ANLY-601
Advanced Pattern Recognition

Spring 2018

L5 – Bayes Classifiers (cont'd)

GEORGETOWN UNIVERSITY

Summary of Dichotomy (Two-Class) Hypothesis Tests

- Bayes least error rate

$$l(x) \equiv \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P_2}{P_1} \equiv \eta$$

- Bayes least cost

$$l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P_2}{P_1}$$

Summary of Dichotomy Hypothesis Tests

- Neyman-Pearson

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \mu \quad \text{with} \quad \left(\int_{L_1(\mu)} p(x|\omega_2)\, d^n x \right) = \mathcal{E}_0$$

- Minimax

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta \quad \text{theshold such that} \quad \mathcal{E}_1 = \mathcal{E}_2$$

Multi-Hypotheses

Suppose there are L classes $\omega_1, \ldots, \omega_L$ and decision costs λ_{ij} for choosing i when j is true. Then the minimal cost decision rule is

pick ω_k where $k = \arg\min_i R_i(x) = \arg\min_i \sum_{j=1}^{L} \lambda_{ij}\; p(\omega_j \mid x)$

When $\lambda_{ii} = 0 \;\; and \;\; \lambda_{ij} = 1, i \neq j$

the cost is just the average error rate, and the decision rule is

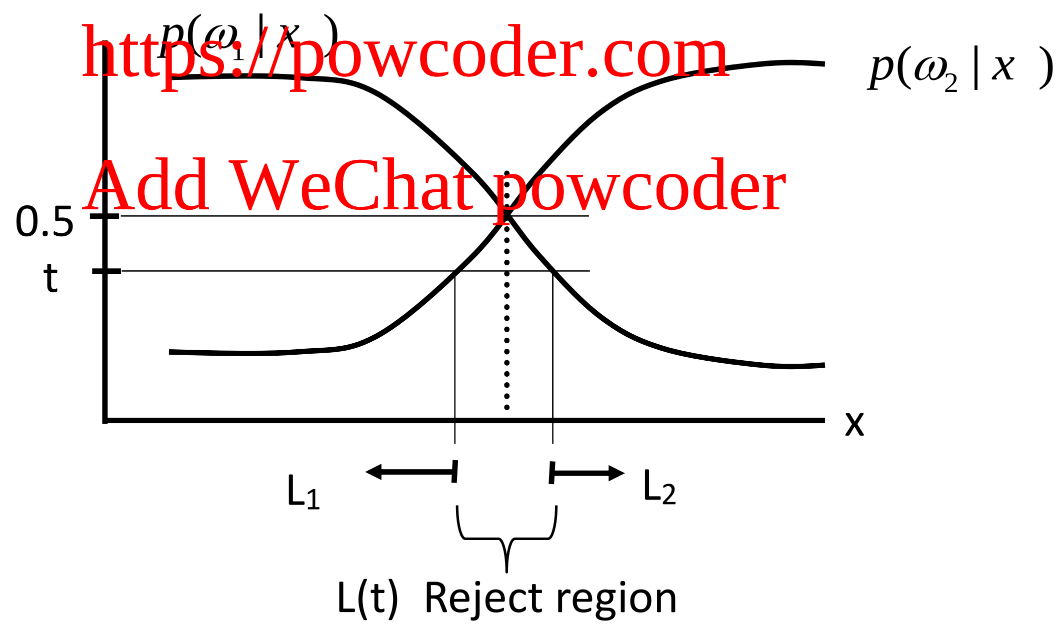
pick ω_k $\quad k = \arg\max_i \; p(\omega_i \mid x)$

