ANLY-601
Advanced Pattern Recognition

Spring 2018

L3-4 – Bayes Classifiers (cont'd)

GEORGETOWN UNIVERSITY

General Cost

- Suppose each of the two classification error types have different cost. What's the ideal decision strategy?

  e.g. In a detection problem (medical, fire alarm), false negatives may be much more costly than false positives.

- Define an average "loss" (or cost) function and devise a decision rule that minimizes it.

Loss Function

Cost incurred for choosing class ω_i when ω_j is the actual class

$$\lambda_{ij}$$

Function to minimize is average cost or "risk"

$$\begin{aligned} R = {} & \lambda_{11} P(choose\ \omega_1 | \omega_1\ is\ true)\ P_1 \\ & + \lambda_{22} P(choose\ \omega_2 | \omega_2\ is\ true)\ P_2 \\ & + \lambda_{12} P(choose\ \omega_1 | \omega_2\ is\ true)\ P_2 \\ & + \lambda_{21} P(choose\ \omega_2 | \omega_1\ is\ true)\ P_1 \end{aligned}$$

Loss Function

Now

$$\lambda_{12}\, P(choose\ \omega_1 | \omega_2\ is\ true\ )\, P_2 \ = \ \lambda_{12} \int_{L1} p(x \mid \omega_2\ )\, P_2\ d^n x$$

and similarly for the other terms

So

$$R \ = \int_{L1} [\ \lambda_{12}\, p(x \mid \omega_2\ )\, P_2 + \lambda_{11}\, p(x \mid \omega_1)\, P_1]\, d^n x$$

$$+ \int_{L2} [\ \lambda_{21}\, p(x \mid \omega_1\ )\, P_1 + \lambda_{22}\, p(x \mid \omega_2)\, P_2]\, d^n x$$

Re-write Loss

$$R = \int_{L1} \left[\lambda_{12}\, p(x \mid \omega_2)\, P_2 + \lambda_{11}\, p(x \mid \omega_1)\, P_1 \right] d^n x$$

$$+ \int_{L2} \left[\lambda_{21}\, p(x \mid \omega_1)\, P_1 + \lambda_{22}\, p(x \mid \omega_2)\, P_2 \right] d^n x$$

Note that

$$\int_{L2} p(x \mid \omega_i)\, d^n x = 1 - \int_{L1} p(x \mid \omega_i)\, d^n x$$

so $$R = \lambda_{21} P_1 + \lambda_{22} P_2$$

$$+ \int_{L1} \left[(\lambda_{11} - \lambda_{21})\, p(x \mid \omega_1) P_1 + (\lambda_{12} - \lambda_{22})\, p(x \mid \omega_2) P_2 \right] d^n x$$

Minimum Loss

$$R = \lambda_{21} P_1 + \lambda_{22} P_2$$

$$+ \int_{L1} \left[\left(\lambda_{11} - \lambda_{21} \right) p\left(x \mid \omega_1 \right) P_1 - \left(\lambda_{22} - \lambda_{12} \right) p\left(x \mid \omega_2 \right) P_2 \right] \; d^n x$$

To minimize this, we want the integrand to be as negative as possible, so

$$\left(\lambda_{22} - \lambda_{12} \right) \, p\left(x \mid \omega_2 \right) P_2 \; > \; \left(\lambda_{11} - \lambda_{21} \right) \, p\left(x \mid \omega_1 \right) P_1 \qquad \forall x \in L_1$$

This tells us how to assign each x to either L1 or L2

Minimum Loss

$$\left(\lambda_{22}-\lambda_{12}\right)\, p\left(x \mid \omega_2\right) P_2 \;>\; \left(\lambda_{11}-\lambda_{21}\right)\, p\left(x \mid \omega_1\right) P_1 \qquad \forall x \in L_1$$

Or (multiply both sides by -1, and change reverse inequality)

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

another likelihood ratio test.

Neyman-Pearson Test

Suppose we don’t know the cost of each type of error, or the priors. How do we proceed?

Can only work with conditional errors

$$P\left(choose\ \omega_1 \mid \omega_2\ is\ true\right) = \int_{L1} p(x \mid \omega_2)\, d^n x \equiv \mathcal{E}_2$$

$$P\left(choose\ \omega_2 \mid \omega_1\ is\ true\right) = \int_{L2} p(x \mid \omega_1)\, d^n x \equiv \mathcal{E}_1$$

A way to proceed is to minimize E_1 subject to some specified acceptable $E_2 = E_o$.

This is a <u>constrained</u> minimization problem that uses the <u>Lagrange multiplier</u> formulation.

