Learning and Adaptation

George Nagy, Rensselaer Polytechnic Institute
Troy, NY

This is a review, from an intuitive rather than a mathematical perspective, of the statistical foundations of adaptive recognition systems. Key considerations for laying siege to adaptive classification are priors, sample size and sampling strategy, labels, statistical dependences, and dimensionality. The small-sample bias and variance of maximum likelihood, maximum a posteriori and Bayes estimators are calculated for small concrete cases. Conventional methods are contrasted with expectation maximization for estimating the sufficient statistics of mixtures in a simple setting. It is argued that correlation among features is sometimes unjustly maligned. A counterintuitive increase in the error rate when a second feature is added 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. The role of rejects in adaptive classification is explored.

Learning and Adaptation

George Nagy

The five most important features for classifying classification problems:
priors, sample size, labels, independence, dimensionality

Supervised training

- Parameter estimation
  - (bias, variance. sufficiency, consistency)
  - Unimodal distributions (Binomial, Normal)
    - ML
    - MAP
    - Bayes
  - Mixture distributions (finite and exponential)
    - Clustering
    - Old methods
    - EM
- Nonparametric classification
  - NN or NN?

Unsupervised learning

- Tracking: Maude, Robert Lucky's adaptive equalizer
- Nearest Neighbors - dynamic pruning
- A self-correcting classifier (1966, 1992)

Hybrid training

- The exponential value of supervised samples (Cover).
- Unsupervised samples cure Hughes' Disease (Landgrebe).
- Chitti Babu strikes again.

Context (move to OCR talk)

- Language
  - Clustering and language context: cipher substitution
- Style
- Pragma

Aphorisms:

- Rejects are where the boundary is.
- Don't waste time on easy discriminations.
- Take advantage of every slap on the wrist.
- If you don't have a small-sample estimation problem,
  - you paid for too many samples.

Don't let the machine sleep.

Candide's Guide, Baird & Nagy, Autonomous

Learning and Adaptation

George Nagy
Rensselaer Polytechnic Institute

Five features for classifying classification problems

Supervised training

- Parameter estimation
  - Properties of ML, MAP & Bayes estimators
  - Mixture estimation, conventional & EM
  - The curse of dimensionality
  - Correlated features
- Nonparametric classification
  - NN or NN?

"Unsupervised" learning

Hybrid training and tracking

- The value of labeled samples
- Experimental observations
- Rejects

Aphorisms

Five features for classifying classification problems

Priors: known or must be estimated?
uniform or disparate?
number of classes fixed?

Sampling: labeled, unlabeled, distribution?
training set representative?
training set static or dynamic?
granularity?

Labels: cost, reliability, specificity, completeness?
machine aided?
from operator corrections?

Independence: symbols (context),
samples (source, sensor),
features?

Dimensionality: too few or too many?

THREE COINS IN THE FOUNTAIN

Prior probability: $P(\theta) = 6\theta(1-\theta)$ $[0<\theta\leq 1]$

Outcome X: **2 Heads, 1 Tail**

$$P(\mathrm{X}=2H,1T \mid \theta)= 3\theta^2(1-\theta)$$

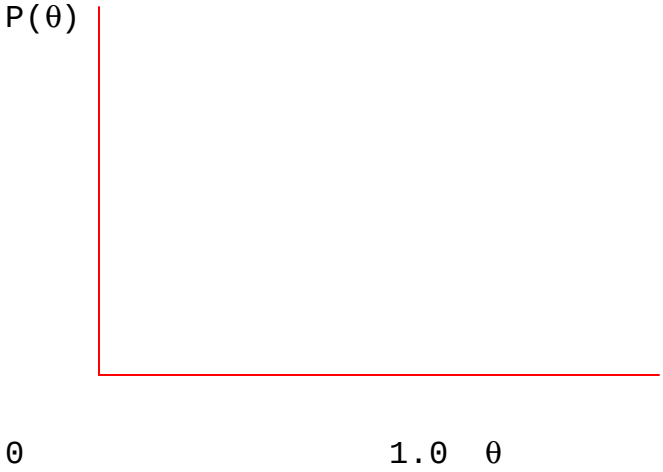
$\hat{\theta}_{ML} = m/n = 2/3 = 0.667$ $(\mathrm{argmax}_\theta\ p(\mathrm{X}|\theta))$