Reject Option

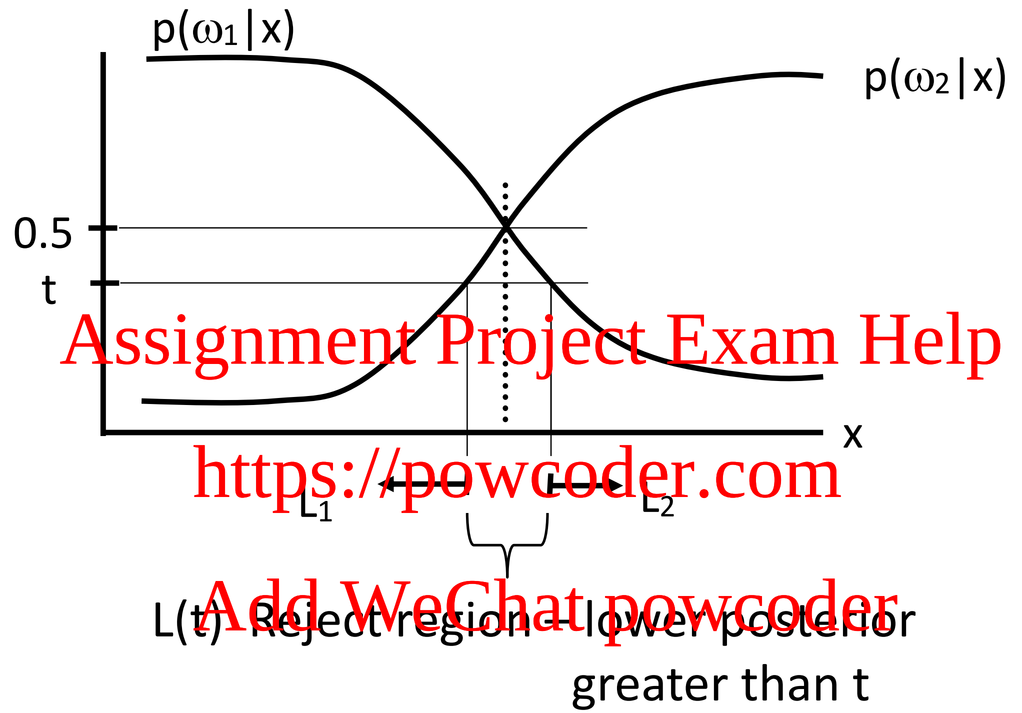
For a 2-class, least error rate problem, when the posteriors are close to 0.5, the error rate will be large

$$\mathcal{E}(x) = \min_i \left(p(\omega_i \mid x) \right)$$

One might want to establish a window for rejection within which we refuse to make a judgment



Reject Option



Reject rate $\quad \mathrm{Prob}\big(x \in L(t)\big) = \int_{L(t)} p(x)\, d^n x$

Error rate $\quad \mathcal{E} = \int_{\bar{L}(t)} \min\left[\; p(\omega_1|x),\; p(\omega_2|x) \;\right] p(x) \quad d^n x$

