L6 – Bounds on Error Rates

Bounding the Bayes Error Rate

Recall the likelihood ratio tests, written in terms of the negative log of the likelihood

$$h(x) \equiv -\log\left(\frac{p(x|\omega_1)}{p(x|\omega_2)}\right) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -\log\eta \equiv t$$

Computation of the Bayes error rate for all but the simplest cases (e.g. Gaussian features x with equal cla es) is difficult.

Calculating the integrals

$$\mathcal{E}_2 = \int_{L1} p(x|\omega_2)\, d^n x = \int_{-\infty}^{t} p(h| \ {}_2$$

is difficult or impossible

$$\mathcal{E}_1 = \int_{L2} p(x|\omega_1)\, d^n x = \int_{t}^{\infty} p(h|\omega_1)\, dh$$

GEORGETOWN UNIVERSITY

Bounding the Bayes Error Rate

Instead, we seek <u>bounds</u> on the error that are

- Hopefully tight
- Easy to calculate

In practice, we estimates, but the theoretical b seful tools to reason with.

The bounds point out factors e error rate. That is, the theoretical framework helps you conceptualize the problem.

A Calculable Case

Let both class-conditional densities be normal, with the same covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$. The neg. log-likelihood ratio is then

$$h(x) = (M_2 - M_1)^T \Sigma^{-1} x + \tfrac{1}{2}\left(M_1^T \Sigma^{-1} M_1 - M_2^T \Sigma^{-1} M_2\right)$$

since x is Gauss h is Gaussian. Its mean and varia

$$E[h \mid \omega_1] = -\tfrac{1}{2}\,(M_2 - M_1)^T \Sigma^{-1} (M_2 - M_1$$

$$E[h \mid \omega_2] = \tfrac{1}{2}\,(M_2 - M_1)^T \Sigma^{-1} (M_2 - M_1) \equiv \theta$$

$$\sigma_i^2 = (M_2 - M_1)^T \Sigma^{-1} (M_2 - M_1) = 2\theta$$

GEORGETOWN UNIVERSITY

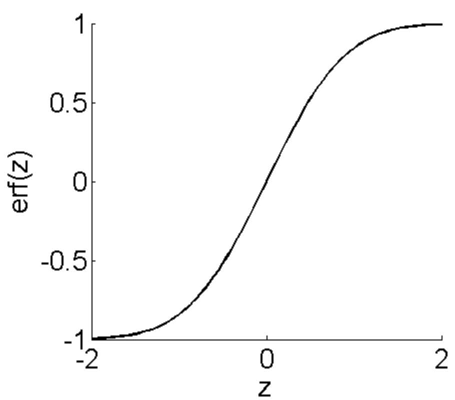
Calculable Case

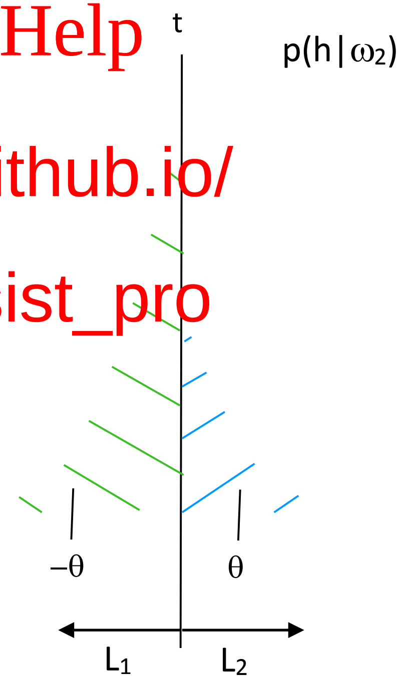
The error rates become

$$\mathcal{E}_2 = \int_{-\infty}^{t} p(h\,|\,\omega_2)\,dh = erf \;\frac{+\theta}{\quad}$$

$$\mathcal{E}_1 = \int_{t}^{\infty} p(h\,|\,\omega_1)\,dh = 1 - erf\left(\frac{t+\theta}{\sigma}\right)$$

$$erf(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\xi^2/2}\, d\xi$$





GEORGETOWN UNIVERSITY

Back to Bounds

Chernoff Bound

Classification error rate is

$$\mathcal{E} = P_1 \int_{L2} p(x|\omega_1)\; d^n x + P_2 \int_{L1} p(x|\omega_2)\; d^n x$$

$$= \int \min\left[P_1\; p(x|\omega_1),\; P_2\; p(x|\omega_2)\right]\; d^n x$$

Use inequality $\min[a,b] \le a^s\, b^{1-s}\;; \quad a,b \ge 0,\; 0 \le s \le 1$

to obtain

$$\boxed{\begin{aligned} \mathcal{E} \le \mathcal{E}_U &= \min_s \left(P_1^s\; P_2^{1-s} \int p^s(x|\omega_1)\; p^{1-s}(x|\omega_2)\; d^n x\right) \\ &\equiv \min_s \left(P_1^s\; P_2^{1-s}\quad e^{-\mu(s)}\right) \end{aligned}}$$

Simplification – Bhattacharyya Bound

Don’t insist on minimizing with respect to s, but take s=1/2

$$\mathcal{E}_U = \sqrt{P_1 \, P_2} \int \sqrt{p(x|\omega_1)\, p(x|\omega_2)}\; d^n x$$

$$= \sqrt{P_1 \, P_2}\; e^{-\mu(1/2)}$$

for x normal, the exponent is

$$\mu(1/2) = \frac{1}{8}\left(M_2 - M_1\right)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \left(M_2 - M_1\right) + \tfrac{1}{2}\ln\left(\frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right)$$

Bhattacharyya Bound – Special Cases

- Equal Means

$$\mu(1/2) = \frac{1}{2}\ln\left(\frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right)$$

Equal Covariances

$$\mu(1/2) = \frac{1}{8}\left(M_2 -\right.$$

Single Hypothesis

May have data predominantly from one class -- e.g. failure analysis, usually have many examples of "healthy" function and only a few failures; may have many different failure modes.