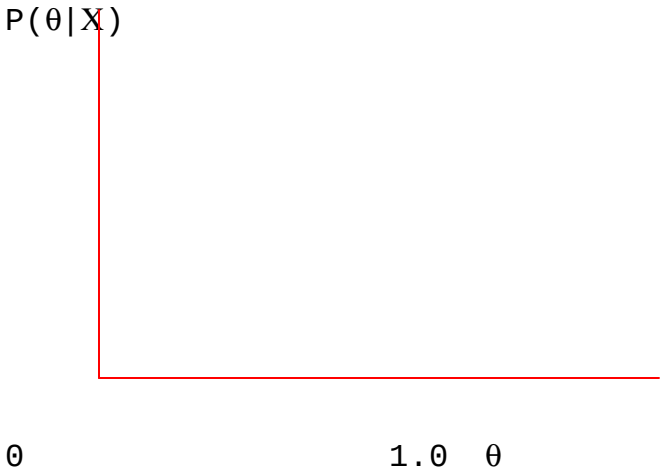
Posterior probability: $p(\theta|\mathrm{X}) = 60\ \theta^3(1-\theta)^2$

$\hat{\theta}_{MAP} =(m+1)/(n+2)= 3/5 = 0.600$ $(\mathrm{argmax}_\theta\ p(\theta|\mathrm{X}))$

$\hat{\theta}_{B} =(m+2)/(n+4)= 4/7 = 0.571$ $(E_{p(\theta|\mathrm{X})}[\theta])$

$$\int_0^1 x^{m-1}(1-x)^{n-1}\,dx = \Gamma(m)\Gamma(n)/\Gamma(m+n), \qquad \Gamma(n+1) = n!$$


P(θ)
0
1.0
θ


P(θ|X)
0
1.0
θ

The Variance of $\underline{\theta}$

m is binomial with parameter θ:

$$E[m] = n\theta \qquad E[m^2] = n\theta(1-\theta) + n^2\theta^2$$

$$E[\underline{\theta}_{ML}] = E[m/n] = 1/n\ E[m] = \theta$$

$$E[\underline{\theta}_{ML}^2] = E[(m/n)^2] = 1/n^2\ E[m^2] = 1/n\ \theta(1-\theta) + \theta^2$$

$$VAR[\underline{\theta}_{ML}] = E[\underline{\theta}_{ML}^2] - (E[\underline{\theta}_{ML}])^2 = 1/n\ \theta(1-\theta) \qquad *$$

$$E[\underline{\theta}_B] = E[(m+2)/(n+4)] = 1/(n+4)\ (E[m] + 2) = (n\theta+2)/(n+4)$$

$$E[\underline{\theta}_B^2] = E[((m+2)/(n+4))^2] = 1/(n+4)^2\ \{E[m^2] +4E[m] + 4\}$$

$$= 1/(n+4)^2\ \{n(n-1)\theta^2 + 5n\theta + 4\}$$

$$VAR[\underline{\theta}_B] = 1/(n+4)^2\ \{n(n-1)\theta^2 +5n\theta+4-(n\theta+2)^2\}$$

$$= n/(n+4)^2\ \theta(1-\theta) \qquad *$$

The Bias of $\underline{\mathbf{q}}$

$BIAS[\underline{\mathbf{q}}] = (E[\underline{\theta}] - \theta)^2$

$BIAS[\underline{\mathbf{q}}_{ML}] = (\theta - \theta)^2 = 0$

$BIAS[\underline{\mathbf{q}}_{B}] = ((n\theta+2)/(n+4)-\theta)^2 = ((2-4\theta)/(n+4))^2$

$\in [0, (2/(n+4))^2]$

e.g., when n=3, θ=0.25, then bias=$(+0.14)^2$

SUMMARY:

ML estimator: unbiased, higher variance fully reflects sample variability.

Bayes and MAP estimators: biased, lower variance because they average prior and posterior distributions.
(Beware when $P(\underline{\theta}_B)=0$!) **Tighter priors decrease variance but increase bias.**

In adaptation, good priors are usually available, so MAP and Bayes estimators are appropriate. ***The priors determine the weighting of new vs. old samples.***

High dimensionality inevitably increases bias/variance.

Consistency and Efficiency

$\underline{\theta}$ is *consistent* if $P(|\underline{\theta}-\theta| \leq \varepsilon) \rightarrow 1$ as $n \rightarrow \infty$.

Practically all estimators are consistent.

An unbiased estimator is *efficient* if
it has the smallest possible variance.

Is the Bayes estimator efficient?

It is *asymptotically efficient* if its deviation from the parameter approaches the normal distribution quickest.

Sufficient Statistic

A statistic is *sufficient* if it has all the necessary information necessary to estimate the parameter.

