Persistent Issues in Learning and Estimation

George Nagy

ELECTRICAL, COMPUTER, AND SYSTEMS ENGINEERING
Rensselaer Polytechnic Institute
Troy, NY 12180-3590, USA
tel.: 518 276-6078
fax.: 518 276-6261
nagy@ecse.rpi.edu

Keywords:

Parameter estimation; Curse of dimensionality; Adaptive classification, Hybrid learning; Labeled and unlabeled samples; Mixture distributions; Correlated features, Expectation Maximization.

Submitted for consideration for ICPR'98, Pattern Recognition and Analysis stream

Persistent Issues in Learning and Estimation

This is a review, from an intuitive rather than a mathematical perspective, of the statistical foundations of adaptive recognition systems. Key considerations in adaptive classification are priors, sample size and sampling strategy, labels, statistical dependencies, and dimensionality. The small-sample bias and variance of maximum likelihood, maximum *a posteriori* and Bayes estimators are compared in a small concrete case. Iterative expectation maximization for estimating the sufficient statistics of mixtures is illustrated in a simple setting. It is shown that correlation among features is sometimes unjustly maligned. A counterintuitive increase in the error rate after adding a second feature is traced to the curse of dimensionality. Adaptive classification is presented in the context of both parametric and non-parametric (nearest neighbors and neural nets) estimation. Some recent theoretical results and not-so-recent experimental observations on hybrid classification (based on both labeled and unlabeled samples) are summarized.

1. Introduction

Adaptive, "unsupervised" classification based partly on unlabeled samples has a long and respectable history in both pattern recognition and statistics, but in some quarters it is still viewed with suspicion. Perhaps the over-ambitious term "unsupervised learning" is responsible for some of the mistrust. Nevertheless, static recognition systems have already been widely explored and success in many new applications will require adaptive techniques that make aggressive use of unlabeled samples. Although there is a large and valuable literature on the theoretical justification of these methods, many of the relevant articles require a degree of statistical sophistication beyond the reach of most practitioners.

The goal of this presentation is to shed light, through examples, on several complex, interacting phenomena that underlie adaptive classification. The examples are kept simple (and therefore artificial) to avoid obscuring the key concepts. They are drawn from the application domains most familiar to the author: optical character recognition, speech recognition, and remote sensing. Other applications exist in biometrics, computer vision, and non-speech time-signals.

The first step towards an engineering solution based on available methodology is an assessment of the *systemic* problem characteristics. Some important features for classifying specific pattern recognition applications from a statistical perspective are the following.

A priori probabilities of the classes of interest (and of "noise" patterns).
Is the number of classes fixed and known, or does it vary from test to test? Are the class populations commensurable (as in the case of digits), or disparate (e.g., words in text or speech)? Are they already known (like letter frequencies) or must they be estimated?

The sampling strategy used to collect training, validation, and test samples.
Is it realistic to consider a reserved subset of a single sample as a test set representative of field conditions? What should be the granularity of the partition? Will some tuning be needed in the field? Are the underlying probability distributions static, or is it desirable to track drifting or cyclic population parameters?

Labels. What is the cost of collecting labeled samples versus that of unlabeled samples? How reliable is the initial labeling process? Is it possible to label every sample accurately, or are there many that must be relegated to some catch-all category (noise, impostor, background, outlier, stammer, blot)? How specific are the labels: are font, writer, or speaker identities known? Can the classifier be used as a clustering tool to facilitate the collection of labeled samples? Is it possible to obtain additional labeled samples under field conditions from operator correction of classifier errors?

Statistical dependence. Are samples statistically independent, or can information be extracted from one sample to help classify another? If there is dependence, is it at the symbolic level, i.e., dependence between the labels, as in language (morphological, lexical, syntactic, semantic) context? Or is it at the level of the observations themselves, as in font-context in printed matter, distinctive styles in writing, and the consistency of an individual's tempo, pitch, prosody, intonation and pronunciation in speech? Is statistical dependence between samples engendered by external factors such as telephone line quality in speech, and scanner characteristics in OCR? Aside from between-sample correlations, do the features themselves exhibit class-conditional dependence?

Dimensionality. The designer of a classification system usually has some latitude in determining the number of features used for classification. Even if additional features are hard to come by, it is always possible to discard some of the available features. Are there too few features or are there too many relative to the sample size?

