ANLY-601
Advanced Pattern Recognition

Spring 2018

L6 – Bounds on Error Rates

GEORGETOWN UNIVERSITY

Bounding the Bayes Error Rate

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

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

Computation of the Bayes error rate for all but the simplest cases (e.g. Gaussian features x with equal class-conditional covariance matrices) is difficult.

Calculating the integrals

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

is difficult or impossible

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

Bounding the Bayes Error Rate

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

- Hopefully tight
- Easy to calculate

In practice, we will use empirical error estimates, but the theoretical bounds here provide useful tools to reason with.

The bounds point out factors entering the 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

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

since x is Gaussian, and h is linear in x, h is Gaussian. Its mean and variance are

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

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

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

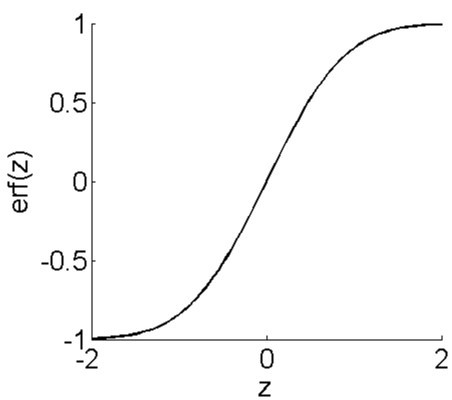
Calculable Case

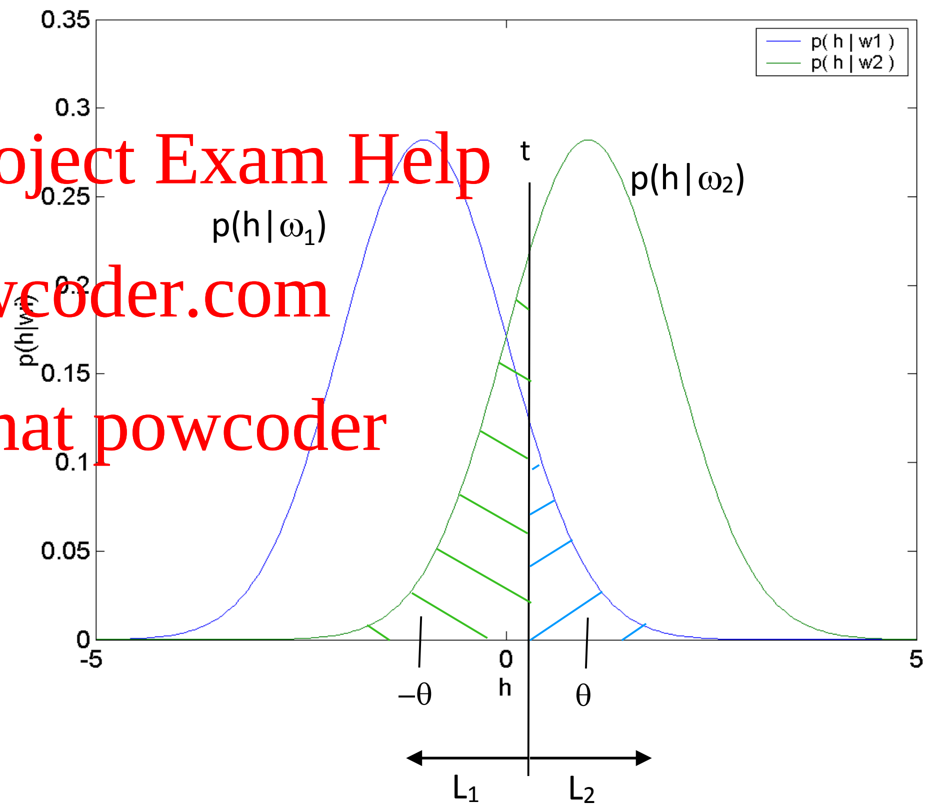
The error rates become

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

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

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





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

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

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

$$\mu(1/2) = \frac{1}{8}\left(M_2 - M_1\right)^T \Sigma^{-1}\left(M_2 - M_1\right)$$

Single Hypothesis Tests

GEORGETOWN UNIVERSITY

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 observation x, want to know if the corresponds to a healthy, or a failed example. Also called “novelty detection.”

May use some kind of distance measure – e.g.

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

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

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 chosen based on both in-class, and out-of-class data, or …

Single Hypothesis

Statistics of d^2

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

$$E[d^2] = E[(x-m)^T \Sigma^{-1} (x-m)]$$
$$= Trace\,E\left[\Sigma^{-1}(x-m)(x-m)^T\right] = Trace\left[\Sigma^{-1}\Sigma\right] = N$$

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

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

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

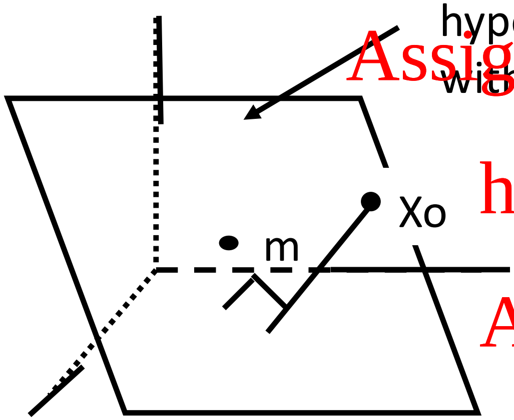
where the parameters a_k, m_k, Σ_k are fit by maximum likelihood. Then classify a new datum x_0 using a threshold test

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

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 new datum x_0 by a threshold test on its perpendicular distance to the hyperplane.

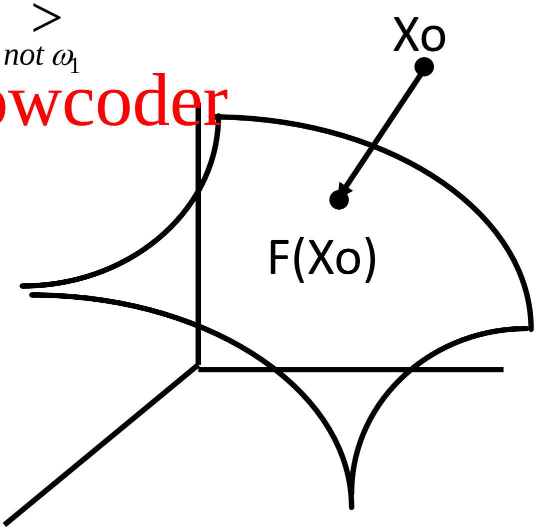
The hyperplane serves as a geometric model of the known class. See Oja, <u>Subspace methods of Pattern Recognition</u>, Wiley and Sons, for other examples.

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 the single hypothesis test because one of the two classes is too sparsely sampled to model well. How does one pick the decision threshold c?

1789

GEORGETOWN UNIVERSITY