Test: **s** is a sufficient statistic for θ
if $P(\mathrm{X}|\mathbf{s})$ can be written in the form $P(\mathrm{X}|\mathbf{s}) = g(s,\theta)\, h(\mathrm{X})$.

for Bernoulli, $h(\mathrm{X})=1$, $s=\Sigma x$:

$$P(\mathrm{X}|\mathbf{s}) = g(s, \theta) = {}^{n}C_{m}\theta^{m}(1-\theta)^{n-m} = {}^{n}C_{m}\,\theta^{m}(1-\theta)^{n-\Sigma x}$$

In our example, number of heads is a sufficient statistic (because we don't need the order).

Mixture Distributions

Clustering Vector Space or Similarity Measure
(SM→VS via multidimensional scaling,
VS→SM via distance function)

Problems:
choice of metric (no individual covariances)
local minima
initialization
necessary constraints:
number of classes or minimum cluster separation
maximum cluster diameter
minimum/maximum cluster population

Maximum Likelihood (identifiable densities)

ML estimation for a mixture differs from that for a single distribution only by the presence of the (unknown) mixture coefficients. The number is components is usually assumed fixed and known.

According to the model, a component is first selected according to the mixture coefficients interpreted as probabilities, then a sample is drawn from the component distribution.

ML by balancing moments of normal densities:

1-D, two components:

$\Sigma x_i = f_1(n_1, n_2, \mu_1, \mu_2)$

$\Sigma x_i^2 = f_2(n_1, n_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$

$\Sigma x_i^3 = f_3(n_1, n_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$

$\Sigma x_i^4 = . . .$

. . .

Five nonlinear equations that require a solution of a **ninth-degree** polynomial!

ML by direct optimization:

Equations for the parameters can be derived by differentiating the log-likelihood wrt each parameter. Need to ensure mixture coefficients that are positive and sum to 1. E.g., Lagrange multipliers: before differentiating, add a term of the form λ(P1 + P2 +.. - 1) .

For 5 features, 10 classes, and 1000 normally distributed samples, we obtain (1+1+15) x 10 = **170 coupled non-linear equations**, each consisting of 1000 ratio terms.

ML by gradient descent:

The equations for the parameters can be rearranged into a form suitable for iterative solution (Duda and Hart, p. 200).

This formulation is **identical to that obtained by Expectation Maximization**.

Expectation Maximization

Example 1. Three partitions with unknown parameter r (simplified from Dempster et al.):

name	x_1	x_2	x_3
fraction	0.5	0.25 r	0.75 r - 0.5
sample population	80		32

Hidden variables: x_1 and x_2.

Sufficient statistic for estimating r: x_1 and x_2.

E-Step: Compute the expected value of the hidden variables given the current parameter(s).

$$x_1^p = 80 \times 0.50 / (0.5 + 0.25\, r^p) \quad (1)$$

$$x_2^p = 80 \times 0.25\, r^p / (0.5 + 0.25\, r^p) \quad (2)$$

M-Step: Find the new parameter(s) from the current value of the hidden variables.

$$X_2^p / (x_2^p + 32) = 0.25 r^{p+1} / (r^{p+1} - 0.5) \text{ or}$$

$$r^{p+1} = 0.5x_2^{p} / (0.75\, x_2^{p} - 8) \qquad (3)$$

Iterative solution:

Initial value of $r = r^0 = 0.7$

E_1: $x_1^0 = 59.25$, $x_2^0 = 20.75$
M_1: $r^1 = 1.37$

E_2: $x_1^1 = 47.50$, $x_2^1 = 32.50$
M_2: $r^2 = 0.992$

E_3: $x_1^2 = 53.47$, $x_2^2 = 26.53$
M_3: $r^3 = 1.115$

Eventually, r^n converges to 1.077

This can also be obtained here in a single step by solving the recursion obtained by combining (2) and (3).

$$r^{p+1} = 10r^p / (13r^p - 4)$$

Example 2. Two-component Gaussian mixture

Model: One of the components is selected according to the mixing parameters, then a sample is drawn from component.

Mixing parameters P_1, P_2 known (½, ½);
variances σ^2 equal and known, means unknown.

Two samples (*incomplete data*): X: $\{x_1 = 0.2, x_2 = 0.7\}$.

Hidden variables: $z_1 = (z_{11}, z_{12})$, $z_2 = (z_{21}, z_{22})$

$z_i = (0,1)$ or $(1,0)$, depending on source of sample x_i.