In the remaining sections, we illustrate some of the statistical consequences engendered by the answers to the above queries.

2. Parameter Estimation

We first illustrate the difference between popular parameter estimators by estimating from a very few (3) samples the probability q that a given pixel in a scanned character is Black (B) or White (W). Suppose that in our three training samples, one particular pixel is Black in two of the samples, and White in the other. The probability of m successes in n tries (the outcome of the experiment is called X) is given by the binomial formula:

$$P(C=2B,1W \mid q)= 3q^2(1-q).$$

The maximum likelihood estimator $\boldsymbol{q}_{ML}$ of q is the value of q that maximizes this probability, $argmax_q\ p(C|q)$. Setting the derivative of the expression equal to zero to find the maximum,

$$\boldsymbol{q}_{ML} = 2/3 = 0.667.$$

In the general case of m successes in n trials, the formula is $q_{ML} = m/n$. Below, we shall compare some properties of the ML estimator with that of others. Assume that we know from experience that the probability that any pixel is Black varies according to the following probability density:

$$p(q) = 6q(1-q) \qquad [0 < q \text{ £ } 1]$$

This is called the *a priori probability,* or prior. We see that this prior is symmetric about the value $q = 0.5$. Now, using Bayes' formula, we calculate the *posterior probability*:

$$p(q|C=2B,1W) = 60\ q^3(1-q)^2.$$

The *maximum a posteriori* (MAP) estimator $\boldsymbol{q}_{MAP}$ is $\text{argmax}_\theta\ p(\theta|\chi)$:

$$\boldsymbol{q}_{MAP} =(m+1)/(n+2)= 3/5 = 0.600,$$

and the Bayes estimator $\boldsymbol{q}_B$ is the expectation of the posterior density, $E_{p(q|C)}[q]$ (or $E[q|\chi]$):

$$\boldsymbol{q}_B = (m+2)/(n+4)= 4/7 = 0.571$$

We have obtained three significantly different estimates, *0.667, 0.600, 0.571*, for the probability that the pixel is Black. Which is best? Although we cannot give a definitive answer to this question, it motivates us to investigate further some of the properties of these estimates.

Each of the estimators is a random variable that takes on a specific value according to the outcome of a (random) experiment. It is customary to characterize the probability distribution of estimators by their means and variances. Since all of our estimators are linear functions of m, their means and variances can be computed from the first and second moments of the binomial distribution that governs m: $E[m] = nq$, and $E[m^2] = nq(1-q) + n^2q^2$.

The *bias* of an estimator is defined as $(E[\underline{q}] - q)^2$. Therefore BIAS$[\underline{q}_{ML}] = (q - q)^2 = 0$, and BIAS$[\underline{q}_B] = ((nq+2)/(n+4)-q)^2 = ((2-4q)/(n+4))^2$. So with three samples and $q=0.25$, the expected value of $\underline{q}_B$ is 0.37, while that of $\underline{q}_{ML}$ would be the correct 0.25.

However, we should also consider the variance, which indicates how much the estimate will fluctuate from sample to sample. Using again the moments of the binomial distribution,

$$\mathrm{VAR}[q_{ML}] = \mathrm{E}[q_{ML}{}^2] - (\mathrm{E}[q_{ML}])^2 = 1/n\ q(1-q), \quad \text{while } \mathrm{VAR}[q_B] = n/(n+4)^2\ q(1-q)$$

We see that for a small number of samples, the variance of q_{ML} is much larger than that of q_B. The relevance of our findings for the design of practical recognition systems is the following.

The ML estimator is unbiased, but its higher variance fully reflects sample variability. The Bayes (and also the MAP) estimator is biased, but has lower variance because it averages prior and posterior distributions. With more features, the bias and variance inevitably increase. For adaptation, where we have good prior estimates, it makes sense to use Bayes (or MAP) estimators. *The variances of the priors determine the weighting of the new vs. old samples.*

3. The curse of dimensionality

Perhaps countertuitively, additional features may actually *increase* the error rate. This phenomenon is named after G.F. Hughes, who contributed the first comprehensive analysis. The root cause is that the parameters are estimated with too few samples, therefore the decision rule is suboptimal. We illustrate this by classifying a scanned character, S_0, into Class ω_1 or Class ω_2 using features consisting of either one pixel (x) or two independent pixels (x_1, x_2). The prior probabilities are known: $P(\omega_1) = P(\omega_2) = 0.5$. For simplicity, let the pixel probabilities be symmetric: if $P(x_i=B|\omega_1) = p_i$, then $P(x_i=B|\omega_2) = 1-p_i$.