Neyman-Pearson Test

We want to minimize E1 subject to the constraint E2=Eo. The Lagrangian (the function to minimize) is

$$r = \mathcal{E}_1 + \lambda \left[\mathcal{E}_2 - \mathcal{E}_0 \right]$$

$$= \int_{L_2} p(x|\omega_1)\, d^n x + \lambda \left[\left(\int_{L_1} p(x|\omega_2)\, d^n x \right) - \mathcal{E}_0 \right]$$

Neyman-Pearson

Re-write the Lagrangian

$$r = \int_{L2} p(x|\omega_1)\; d^n x \; + \lambda \left[\left(\int_{L1} p(x|\omega_2)\, d^n x \right) - \mathcal{E}_0 \right]$$

$$= \left(1 - \lambda\, \mathcal{E}_0\right) + \int_{L1} \left(\lambda\, p(x|\omega_2) - p(x|\omega_1) \right) d^n x$$

r will be minimized when the integrand is kept negative, so the decision rule is

$$\boxed{\frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \lambda}$$

another likelihood ratio test.

Neyman-Pearson Test

- What's the threshold λ? It's set by requiring that the constraint be satisfied

$$\left(\int_{L1} p(x|\omega_2)\, d^n x \right) = \mathcal{E}_0$$

Hypothesis Test
Continued

Minimax Test --

We've been using likelihood ratio tests like

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

but what happens if the priors change after the test is designed?

One approach - construct test so that its performance is no worse than the worst possible Bayes test.

Minimax Test

Rewrite the expected loss, using $P_1 + P_2 = 1$

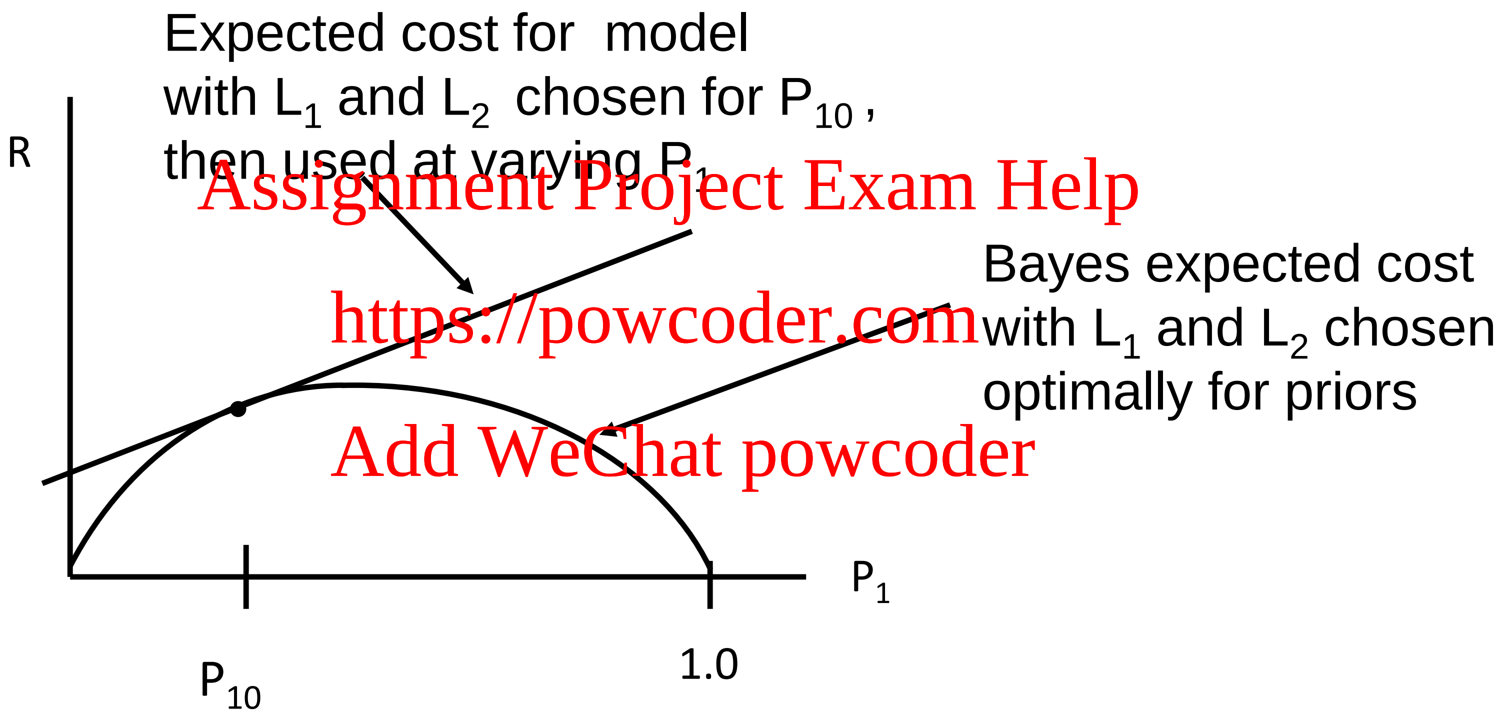
$$R = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{L_1} p(x \,|\, \omega_1)\, d^n x$$