Sufficient statistic: z_1, z_2

Complete data: $\{(x_1, z_1), (x_2, z_2)\}$

Initial estimate of the means: $\mu_1^0 = 0.0$, $\mu_2^0 = 1.0$

E-step: Determine $E[z_1, z_2]$ given $x_1, x_2, \mu_1^0, \mu_2^0$

$E[z_1, z_2] = E[z_1], E[z_2] = (E[z_{11},], E[z_{12}]);$ $(E[z_{21},], E[z_{22}])$

z_{ij} is the probability that sample *i* was from component *j*

M-step: Determine μ_1, μ_2 given z_1, z_2

First two steps

E-Step:

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

$$E[\mathbf{z}_1^0] = \frac{ce^{-(0.2-m_1^0)^2/2s^2}}{ce^{-(0.2-m_1^0)^2/2s^2} + ce^{-(0.2-m_2^0)^2/2s^2}}, \frac{ce^{-(0.2-m_2^0)^2/2s^2}}{ce^{-(0.2-m_1^0)^2/2s^2} + ce^{-(0.2-m_2^0)^2/2s^2}}$$

$$if\ s^2 = 0.10, \quad c = \frac{0.5}{\sqrt{2p \times 0.10}}$$

$$= \frac{0.8187}{0.8187 + 00408}, \frac{00408}{0.8187 + 00408} = (0.9525, 0.0475)$$

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

M-Step:

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

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

The process converges to $\mu_1 = 0.2$, $\mu_2 = 0.7$ for small σ^2, and to $\mu_1 = \mu_2 = 0.45$ for large σ^2.

Convergence of EM Algorithm

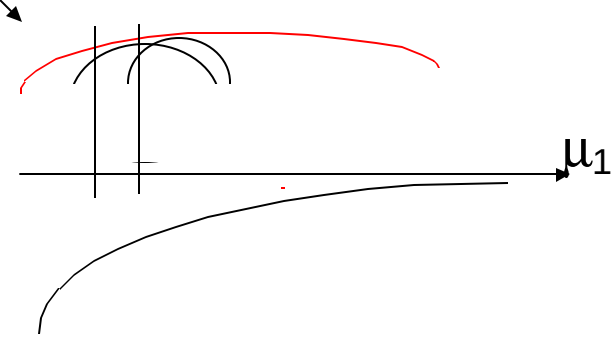
Convergence to a (local) maximum of the Log Likelihood Function is proved by showing that it is *greater* than the log of expectation of the Likelihood of the Complete Data.

(Jensen's Inequality for Convex Functions like the Logarithm:)

$$\log E[Y] \geq E[\log Y]$$

☹

$$\log [z_{11}\, p(x_1,z_{11}),+z_{21}\, p(x_2,z_{21})] \geq z_{11} \log p(x_1,z_{11})+z_{21} \log p(x_2,z_{21})$$



The upper curve is the log likelihood of the Incomplete Data, as a function of μ_1.

The lower curves are the expectations of the Complete Data based on the current mixture coefficients.

Correlated Features

It is widely believed that statistical dependence among the features increases the error rate. Not necessarily:

$P(A) = ½ P(B) = ½$

x_1 , x_2 are binary features, such that

$P(x_1 = 1 \mid A) = P(x_2 = 1 \mid A)$

$= \quad P(x_1 = 1 \mid B) = P(x_2 = 1 \mid B) = ½$

The Bayes error P* = 0 if

$P(x_1 = 1, x_2 = 1 \mid A) = 1, \quad P(x_1 = 0, x_2 = 0 \mid B) = 1$!

Correlation helps if it is different for each class.

The Curse of Dimensionality

In nonparametric classification, sample size must increase *exponentially* with dimensionality.

Even in parameteric classification, the sample size must grow polynomially.

The Hughes Phenomenon

Additional dimensions may *increase* the error rate! The root cause is that the parameters are estimated with too few samples, therefore the decision rule is suboptimal.

A simple example:

Two classes, $P(\omega_1) = P(\omega_2) = 0.5$;

m=1, m=2 training samples, symmetric feature probabilities.

