Bayesian decision theory

Nuno Vasconcelos

ECE Department, UCSD

Bayesian decision theory

- recall that we have
  - Y – state of the world
  - X – observations
  - g(x) – decision function
  - L[g(x),y] – loss of predicting y with g(x)
- the expected value of the loss is called the risk

$$\boxed{Risk = E_{X,Y}\left[L(X,Y)\right]}$$

  - which can be written as

$$Risk = \int \sum_{i=1}^{M} P_{Y,X}(i,x) L[g(x),i] dx$$

Bayesian decision theory

- from this

$$Risk = \int \sum_{i=1}^{M} P_{Y,X}(i,x)L[g(x),i]dx$$

- by chain rule

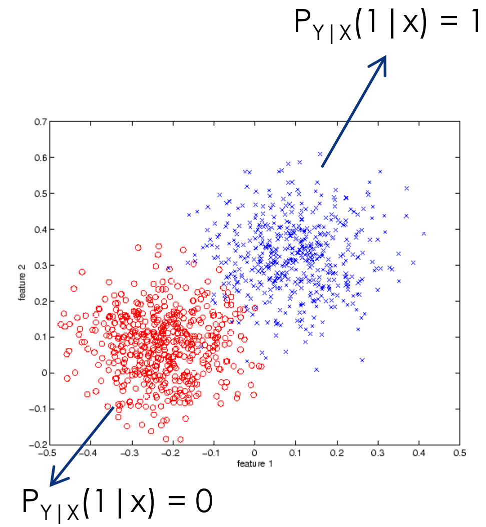
$$Risk = \int P_X(x) \sum_{i=1}^{M} P_{Y|X}(i\,|\,x)L[g(x),i]dx$$

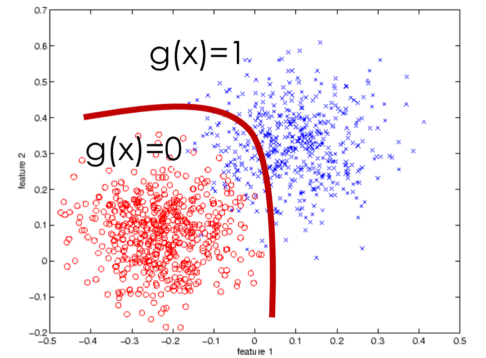
$$= \int P_X(x)R(x)dx = E_X[R(x)]$$

- where

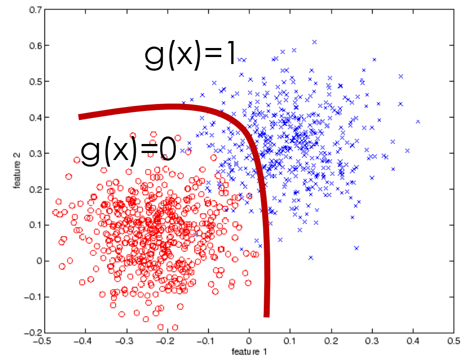
$$\boxed{R(x) = \sum_{i=1}^{M} P_{Y|X}(i\,|\,x)L[g(x),i]}$$

- is the conditional risk, given the observation x





Bayesian decision theory



- since, by definition,

$$L[g(x), i] \geq 0, \quad \forall x, y$$

- it follows that

$$R(x) = \sum_{i=1}^{M} P_{Y|X}(i \mid x) L[g(x), i] \geq 0, \quad \forall x$$

- Also, the

$$\boxed{Risk = E_X[R(X)]}$$

is minimum if we minimize R(x) at all x, i.e., if we use pick the decision function

$$\boxed{g^*(x) = \arg\min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i \mid x) L[g(x), i]}$$

Bayesian decision theory

- this is the Bayes decision rule

$$g^*(x) = \arg\min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i|x)L[g(x),i]$$

- the associated risk

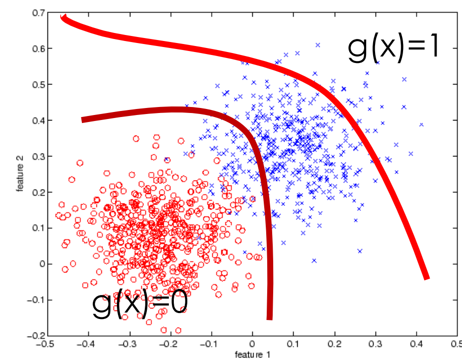
$$R^* = \int \sum_{i=1}^{M} P_{Y,X}(i,x)L[g^*(x),i]dx$$

- or

$$R^* = \int P_X(x) \sum_{i=1}^{M} P_{Y|X}(i|x)L[g^*(x),i]dx$$

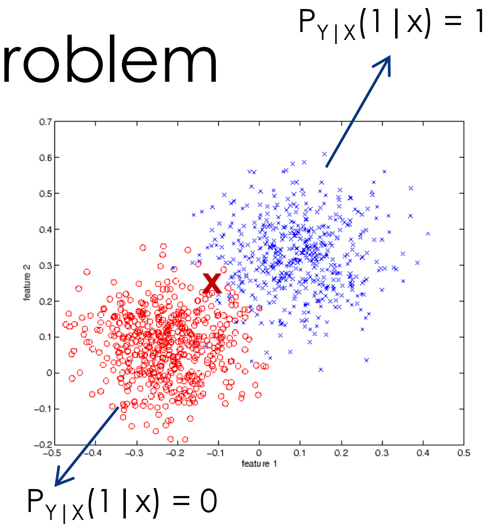
- is the Bayes risk, and cannot be beaten

which decision rule has smaller conditional risk?



