SOME REMARKS ON THE GAUSSIAN DISCRIMINANT

UNCLASSIFIED

END

DATE

12-MAY

DTIC
Some Remarks on the Gaussian Discriminant

Lincoln Laboratory
SOME REMARKS ON THE GAUSSIAN DISCRIMINANT

L. K. JONES
Group 92

TECHNICAL NOTE 1980-47

3 OCTOBER 1980

Approved for public release; distribution unlimited.
ABSTRACT

We comment on the performance of the Gaussian discriminant function with (possibly) non-Gaussian underlying distributions. An asymptotic expression for the probability of error for the Gaussian case is given with a formal convergence proof.
I. INTRODUCTION

For many practical problems in two class pattern recognition, one has (reliable estimates of) the first two moments of each class (mean vectors in \( \mathbb{R}^n \) - \( M_1 \), \( M_2 \) and covariance matrices - \( \Sigma_1 \), \( \Sigma_2 \)). Whether or not the underlying distributions are indeed Gaussian, one proceeds to apply the standard Gaussian hypothesis test to classify new data. More precisely, one uses the Gaussian discriminant function \( h(X) = \log \frac{p_2(X)}{p_1(X)} \), where \( p_1, p_2 \) are multivariate normal with the same first two moments as the underlying distributions. Applying an affine transformation to our problem (which has no effect on the discriminant \( h \)) that simultaneously diagonalizes \( \Sigma_1 \) and \( \Sigma_2 \) \( (\Sigma_1 \rightarrow \mathbf{I}, \Sigma_2 \rightarrow \Lambda, M_2 \rightarrow \overline{\mathbf{0}}, M_1 \rightarrow (d_1, d_2, \ldots, d_n) \) with \( d_k \geq 0 \), we have

\[
(1) \quad h(X) = \frac{1}{2} \sum_{k=1}^{n} \left[ (x_k - d_k)^2 - x_k^2/\lambda_k + \ln(1/\lambda_k) \right]
\]

In this correspondence, we first present some elementary inequalities in \( h \), valid regardless of the class distributions; and then we demonstrate the asymptotic result:

\[
(2) \quad p_{\text{error}} \approx \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{J}{2}}}^{\infty} e^{-\frac{1}{2}x^2} \, dx \quad \text{(with } J \text{ the divergence)}
\]
for the case of equal priors, Gaussian distributions, and all 
\( \lambda_k \) close to 1. We note that the above does not follow from 
the elementary fact that, for fixed \( n \), \( h(x) \) is a linear function 
as all \( \lambda_k \to 1 \); for all \( \lambda_k \) may be close to 1 but the quadratic 
part of \( h = \frac{1}{2} \sum_{k=1}^{n} x_k^2 \left( 1 - \frac{1}{\lambda_k} \right) \) may not approach 0 if \( n \) becomes large.

II. THE GAUSSIAN DISCRIMINANT FOR ARBITRARY CLASS DISTRIBUTIONS

Calculating the first moments of \( h \) under each hypothesis, 
we have, regardless of the underlying distributions:

\[
E_1(h) = \frac{1}{2} \sum_{k=1}^{n} \left[ (1 - \frac{1}{\lambda_k}) - \frac{d_k^2}{\lambda_k} + \ln(1/\lambda_k) \right]
\]

\[
E_2(h) = \frac{1}{2} \sum_{k=1}^{n} \left[ (\lambda_k - 1) + d_k^2 + \ln(1/\lambda_k) \right]
\]

Since \( Z - 1 + \ln(1/Z) \geq 0 \) for all \( Z > 0 \), we see immediately that

\[
E_2(h) \geq \frac{1}{2} \sum_{k=1}^{n} d_k^2 = \frac{1}{2} D^2.
\]

Noting that the maximum value of \( f(Z) = 1 - \frac{1}{Z} - \frac{1}{Z^2} + \ln(1/Z) \)
for \( Z > 0 \) occurs at \( Z = 1 + \gamma^2 \), we have \( f(Z) \leq 1 - \frac{1}{1 + \gamma^2} \)
+ \( \ln\left(\frac{1}{1 + \gamma^2}\right) - \frac{\gamma^2}{1 + \gamma^2} \leq - \frac{\gamma^2}{1 + \gamma^2} \). Hence
(6) \( E_1(h) \leq -\frac{1}{2} \sum_{i=1}^{n} d_i^2/1 + d_i^2 \)

which is \( \approx -\frac{1}{2} D^2 \) if each component \( d_i \) is small. Therefore, in many practical problems \( E_2(h) - E_1(h) \geq D^2 = \sum_{i=1}^{n} d_i^2 \). \( D^2 \) is then a first order measure of the performance of \( h \). If \( n \) is large, the \( \lambda_i \) are close to one, the \( d_i \) are small, and the sequence of random variables \( x_i \) is \( k \) dependent for small \( k \), then we could apply the central limit theorem and obtain estimates of the error probability of \( h \) by calculating \( \text{Var}_1(h) \) and \( \text{Var}_2(h) \) from sample data.