For m=1, S_1 Î ω_1 with Prob = ½
For m=2, S_1 Î ω_1, S_2 Î ω_2

Conditional feature probabilities	P*	P^1	P^2
$P(x_1\|\omega_1) = 0.6.$	0.40	0.48000	0.4800
$P(x_1\|\omega_1) = 0.6,\ P(x_2\|\omega_1) = 0.60$	0.40	0.49000	0.4800

$P(x_1\|\omega_1) = 0.6,$	$P(x_2\|\omega_1) = 0.55$	0.40	0.49375	0.4875
$P(x_1\|\omega_1) = 0.6,$	$P(x_2\|\omega_1) = 0.50$	0.40	0.49500	0.4900

1 feature, 2 training samples (S1 Î w1, S2 Î w2)

Training set T_1: $S_1 = 1$, $S_2 = 0$;
Training set T_1: $S_1 = 0$, $S_2 = 1$;
Training set T_1: $S_1 = 1$, $S_2 = 1$;
Training set T_1: $S_1 = 0$, $S_2 = 0$;

$P(T_1) = 0.36$ $P(error|T_1) = 0.4$
$P(T_2) = 0.16$ $P(error|T_2) = 0.6$
$P(T_3) = 0.24$ $P(error|T_3) = 0.5$
$P(T_4) = 0.24$ $P(error|T_4) = 0.5$

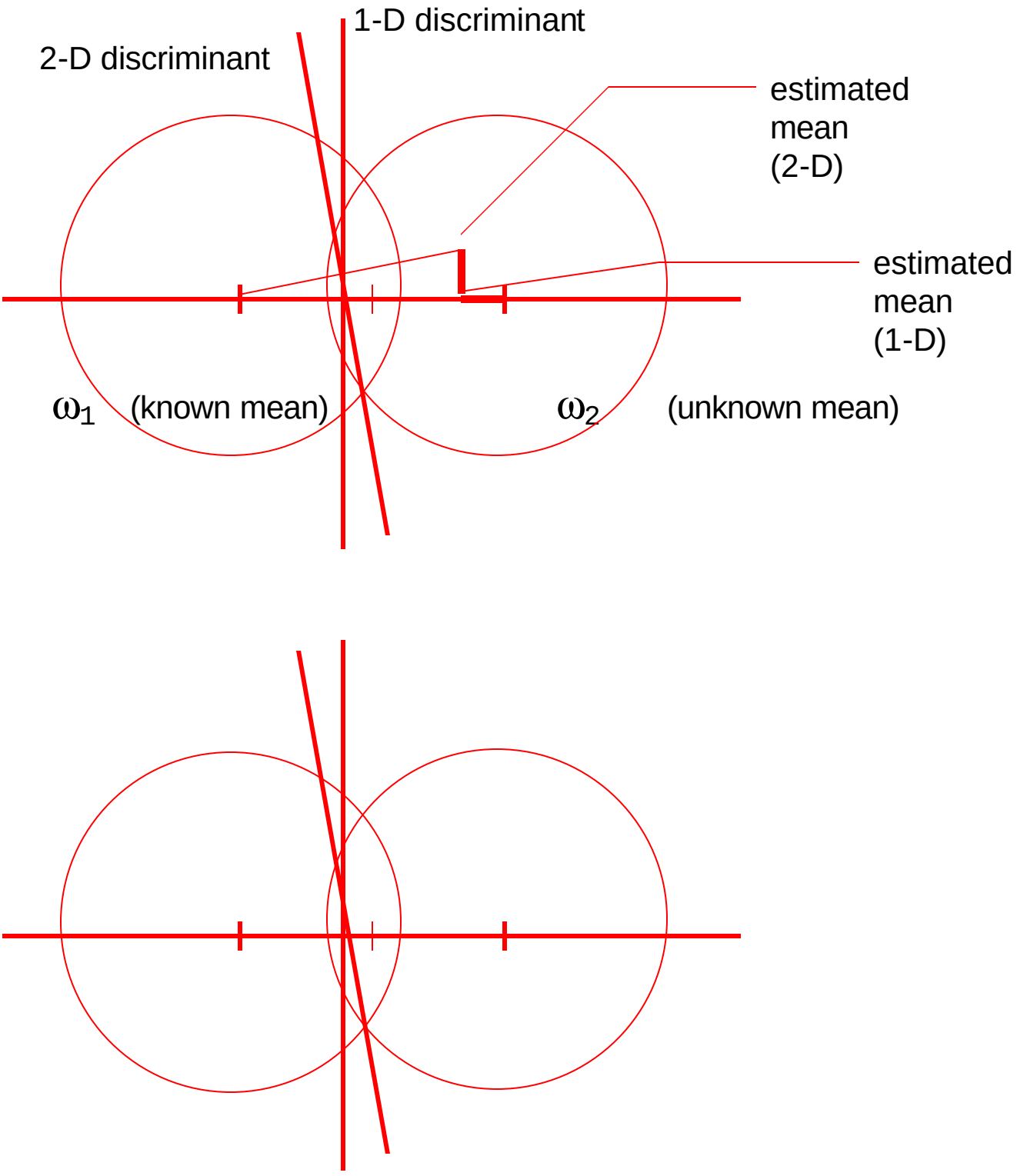
$P[error] = \Sigma\, P(T_i)\, P(error|T_i)$ $= 0.48$,
(Bayes error $= 0.40$).