$$+ P_1 \left\{ (\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{L_2} p(x \,|\, \omega_1)\, d^n x - (\lambda_{12} - \lambda_{22}) \int_{L_1} p(x \,|\, \omega_2)\, d^n x \right\}$$

eqn (***)

What's this look like as a function of P_1 ?

Minimax Test

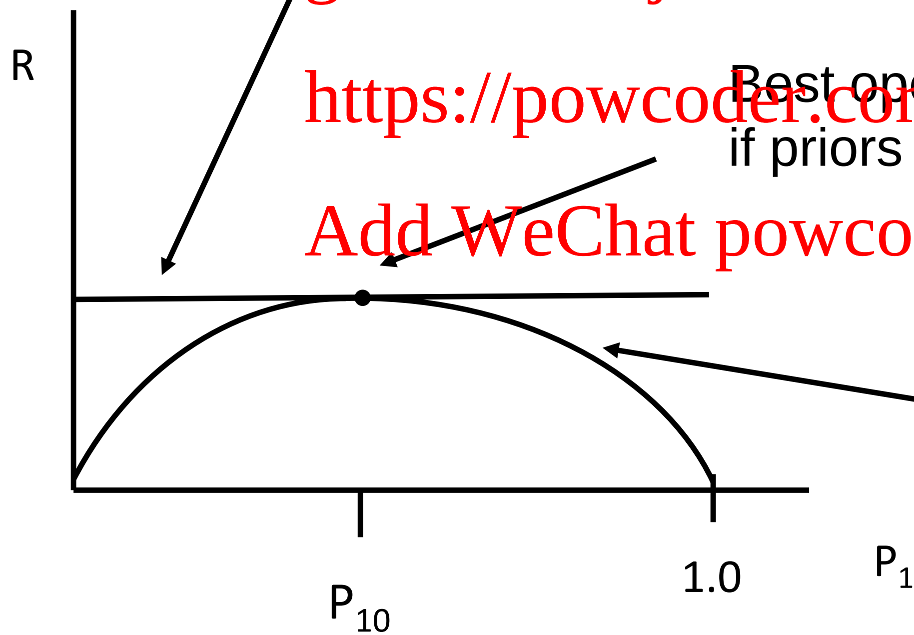


Minimax Test

Expected cost for model with L_1 and L_2 chosen for P_{10} , then used at varying P_1

Performance is no worse than worst Bayes performance

Best operating point if priors are not known



Bayes expected cost with L_1 and L_2 chosen optimally for priors

Want dR/dP_1 for CONSTANT L_1 to be zero.

Minimax Test

From eqn(***)

$$\left.\frac{dR}{dP_1}\right|_{Fixed\ L_1,\,L_2} = 0$$

$$\longrightarrow (\lambda_{21}-\lambda_{11})\,\mathcal{E}_1 = (\lambda_{11}-\lambda_{22}) + (\lambda_{12}-\lambda_{22})\,\mathcal{E}_2$$

For example, if the cost for each kind of correct decision are the same, and the costs for each type of error are the same

$$\lambda_{11} = \lambda_{22} \equiv \lambda_c$$

$$\lambda_{12} = \lambda_{21} \equiv \lambda_I$$

and the above condition becomes $\boxed{\mathcal{E}_1 = \mathcal{E}_2}$ choose the operating point that gives equal rates for both kinds of error.

ROC Curves

- Likelihood ratio tests are threshold tests, with the threshold defined by the priors, and the decision costs. As the priors and decision costs change, the threshold changes and the rate of each kind of error changes.
- The Receiver Operating Characteristic (or ROC) curve shows the system behavior over the full range of thresholds.
- The ROC is determined only by the class conditional probability distributions for the measured features.