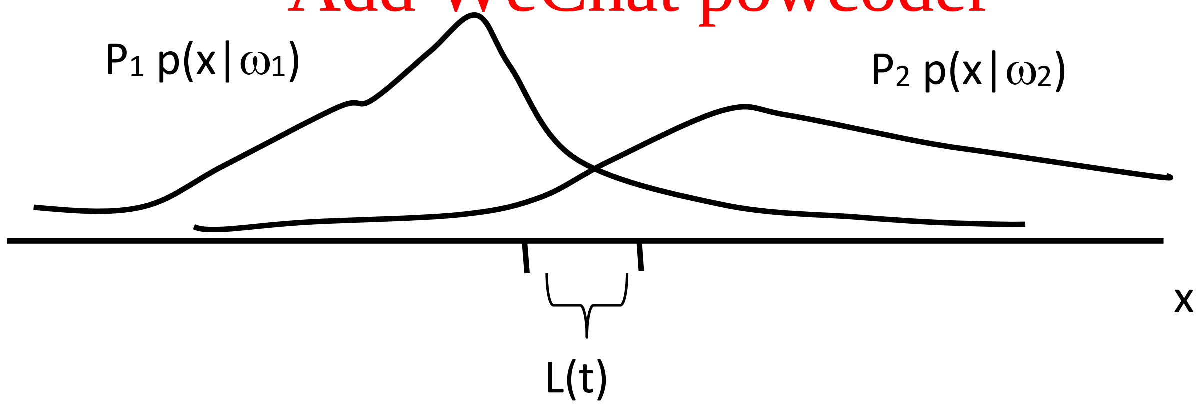
GEORGETOWN UNIVERSITY

Reject Option

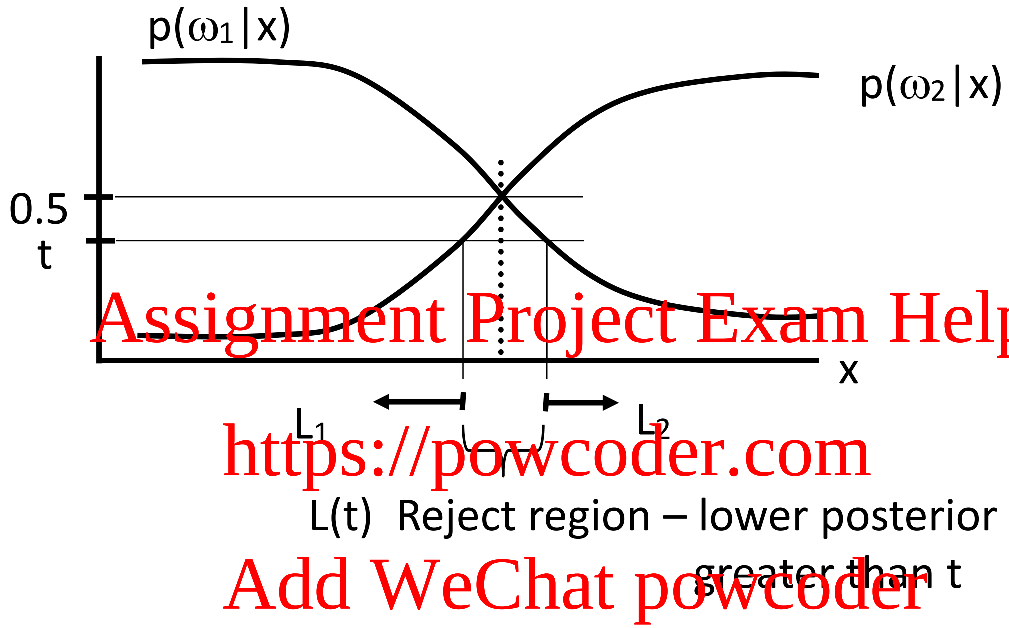
Error rate

$$\mathcal{E} = \int_{\overline{L(t)}} \min \left[\, p(\omega_1 \mid x), \; p(\omega_2 \mid x) \, \right] \; p(x) \; d^n x$$

$$= \int_{\overline{L(t)}} \min \left[\, P_1 \; p(x \mid \omega_1), \; P_2 \; p(x \mid \omega_2) \right] d^n x \;\; \equiv \;\; P_1 \mathcal{E}_1 \; + \; P_2 \mathcal{E}_2$$



Reject Option



- Reject option lowers error rate by refusing to make decisions on feature values *x* where the error rate is high (near the crossing of the posterior curves).
- A larger reject region (smaller *t*) lowers the error, and increases the rate at which we refuse to make a decision.

Sequential Hypothesis Tests

Have sequence of observations

$$x_1,\ x_2,\ \ldots,\ x_n$$

assumed to be independent and identically distributed (i.i.d.). May be from a timeseries, e.g. speech segments, manufacturing production run …

Each sequence is from one of two possible classes.

Suppose we want to continue to accrue information from this sequence until we have enough information to make a decision -- e.g. maybe we have a reject threshold to overcome.

It seems clear that if we make many measurements (e.g. on consecutive items in a manufacturing production run) that we'll improve our classification results.

Sequential Hypothesis Tests

- log likelihood ratio

$$H(x_1, x_2, \ldots, x_m) \equiv -\ln \frac{p(x_1, x_2, \ldots, x_m | \omega_1)}{p(x_1, x_2, \ldots, x_m | \omega_2)}$$

$$= -\sum_{i=1}^{m} \ln \left(\frac{p(x_i | \omega_1)}{p(x_i | \omega_2)} \right) = \sum_{i=1}^{m} h(x_i)$$

How does H behave relative to h_i ? Let's look at its mean and variance

$$E[H | \omega_i] = m\ E[h | \omega_i] \equiv m\ \mu_i$$

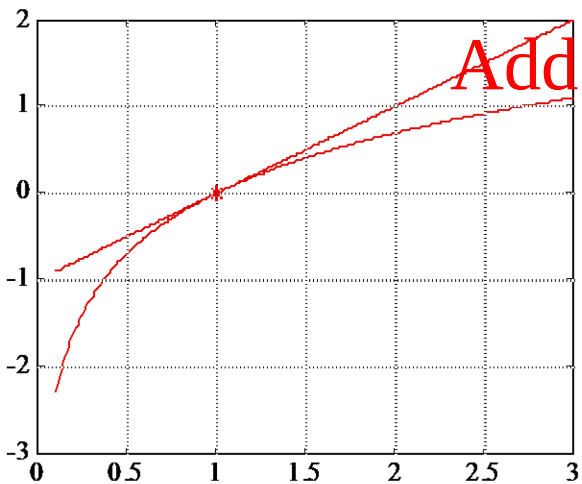
$$\mathrm{var}[H | \omega_i] = \mathrm{var}\left[\sum_{i=1}^{m} h(x_i)\ | \omega_i \right] = m\ \mathrm{var}[h | \omega_i] \equiv m\ \sigma_i^2$$