2 features, 2 training samples (S_1 Î w_1, S_2 Î w_2)

	Case 1	**Case 2**	**Case 3**
	$P(x_1\|\omega_1)=0.60$	$P(x_1\|\omega_1)=0.60$	$P(x_1\|\omega_1)=0.60$
	$P(x_2\|\omega_1)=0.60$	$P(x_2\|\omega_1)=0.55$	$P(x_2\|\omega_1)=0.50$
	$P(x_1\|\omega_2)=0.40$	$P(x_1\|\omega_2)=0.40$	$P(x_1\|\omega_2)=0.40$
	$P(x_2\|\omega_2)=0.40$	$P(x_2\|\omega_2)=0.45$	$P(x_2\|\omega_2)=0.50$
T_1	**S_1=(1,1)**	**S_2=(0,0)**	
$P(T_1)$	0.1296	0.1089	0.0900
$P(error\|T_1)$	0.4000	0.4250	0.4500
T_2	**S_1=(0,0)**	**S_2=(1,1)**	
$P(T_2)$	0.0256	0.0324	0.0400
$P(error\|T_2)$	0.6000	0.5940	0.5500
T_3	**S_1=(0,1)**	**S_2=(0,1)**	
$P(T_3)$	0.0576	0.0594	0.0600

$P(error|T_3)$	0.5000	0.5000	0.5000
P[error]	0.4800	0.4875	0.4900

Graphical illustration of Hughes phenomenon:



Adding a useless dimension increases the error if sample size is finite.

Nonparametric Classification

Neural networks

Easy to modify to use decision-based reinforcement. Batch or sample based?

Nearest Neighbors

Use KNN and deport minority voters. Tag contribution of each reference sample, then omit long-unused references. (Prune and preprocess for speed.)

Validation set necessary for *all* adaptive classifiers.

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

Hybrid Classification

On the exponential value of labeled samples

V. Castelli and T. M. Cover, PRL 16, 105-111, 1995

2 classes, mixture coefficients P_A, $1-P_A$, identifiable densities

Labeled training samples: $X_1, X_2, \ldots X_m$
Unlabeled training samples: $X'_1, X'_2, \ldots$
Test sample: X_0

# of labeled samples	# of unlabeled samples	Probability of error
0	u	0.5
1	0	$2P_A(1-P_A)$ (1)
∞	u	P*
1	∞	2P*(1-P*)
m	∞	$P^* + e^{-am}$

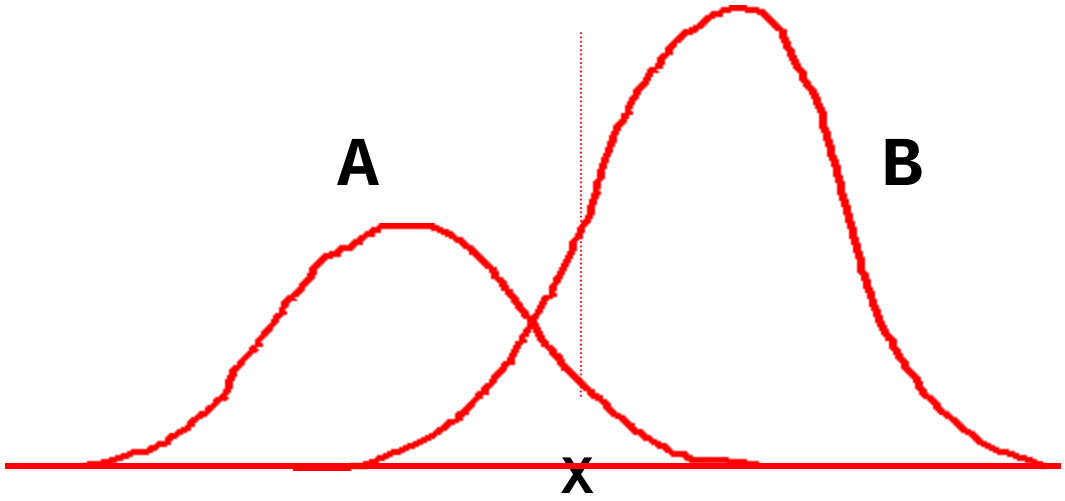
(1) For (1,0) decision rule is to give X_0 the same label as X_1.