ROC Curves

Recall the error rates

$$\mathcal{E}_j = \mathrm{P}(\,choose\;\omega_{i\neq j}\,|\,\omega_j\;is\;true\,)$$
$$= \int_{L_{i\neq j}} p(x\,|\,\omega_j)\,d^n x$$

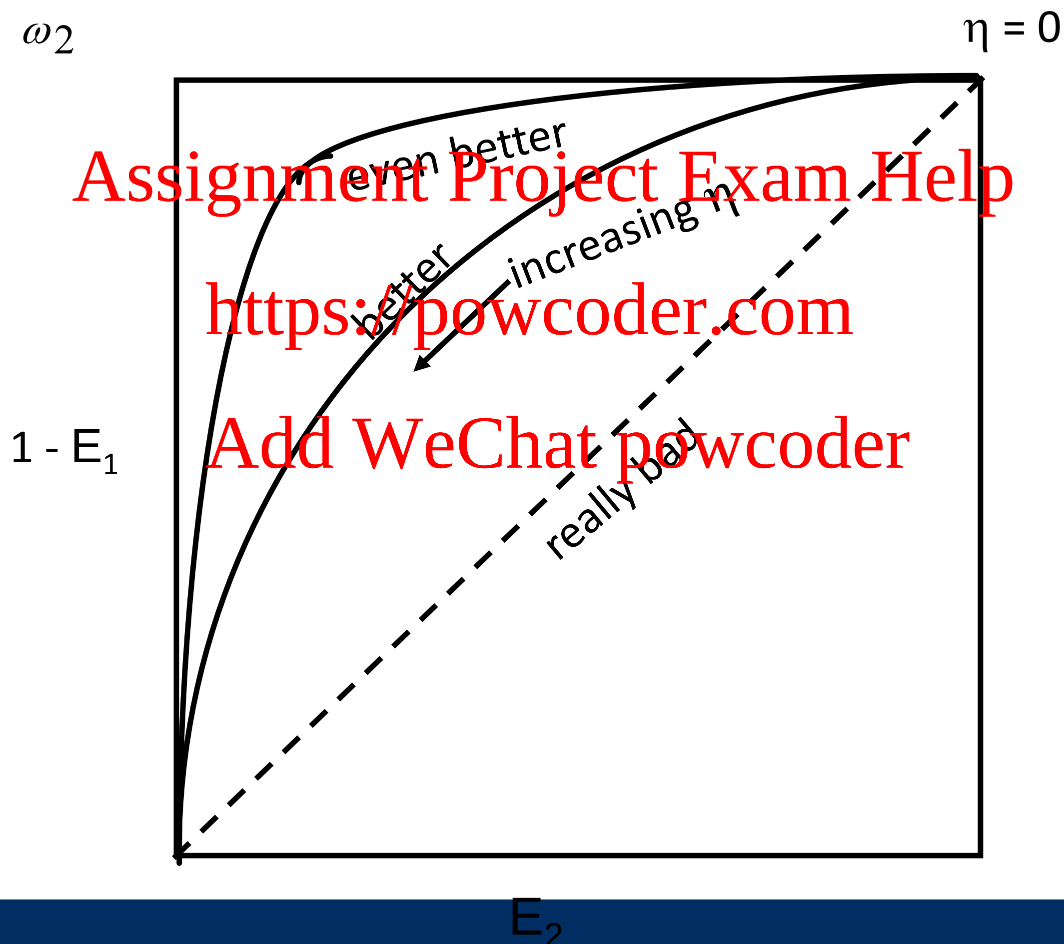
where the different regions L1 and L2 are defined by the likelihood ratio test

$$\frac{p(x\,|\,\omega_1)}{p(x\,|\,\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta$$

To display the system performance at a glance, we'll plot the error rates as a function of threshold. This is the ROC curve.

ROC Curves

$$\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta$$



ROC Curves

- Concave downward
- ROCs are above the diagonal
- Slope = threshold

GEORGETOWN UNIVERSITY

Log-Likelihood

Sometimes it's more convenient to work with the conditional distribution of the negative log-likelihood than the conditional distribution of the features x.

Recall the likelihood ratio test with threshold η

$$\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \eta$$

In terms of negative log-likelihood $h = -\log \dfrac{p(x \mid \omega_1)}{p(x \mid \omega_2)}$

$$h \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} -\log \eta$$

x is a random vector, so h is a random scalar, and has distribution

$p(h \mid \omega_i)$ when ω_i is true

Log Likelihood

We can rewrite the error probabilities in terms of integrals over the distribution for h

$$h \;\underset{\omega_2}{\overset{\omega_1}{\lessgtr}}\; -\log \eta$$

$$E_1 = \int_{L_2} p(x \mid \omega_1)\; d^n x = \int_{-\log \eta}^{\infty} p(h \mid \omega_1)\; dh$$

$$E_2 = \int_{L_1} p(x \mid \omega_2)\; d^n x = \int_{-\infty}^{-\log \eta} p(h \mid \omega_2)\; dh$$

GEORGETOWN UNIVERSITY