference on Neural Networks Special Symposion on Computational Intelligence: Imitating Life*, IEEE Service Center, NJ, 1994.

[7] J. Larsen: *Design of Neural Network Filters*, Ph.D. Thesis, Electronics Institute, The Technical University of Denmark, March 1993. Available from ftp://ei.dtu.dk/dist/PhD_thesis/jlarsen.thesis.ps.Z.

[8] J.R. Moeller, S.C. Strother, J.J. Sidtis, and D.A. Rottenberg "Scaled Subprofile Model: A Statistical Approach to the Analysis of Functional Patterns in Positron Emission Tomographic Data," *J. Cereb. Blood Flow Metab.*, vol. 7, pp. 649–658, 1987.

[9] J. Moody & C.J. Darken: "Fast Learning in Networks of Locally-Tuned Processing Units," *Neural Computation*, vol. 1, pp. 281–294, 1989.

[10] N. Murata, S. Yoshizawaand & S. Amari: "Network Information Criterion — Determining the Number of Hidden Units for an Artificial Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 865–872, 1994.

[11] E. Oja: "Neural Networks, Principal Components, and Subspaces," *International Journal of Neural Systems*, vol. 1, pp. 61–68, 1989.

[12] C. Svarer, L.K. Hansen, and J. Larsen: "On Design and Evaluation of Tapped-Delay Neural Network Architectures," in H.R. Berenji *et al.* (eds.) *Proceedings of the 1993 IEEE Int. Conference on Neural Networks*, IEEE Service Center, NJ, vol. 1, pp. 46–51, 1993.

[13] Y. Wong : "Clustering Data by Melting," *Neural Computation*, vol. 5, pp. 89–104, 1993.

[14] J. Zhang & J.W. Modestino: "A Model Fitting Approach to Cluster Validation With Application to Stochastic Model-Based Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 1009–1017, 1990.