Error(m,¥)

= P[training sample OK] × P[test sample wrong]
+ P[training sample wrong] × P[test sample OK]

= P* + (1-2P*) × P[training sample wrong]

$$\therefore \text{Error}(1,\infty) = P^* + (1-2P^*)P^* = 2P^*(1-P^*).$$



Error rate for m training samples:

if class X_i = l, P[training sample wrong]

= P[$\prod P_{class(l)} P(X_i|\text{class l}) < \prod P_{class(k)} P(X_i)|\text{class k})$]

(e.g., if X_i in A, = $P_A P(X_i|A) < P_B P(X_i|B)$)

Now bring products to one side, take logs, and consider each term in the resulting sum,

$$\log[P_{class(l)} P(X_i|\text{class l}) / P_{class(k)} P(X_i)|\text{class k})],$$

as an iid random variable. Then, for m>>1, "it can be shown"
that P[training sample wrong]

$$= e^{-m \log\{2(P_A P_B)^{1/2} \int [p(x|A)p(x|B)]^{1/2}\}} + o(m)$$

Hence the error, with m labeled samples, converges to the Bayes error at an exponential rate that depends on the overlap (*Bhattacharyya distance*) between the two densities (weighted by the *a prioris*).

MAUDE

B. Gold, Machine Recognition of Hand-Sent Morse Code, IRE-IT 17-24, March 1989

Adaptation to running average of the mean of short and long spaces after elimination of shortest and longest (out of 6) spaces.

Results on
184 messages, 45,000 characters, 53 operators
~16% of messages with >6% character error

Causes of error:
1. duration of marks and spaces
2. missing and extra marks and spaces (operator error)

Human performance (12 operators):
~ 1% on English, letter cipher, or number cipher text

Simulation on IBM 704:
adaptive estimation of variance would reduce pertinent character errors three-fold

Citation: "… speech recognizers using only the waveform of speech are bound to be severely limited,"

Adaptive Equalization

R.W. Lucky, Techniques for Adaptive Equalization of Digital Communication Systems, BSTJ XLV, 2, 255-286, February 1966

Adjust tap gains of transversal filter to preserve pulse waveform and keep eye diagram open
(assume output symbol recognized correctly)

16-level quantization for 9600 baud digital transmission over telephone lines

Typical error rate << 1%, but works well even at 10%

Delayed feedback depending on magnitude of deviation from ideal - adaptation time of the order of seconds

Up-down counters to store adaptation weights:
tap gain adjusted on underflow or overflow.

For binary operation, impossible to find a setting or disturbance from which equalizer does not recover

The equalizer is preset with a short sequence of known test pulses. Averaged tenfold gain by adaptation over preset equalizer.

Use of external error control also examined.

The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon

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

Data: AVIRIS multispectral (210 band) scanner

Four classes: soil, wheat, soybean, corn

Sample size: ~2500 pixels

Training set: 100 labeled samples/class (X 10)

Dimensionality: 1-18 bands
(bands ranked by Bhattacharyya distance)

Classifiers: 1. Gaussian quadratic 2. Minimum distance
(ML for labeled, MAP for unlabeled)

Training:
1. 20 labeled samples per class
2. 100 labeled samples per class
3. 20 labeled + 500 unlabeled samples
4. 20 labeled + 1000 unlabeled samples

EM ML estimation with unlabeled samples, using estimates based on labeled samples as starting point.

Experimental observations (Shashani and Landgrebe):

The quadratic classifier yields higher accuracy

The accuracy on test data at first increases, then *decreases* with dimensionality