III. ASYMPTOTIC APPROXIMATION TO ERROR PROBABILITY

To justify the claim in I, we state and prove the following theorem:

Theorem: Let a sequence of decision problems, with underlying Gaussian distributions described by means \( D_i, \bar{0} \) in \( \mathbb{R}^{n_i} \) and covariances \( I, \Lambda_i \), be given. Then, if \( \max_{i \leq k \leq n_i} |\lambda_i^k - 1| \to 0 \) as \( i \to \infty \),

\[
\left| P_{\text{error}}^i - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} \, dx \right| \to 0
\]

for the equal prior case.
Proof: We shall apply a central limit theorem for arrays of random variables and use the first two moments of \( h^i \) to obtain an asymptotic expression for the error probability. Calculating the variances under each hypothesis of \( h^i \), we obtain

\[
\text{Var}_1(h^i) = \lambda \sum_{l}^{n_i} \left[ \left(1 - \frac{1}{\lambda_{l}^i} \right)^2 + \frac{2(d_{l}^i)^2}{\lambda_{l}^i} \right] \tag{7}
\]

\[
\text{Var}_2(h^i) = \lambda \sum_{l}^{n_i} \left[ \left(\lambda_{l}^i - 1 \right)^2 + 2\lambda_{l}^i \left(d_{l}^i \right)^2 \right] \tag{8}
\]

Using (3), (4), (7) and (8), and noting by elementary calculus that

\[
\frac{(1-1/\lambda_{l}^i)^2}{1-1/\lambda_{l}^i + \ln(1/\lambda_{l}^i)} \rightarrow -2 \left(1-1/\lambda_{l}^i \right) \frac{(\lambda_{l}^i - 1)}{(\lambda_{l}^i - 1)} \rightarrow -2
\]

\[
\frac{(\lambda_{l}^i - 1)^2}{\lambda_{l}^i - 1 + \ln(1/\lambda_{l}^i)} \rightarrow \frac{2(\lambda_{l}^i - 1)}{(1-\lambda_{l}^i)} = 2
\]

\[
\frac{2(d_{l}^i)^2}{\lambda_{l}^i} \rightarrow -\frac{(d_{l}^i)^2}{\lambda_{l}^i} = -2
\]

\[
\frac{2\lambda_{l}^i(d_{l}^i)^2}{(d_{l}^i)^2} \rightarrow +2
\]
we have
\[ \frac{\text{VAR}_1(h^i)}{2E_1(h^i)} \rightarrow -1 \]
and
\[ \frac{\text{VAR}_2(h^i)}{2E_2(h^i)} \rightarrow +1 \]

Furthermore
\[ -\frac{(d_x^i)^2}{\frac{1}{\lambda_x^i}} \rightarrow -1 \]

and
\[ \frac{(1-\frac{1}{\lambda_x^i}) + \ln \left(1/\lambda_x^i\right)}{(1-\frac{1}{\lambda_x^i}) + \ln \left(1/\lambda_x^i\right)} \rightarrow \frac{1-1/\lambda_x^i}{1/(\lambda_x^i)^2 - 1/\lambda_x^i} \rightarrow -1 \]

imply that
\[ \frac{E_2(h^i)}{E_1(h^i)} \rightarrow -1 \]
or equivalently
\[ \frac{E_2(h^i)}{J^i} \rightarrow +\frac{1}{5} \]
\[ \frac{E_1(h^i)}{J^i} \rightarrow -\frac{1}{5} \]

We now proceed with the main proof. We may assume (by passing to subsequences if necessary) that both $J^i$ and $p_{\text{error}}^i$ are convergent sequences (possibly to $\pm \infty$ in the case of $J^i$). We divide the argument into several cases:
CASE (1) 

It suffices to show that $P_{\text{error}} \to \frac{1}{2}$. This is actually true in general. Consider any 2 positive density functions, $p, q$, on some probability space. Then, if for some real $\delta > 0$, there is no measurable set whose $q$ measure is greater than $\delta$ and such that on this set $q/p > 1 + \delta$, it follows that

$$P_{\text{error}} = \frac{1}{2} \left( \int_{q \leq p} q + \int_{q > p} p \right) =$$

$$\frac{1}{2} \left[ \int_{q \leq p} q + \int_{q > 1 + \delta} p + \int_{1 < q/p < 1 + \delta} p \right] \geq$$

$$\frac{1}{2} \left[ \int_{q \leq p} q + \int_{1 < q/p \leq 1 + \delta} (p/q) q \right] >$$

$$\frac{1}{2} \left[ \frac{1}{1 + \delta} \int_{q \leq p} q + \frac{1}{1 + \delta} \left( \int_{q > 1} q - \int_{q > 1 + \delta} q \right) \right] \geq \frac{1 - \delta}{2(1 + \delta)}.$$  