With m binary features and k training samples, there are $(2^m)^k$ conceivable configurations of the training set. Given the class-conditioned pixel-value probabilities, we can compute, in principle, (i) the probability of each configuration, (ii) the resulting probability of misclassification, and hence (iii) by aggregation, the expected error rate. The expected error rate is shown for a few cases.

Conditional feature probabilities	P*	P^1	P^2

$P(x = B\|\omega_1) = 0.6$	0.40	0.48000	0.4800
$P(x_1=B\|\omega_1) = 0.6$, $P(x_2=B\|\omega_1) = 0.60$	0.40	0.49000	0.4800
$P(x_1=B\|\omega_1) = 0.6$, $P(x_2=B\|\omega_1) = 0.55$	0.40	0.49375	0.4875
$P(x_1=B\|\omega_1) = 0.6$, $P(x_2=B\|\omega_1) = 0.50$	0.40	0.49500	0.4900

It is instructive to compare the results of the following scenarios: (i) There is only a single training sample selected with equal probability from ω_1 or ω_2. (ii) There are two training samples S_1 and S_2 such that S_1 Î ω_1, S_2 Î ω_2. The appropriate decision rule is to classify S_0 into the same class as the training sample(s) with the same feature values. If there is ambiguity (for instance S_1 and S_2 both have (x_1, x_2) = (W,W), or S_1 = (B,B) and S_2 = (W,W) but S_0= (W,B)), then S_0 is arbitrarily labeled ω_1. The Bayes error, P*, which is the minimum error rate achievable when the priors and the feature probabilities are estimated perfectly, is 0.40 in each case. But the error P^k based on k training samples increases when a second feature is added, even if this feature contributes useful information. The additional information is negated by additional "noise" in the estimates. This effect is most pronounced for $P(x_2=B|\omega_1) = P(x_2=B|\omega_2) = 0.5$, where the second feature contributes no useful information. The error increases less when the training set is larger.

4. Mixture Populations

When the feature distribution is multimodal, or when the distribution of the test set is different from that of the training set and must be estimated using *unlabeled samples*, mixture estimation techniques come into play. *Clustering* methods partition a set of samples into mutually exclusive categories, while methods based on maximum likelihood can cope with overlapping densities. We illustrate the iterative formulas for estimating the parameters in the popular Expectation Maximization (EM) formulation. We note, however, that these formulas are identical to the classical mixture estimators as presented in Duda and Hart.

Consider a two-component scalar Gaussian mixture. One of the components is selected according to the mixing parameters, then a sample is drawn from the component. The mixing parameters P_1, P_2 are known (½,½); the variances s^2 are equal and known, but the means are unknown. The two samples (*incomplete data*) are: X: $\{x_1=0.2, x_2=0.7\}$. The key to EM is to postulate *hidden variables* $\mathbf{z_1} = (z_{11}, z_{12})$, $\mathbf{z_2} = (z_{21}, z_{22})$. The value of $\mathbf{z_i}$ is (0,1) or (1,0), depending on source of sample x_i, but we shall estimate its *expectation* conditioned on the *complete data*, {(x_1, $\mathbf{z_1}$), (x_2, $\mathbf{z_2}$)}, using our current best estimate of the parameters (*E-Step*). Once we have an estimate of the hidden variables, we can find the maximum likelihood estimates of the unknown parameters, which here are only *m_1 and m_2* (*M-Step*). If the initial estimates of the means are $m_1^0 = 0.0$, $m_2^0 = 1.0$, the first two steps are:

E-step: Determine E[$\mathbf{z_1^0}$, $\mathbf{z_2^0}$] (z_{ij} is the probability that sample *i* was from component *j*):

$$E[\mathbf{z}_1^0] = E[z_{11}^0, z_{12}^0] = p(z_{11}^0|x_1),\ p(z_{12}^0|x_1)$$

$$= \frac{ce^{-(0.2-m_1^0)^2/2s^2}}{ce^{-(0.2-m_1^0)^2/2s^2} + ce^{-(0.2-m_2^0)^2/2s^2}}, \frac{ce^{-(0.2-m_2^0)^2/2s^2}}{ce^{-(0.2-m_1^0)^2/2s^2} + ce^{-(0.2-m_2^0)^2/2s^2}}$$

$$= \frac{0.8187}{0.8187 + 0.0408}, \frac{0.0408}{0.8187 + 0.0408} = 0.9525, 0.0475 \quad (s^2 = 0.10, \quad c = \frac{0.5}{\sqrt{2p \times 0.10}})$$

$$E[\mathbf{z}_2^0] = E[z_{21}^0, z_{22}^0] = \frac{0.0863}{0.0863 + 0.6376}, \frac{0.6376}{0.0863 + 0.6376} = 0.1192, 0.8808$$

M-step: Determine m_1, m_2 given $\mathbf{z_1}$, $\mathbf{z_2}$**:**

$$m_1^1 = \frac{z_{11}^0}{z_{11}^0 + z_{21}^0} x_1 + \frac{z_{21}^0}{z_{11}^0 + z_{21}^0} x_2 = 0.8888 \times 0.2 + 0.1112 \times 0.7 = 0.2586,$$

$$m_2^1 = \frac{z_{12}^0}{z_{12}^0 + z_{22}^0} x_1 + \frac{z_{22}^0}{z_{12}^0 + z_{22}^0} x_2 = 0.0511 \times 0.2 + 0.9489 \times 0.7 = 0.6744$$

The estimates converge to $m_1 = 0.2$, $m_2 = 0.7$ for small values of σ^2 (which make large deviations from the mean unlikely), and to $m_1 = m_2 = 0.45$ for large σ^2.

5. Non-parametric classification

Trainable neural networks are easy to modify for decision-based reinforcement, where the output of the classifier produces labeled samples for training. The jury is still out on whether it is better to adapt the weights on each sample, or to average them first.

A validation set, which is often used in Neural Network classification to prevent over-training, is necessary for *all* adaptive classifiers to avoid the possibility that some anomalous set of samples corrupts the classifier. If this happens, the results on the validation set will catch it, and the previous state of the classifier can be simply restored. Single Nearest Neighbor classification cannot be used directly, because a mislabeled reference pattern will simply surround itself with new mislabeled samples. Instead, we use K Nearest Neighbors and deport minority voters. We can tag the contribution of each reference sample, then omit long-unused references to track changes in the population statistics. (Prune and preprocess for speed.)

Nonparametric techniques are easy to program and hard to analyze. Small-sample and dimensionality problems are *hidden*.

6. Correlated features

Are statistically independent features always best? Not necessarily. Suppose that $P(\omega_1) = \frac{1}{2}$, $P(\omega_2) = \frac{1}{2}$, and x_1 , x_2 are binary pixels such that

$$P(x_1=B|\omega_1) = P(x_2=B|\omega_1) = P(x_1= B|\omega_2) = P(x_2=B|\omega_2) = \frac{1}{2},$$

$$P(x_1=B, x_2=B|\omega_1)=P(x_1=W, x_2=W|\omega_1) = \frac{1}{2}, \text{ and } P(x_1=B, x_2=W|\omega_2) = P(x_1=W, x_2=B|\omega_2) = \frac{1}{2}$$

The Bayes error P* is 0. Correlation actually *helps* if it is different for each class.

7. Partially supervised classification

The value of unlabeled samples for obtaining better estimates of mixture densities was recently explored by Vittorio Castelli and Thomas Cover. Their conclusion is that labeled samples are exponentially more valuable than unlabeled samples, because only labeled samples can reveal which mixture component belongs to which class. Once that has been accomplished, however, unlabeled samples are just as useful for characterizing the densities and the mixture coefficients.

Among the earliest demonstrations (in 1959) of the power of adaptive classification was Bernard Gold's experiment on recognizing hand-sent Morse signals. He kept a running total of the observed lengths of three classes of spaces and two classes of marks and showed a significant gain through adapting the classifier parameters to each operator.