The decrease is more pronounced for small training set

The decrease occurs earlier for the quadratic classifier

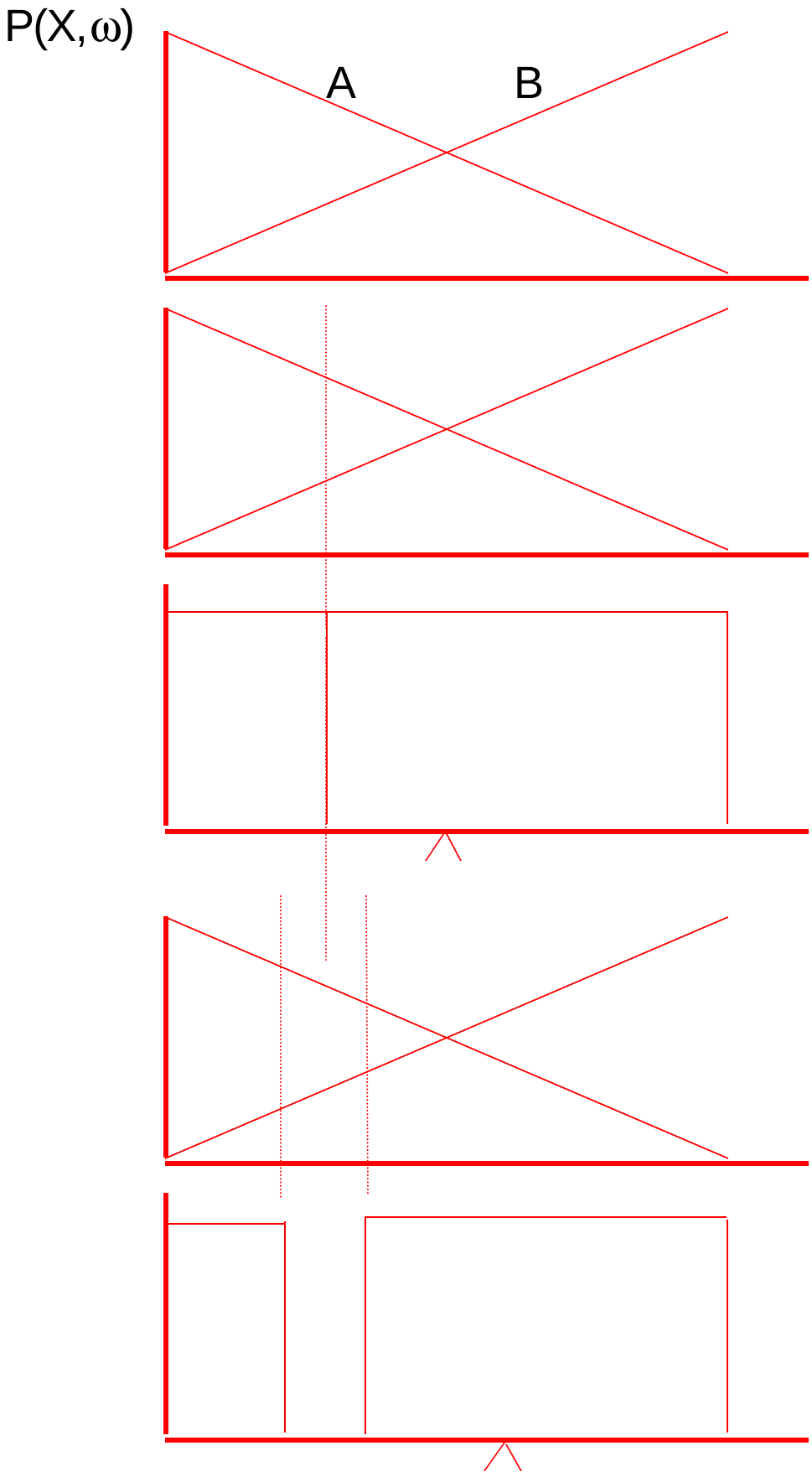
Additional unlabeled training samples reduce the error, and inrease the maximum useful dimensionality (5.4% at d=7 vs. 3.2% at d=13)

Part of the observed effect is due to MAP classification

Unlabeled samples are particularly useful when labeled training set drawn from adjacent pixels

(The paper does not state whether multimodal classifiers were used for any class)

Rejection in adaptation



SUMMARY

Both explicit (parametric classifiers) and implicit (nonparametric classifiers) require *estimating* the characteristics of the data from a training set. The single most important set of parameters are usually 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 (HMM) distributions require mixture estimation techniques.

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

There is no such thing as completely *unsupervised* learning or classification. But the effective sample size can be increased by taking advantage of *unlabeled samples*.

When the training set is not representative, use *adaptive methods* that exploit mostly-correct classification.

Nevertheless, *labeled samples* are valuable: endeavor to obtain more.

Don't waste *rejects*: they may contain useful information.

Don't *ever* let the machine rest.

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, 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, 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 17-24, March 1989.

Lucky, R.W., Automatic equalization for digital communication, BSTJ XLIV, 4, 547-588, 1965.

Lucky, R.W., Techniques for adaptive equalization of digital communication systems, BSTJ XLV, 2, 255-286, February 1966.

Nagy, G., and 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, 1087-1095, 1994.