Sequential Hypothesis Test

Conditional mean

$$\mu_i \equiv E\left[h \mid \omega_i\right] \equiv \int -\ln\left(\frac{p(x\mid\omega_1)}{p(x\mid\omega_2)}\right) p(x\mid\omega_i)\; d^n x$$

Can bound μ_i even for arbitrary density by appeal to the inequality ln z <= z-1



This gives

$$\mu_1 \equiv \int \ln\left(\frac{p(x\mid\omega_2)}{p(x\mid\omega_1)}\right) p(x\mid\omega_1)\; d^n x$$

$$\leq \int \left[\frac{p(x\mid\omega_2)}{p(x\mid\omega_1)} - 1\right] p(x\mid\omega_1)\; d^n x \;=\; 1 - 1 = 0$$

$$\textit{So} \quad \mu_1 \leq 0$$

$$\textit{Similarly}, \quad \mu_2 \geq 0.$$

GEORGETOWN UNIVERSITY

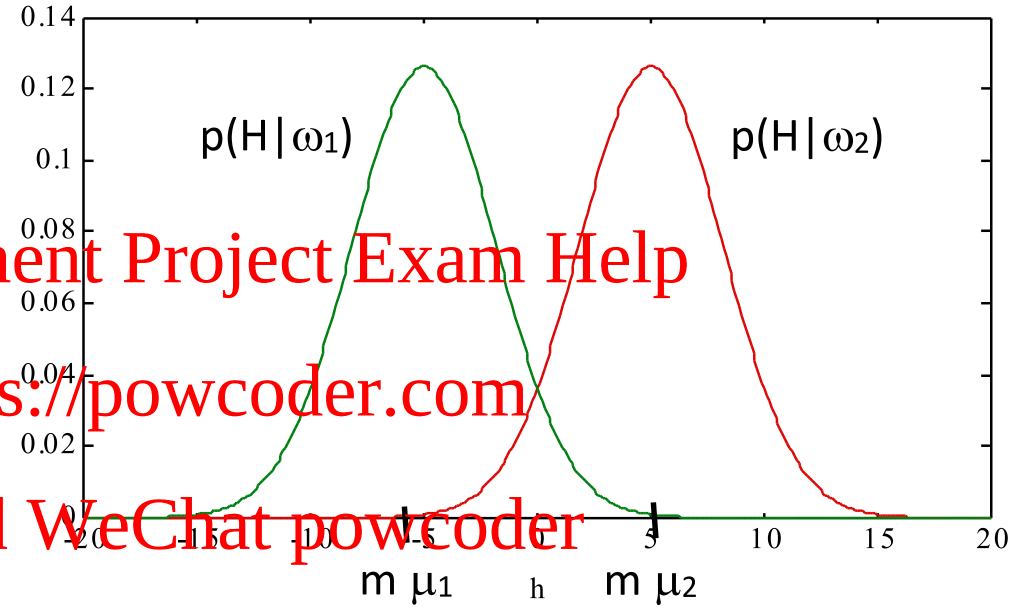
Sequential Hypothesis Test

We have

$$\mu_1 \leq 0, \quad \mu_2 \geq 0$$

$$E\left[H \mid \omega_i\right] = m \ \mu_i$$

$$\operatorname{var}\left[H \mid \omega_i\right] = m \ \sigma_i^2$$



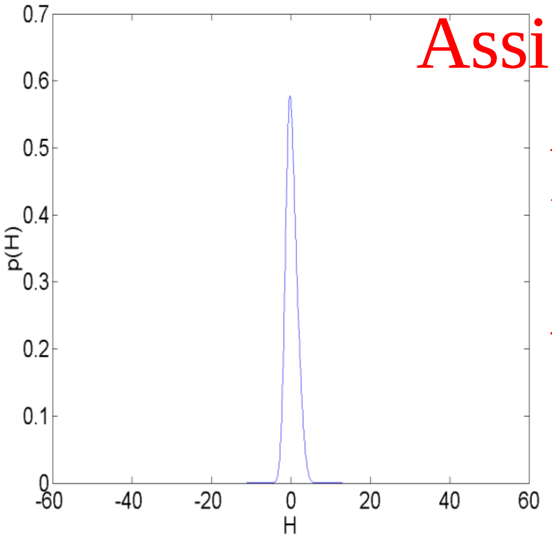
A convenient measure of separation between the two classes is