In 1966, Robert Lucky used adaptive transversal filters for precise adjustment of tap gain settings in transversal filters in telephone line equalization for digital transmission. He showed that the filters recover their optimal setting even with a pulse misclassification rate of 10%, and that adaptive equalization lowered the raw symbol error rate by a factor of 10.

At the same time, George Nagy and Glen Shelton obtained a fivefold decrease in printed character classification with an omnifont classifier whose parameters were averaged over a set of unlabeled characters from the same font. Baird and Nagy replicated the experiment 25 years later with 100 different fonts.

Behzad Shahshani and David Landgrebe (1994) combined unlabeled samples with a small labeled sample to improve crop classification rates. They demonstrated that EM estimation based on unlabeled samples extends the number of features before Hughes deterioration set in.

The use of unlabeled samples is currently gaining popularity in speech recognition, where large vocabularies and regional, individual and line characteristics induce high error rates and representative labeled samples are difficult to obtain.

8. Summary

Both parametric and nonparametric classifiers require estimating the characteristics of the data from a training set. The single most important set of parameters are often the priors. Even if the training set is representative, finite sample size introduces bias and variance, especially when many parameters must be estimated. Multimodal and composite (Hidden Markov) distributions require mixture estimation techniques, for which Expectation Maximization provides a sound but not infallible basis. Unlabeled samples can improve the estimates.

Correlation among features can help or hinder. Although independence assumptions are seldom justified, they are preferable to biased or high-variance estimates of second-order parameters from small samples.

There is no such thing as completely unsupervised adaptation or learning. But the effective sample size can be increased by taking advantage of unlabeled samples. When the training set is not representative, one may use adaptive methods that exploit mostly-correct classification. Nevertheless, labeled samples are valuable: endeavor to obtain more. Don't waste rejects: they contain useful information about the decision boundaries. Above all, don't ever let the machine rest. After the day of work is done, make sure it assimilates everything that it has seen during the day, including operator corrections, to improve its performance for the morrow.

9. Bibliography

Baird, H.S., Nagy, G., A self-correcting 100-font classifier, *Proc. SPIE Conference on Document Recognition, Volume SPIE-2181*, pp. 106-115, San Jose, CA, 1994.

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

Brunk, H.D., An Introduction to Mathematical Statistics, Ginn & Co., Boston, 1960.

Castelli, V., Cover, T.M., On the exponential value of labeled samples, *PRL 16*, pp. 105-111, 1995.

Castelli, V., Cover, T.M., The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter, *IEEE-IT 42*, 6, pp. 2101-2117, 1996.

Chittenini, C.B., Some approaches to optimal cluster labeling with applications to remote sensing, *Pattern Recognition 15*, 3, pp. 201-216, 1982.

DeGroot, M., Optimal Statistical Decision, McGraw-Hill 1970 (courtesy of D. Thomas).

Dempster, A.P., Laird, N.M., Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm, J. *Royal Statistical Society Series B, 39*, pp. 1-38, 1977.

Duda, R., Hart, P., Pattern Recognition, Wiley, 1973.

Geman, S., Bienstock, E., Doursat, R., Neural networks and the bias/variance dilemma, *Neural Computation 4*, 1, pp. 1-58, 1992.

Gold, B., Machine recognition of hand-sent Morse code, *IRE-IT 5*, pp. 17-24, March 1989.

Lucky, R.W., Automatic equalization for digital communication, *BSTJ XLIV*, 4, pp. 547-588, 1965: see also Techniques for adaptive equalization of digital communication systems, *BSTJ XLV*, 2, pp. 255-286, 1966.

Nagy, G., Shelton, G.L., Self-corrective character recognition system, *IEEE-IT 12*, 3, pp. 215-222, 1966.

Raudys, S.J., Jain, A.K., Small sample effects in statistical pattern recognition: Recommendations for practitioners, *IEEE-PAMI 13*, 3, pp. 252-263, 1991.

Redner, R.A., Walker, H.F., Mixture densities, maximum likelihood, and the EM algorithm, *SIAM Review 26*, 2, pp. 195-235, 1984.

Shahshani B.S., Landgrebe, D.A., *IEEE-Trans. Geoscience & Remote Sensing 32*, 5, pp. 1087-1095, 1994.