Given an observati corresponds to a healthy, or a failed ovelty detection."

May use some kind of distance m .g.

$$d^2 = (x-m)^T \ \Sigma^{-1} \ (x-m)$$

where Σ and m are the covariance and mean of the class <u>ω_1 that one is able to model well.</u>

Single Hypothesis

To use this in classification

Given x, compute d^2 and compare with a threshold

$$d^2 \underset{not\ \omega_1}{\overset{\omega_1}{\lessgtr}} c$$

where the threshold c is ch[illegible] on both in-class, and out-of-class data, or …

GEORGETOWN UNIVERSITY

Single Hypothesis

Statistics of d^2

d^2 is a random variable, with mean and variance

$$E[d^2] = E[(x-m)^T \Sigma^{-1} (x$$

$$= Trace\, E\left[\Sigma^{-1} (x -\right.$$

$$\mathrm{var}[d^2] = E\left[d^4\right] - \left(E\left[d^2\right]\right)^2$$

Note: d^2 is the sum of N random identically distributed random variables. For large dimension N, we can invoke the central limit theorem with the result that d^2 becomes normally distributed.

Single Hypothesis

Other "distance" measures

– Model the probability distribution for objects in the known class - e.g. as a Gaussian mixture

$$p(x \mid \omega_1) = \sum_{k=1}^{Q} a_k \frac{1}{\sqrt{(2\pi)^n \, |\Sigma_k|}} \exp\left(-\tfrac{1}{2} (x - m_k) \, \Sigma_k^{-1} \, (x - m_k) \right)$$

where the parameters a_k, m_k, Σ_k fit by maximum likelihood. Then classify a new datum x_0 using a threshold test

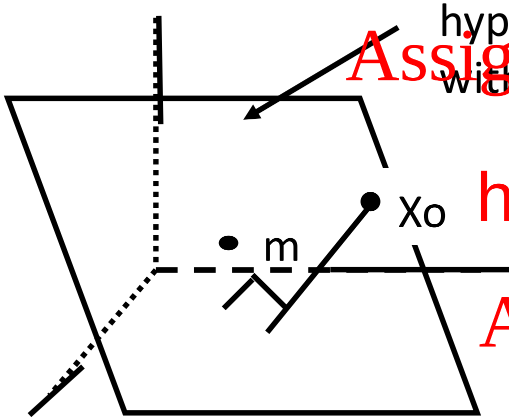
$$-\ln \; p(x \mid \omega_1) \; \underset{not\,\omega_1}{\overset{\omega_1}{\lessgtr}} \; c$$

GEORGETOWN UNIVERSITY

Single Hypothesis

Other “distance” measures, continued

Model known class by PCA subspace



hyperplane spanned by leading $m<n$ eigenvectors of Σ (those with highest eigenvalues -> largest variance directions.

Then classify a by a threshold test on its perpendicular hyperplane.

The hyperplane serves as a geometric model of the known class. See Oja, <u>Subspace methods of Pattern Recognition</u>, Wiley and Sons, for other examples.

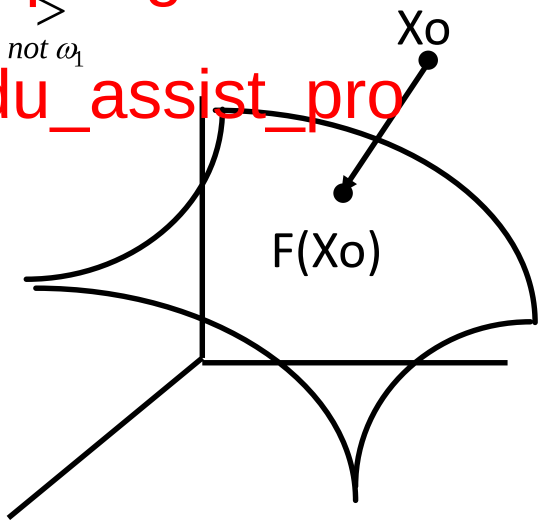
GEORGETOWN UNIVERSITY

Single Hypothesis

Other distance measures (cont'd)

Model known class as a <u>curved manifold</u> (constructed e.g. with a neural network). Measure distance between data x_0 and its projection $F(x_0)$ onto the manifold.

The threshold test is $d^2 = \|x_0 - F(x_0;\theta)\|^2 \underset{not\ \omega_1}{\overset{\omega_1}{\lessgtr}} c$



Problems with Single Hypothesis Tests

- Usually perform worse than full Bayesian tests. If you have enough data to model both classes, performance should improve.

- We turned to is test because one of the two classes is ely sampled to model well. How does one pick the decision threshold c?

GEORGETOWN UNIVERSITY