Hence if $P_{\text{error}}$ does not approach $\frac{1}{2}$, such a $\delta$ exits. But then the divergence $J^i(p,q) =$

$$\int_{p \geq q} \ln(p/q)(p-q) + \int_{q > p} \ln(q/p)(q-p)$$

6
\[
\ln(1+\delta) \left[ \left( 1 - \frac{1}{1+\delta} \right) \delta \right] = \frac{\delta^2 \ln(1+\delta)}{1+\delta} > 0.
\]

**CASE (2) \( J^i \rightarrow J \neq 0 \)**

Let's rewrite

\[
h^i(x) = b \sum_{l}^{n_i} \left[ \left( x^i_{l} \right)^2 \left( 1 - 1/\lambda^i_{l} \right) - 2x^i_{l} d^i_{l} \right] + k_i
\]

where we reorder the \( d^i_{l} \) such that

\[
d^i_{l} \geq d^i_{l+1}.
\]

**Subcase (a) \( \sup_{l} \left( \sum_{1}^{n_i} (d^i_{l})^2 \right) = +\infty \).**

Clearly from (5) \( J = +\infty \). Consider the (sub-optimal) discriminants \( g^i = \sum_{1}^{n_i} x^i_{l} d^i_{l} \). These are normally distributed with means, \( \sum_{1}^{n_i} (d^i_{l})^2 \) and 0, and standard deviations \( \sqrt{\sum_{1}^{n_i} (d^i_{l})^2} \) and \( \sqrt{\sum_{1}^{n_i} \lambda^i_{l} (d^i_{l})^2} \). One can then find arbitrarily large \( i \) for which \( g^i \) has arbitrarily small error probability. Since \( h^i \) is optimal, it has arbitrarily small error for these \( i \) and hence, \( p^i_{\text{error}} \rightarrow 0. \)
Subcase (b) \[ \sup \left( \sum_{i=1}^{n_i} (d^i_{\lambda_i})^2 \right) < +\infty. \]

We first note that \( \text{Var}(h^i) \to J \neq 0 \) under either hypothesis.

Let us rewrite \( h^i = \frac{1}{\sqrt{n_i}} \sum_{i=1}^{n_i} \left[ (x^i_{\lambda_i})^2 \left( 1 - 1 / \frac{i}{\lambda_i} \right) - 2x^i_{\lambda_i} d^i_{\lambda_i} \right] \)

\[ + \frac{1}{\sqrt{n_i}} \sum_{i=n_i+1}^{n_i} \left[ (x^i_{\lambda_i})^2 \left( 1 - 1 / \frac{i}{\lambda_i} \right) - 2x^i_{\lambda_i} d^i_{\lambda_i} \right] + F_{i=1} = F_1^i + F_2^i + K_i \text{ with} \]

\( n_i \) chosen such that \( n_i \to \infty \) but that \( \sum_{i=1}^{n_i} |1 - 1 / \frac{i}{\lambda_i}| \to 0 \). We may now apply a central limit theorem, for instance Corollary 4.2 on page 232 of \([1]\): For any \( \beta > 0 \), either \( F_2^i \) has variance \( < \beta \), or \( F_2^i \) becomes normal in distribution for large \( i \). This follows from the central limit theorem for arrays mentioned above, provided the variances of the terms in the summand of \( F_2^i \) become arbitrarily small and this follows if

\[ \sup_{i} \left\{ d^i_{\lambda_i}; \lambda > \sqrt{n_i} \right\} \to 0. \] But if this were not the case, \( d^i_{\lambda_i} > \gamma > 0 \) for infinitely many \( i \) and hence, since \( n_i \to \infty \), \( \sum_{i=1}^{n_i} (d^i_{\lambda_i})^2 > \gamma^2 \)

contradicts our initial assumption. Further, \( F_1^i \) either has variance \( < \beta \) or approaches a normal random variable in distribution since its linear part is normal and its nonlinear part has variance approaching 0. Since \( \beta \) was arbitrary, \( J > 0 \), and \( F_1^i \) is independent of \( F_2^i \); \( h^i \) approaches a normal random variable in distribution and we obtain the asymptotic error
Finally we note that, in (2), we could replace $J$ by $8B$ where $B$ is the Bhattacharyya distance. This follows from the simply verified fact that $\frac{8B}{J} \to 1$ as all $\lambda_\ell \to 1$. 
REFERENCES


We comment on the performance of the Gaussian discriminant function with (possibly) non-Gaussian underlying distributions. An asymptotic expression for the probability of error for the Gaussian case is given with a formal convergence proof.