$$\frac{E\left[H \mid \omega_2\right] - E\left[H \mid \omega_1\right]}{\sqrt{\operatorname{var}\left[H \mid \omega_2\right] + \operatorname{var}\left[H \mid \omega_1\right]}} = \sqrt{m} \ \frac{\mu_2 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \longrightarrow$$

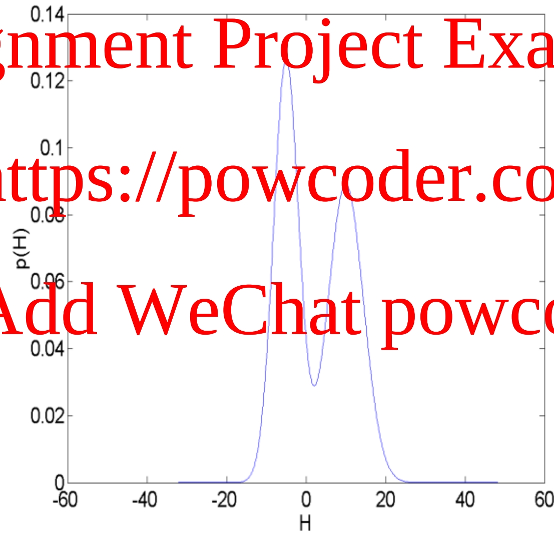
Separation increases with increasing number of observations as $m^{1/2}$.

Sequential Hypothesis Tests

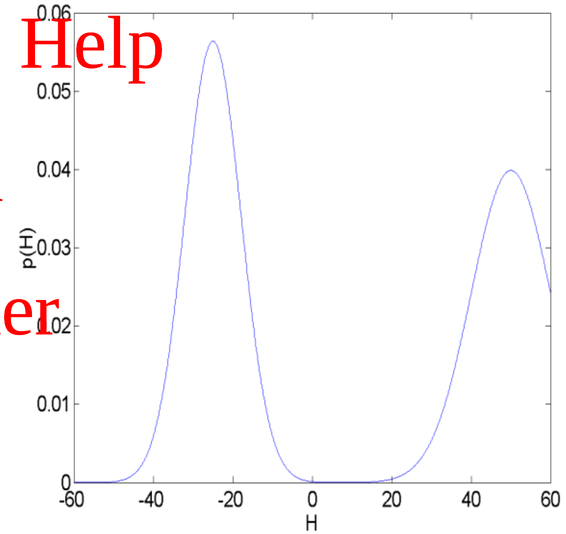
m=1



m=10



m=50



Wald Test for Sequential Observations

$$H_m \equiv \sum_{k=1}^{m} h(x_k) \qquad E[H \mid \omega_1] = m\,\mu_1 < 0$$

$$E[H \mid \omega_2] = m\,\mu_2 > 0$$

Terminate sequence of observations when H reaches some threshold -- e.g. when

$$H_m \leq a \quad choose \;\; \omega_1$$

$$or$$

$$H_m \geq b \quad choose \;\; \omega_2$$

otherwise, continue gathering measurements.

Wald Test

- Wald showed that
  - Error rates: When h(x) is small

$$\mathcal{E}_1 \approx \left(\frac{B(A-1)}{A-B}\right), \quad \mathcal{E}_2 \approx \left(\frac{1-B}{A-B}\right)$$

$$A \equiv e^{-a}, \quad B \equiv e^{-b}$$

  - Average sequence length to reach threshold is

$$E[m \mid \omega_1] = \frac{a(1-\mathcal{E}_1) + b\,\mathcal{E}_1}{\mu_1}$$

$$E[m \mid \omega_2] = \frac{a\,\mathcal{E}_2 + b(1-\mathcal{E}_2)}{\mu_2}$$

17 89

GEORGETOWN UNIVERSITY