Example

- let's consider a binary classification problem



$$g^*(x) \in \{0,1\}$$

- for which the conditional risk is

$$R(x) = \sum_{i=0}^{1} P_{Y|X}(i|x) L[g(x), i]$$

$$= P_{Y|X}(0|x) L[g(x), 0] + P_{Y|X}(1|x) L[g(x), 1]$$

- we have two options

$$g(x) = 0 \Rightarrow R_0(x) = P_{Y|X}(0|x) L[0,0] + P_{Y|X}(1|x) L[0,1]$$

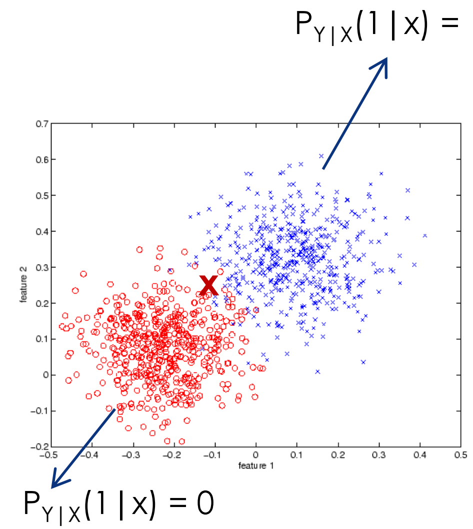
$$g(x) = 1 \Rightarrow R_1(x) = P_{Y|X}(0|x) L[1,0] + P_{Y|X}(1|x) L[1,1]$$

- and should pick the one of smaller conditional risk

Example

- i.e. pick g(x) = 0 if $R_0(x) < R_1(x)$ and g(x)=1 otherwise
- this can be written as, pick 0 if

$$P_{Y|X}(0|x)L[0,0] + P_{Y|X}(1|x)L[0,1] <$$
$$< P_{Y|X}(0|x)L[1,0] + P_{Y|X}(1|x)L[1,1]$$



- or

$$P_{Y|X}(0|x)\{L[0,0] - L[1,0]\} <$$
$$< P_{Y|X}(1|x)\{L[1,1] - L[0,1]\}$$

- usually there is no loss associated with the correct decision

$$L[1,1] = L[0,0] = 0$$

- and this is the same as

$$P_{Y|X}(0|x)L[1,0] > P_{Y|X}(1|x)L[0,1]$$

$$P_{Y|X}(y|x) = \frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)}$$

Example

- or, "pick 0" if

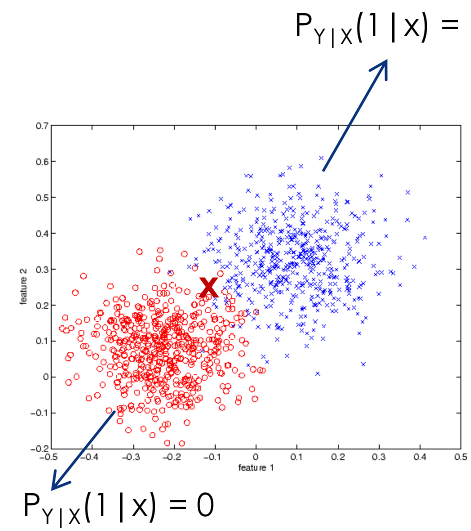
$$\frac{P_{Y|X}(0|x)}{P_{Y|X}(1|x)} > \frac{L[0,1]}{L[1,0]}$$

- and applying Bayes rule

$$\frac{P_{X|Y}(x|0)P_Y(0)}{P_{X|Y}(x|1)P_Y(1)} > \frac{L[0,1]}{L[1,0]}$$

- which is equivalent to "pick 0" if

$$\frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} > T^* = \frac{L[0,1]P_Y(1)}{L[1,0]P_Y(0)}$$



  - i.e. we pick 0, when the probability of X given that Y=0 divided by that given Y=1 is greater than a threshold
  - the optimal threshold T* depends on the costs of the two types of error and the probabilities of the two classes

BDR for 0-1 loss

- let's consider the "0-1" loss

$$L[g(x), y] = \begin{cases} 1, & g(x) \neq y \\ 0, & g(x) = y \end{cases}$$

  - in this case the optimal decision function is

$$\begin{aligned} g^*(x) &= \arg\min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i|x) L[g(x), i] \\ &= \arg\min_{g(x)} \sum_{i \neq g(x)} P_{Y|X}(i|x) \\ &= \arg\min_{g(x)} \left[1 - P_{Y|X}(g(x)|x)\right] \\ &= \arg\max_{g(x)} P_{Y|X}(g(x)|x) \\ &= \arg\max_{i} P_{Y|X}(i|x) \end{aligned}$$

BDR for 0-1 loss

- for the "0-1" loss the optimal decision rule is the maximum a-posteriori probability rule

$$g^*(x) = \arg\max_i P_{Y|X}(i|x)$$

- what is the associated risk?

$$\begin{aligned}
R^* &= \int P_X(x) \sum_{i=1}^{M} P_{Y|X}(i|x) L[g^*(x), i] dx \\
&= \int P_X(x) \sum_{i \neq g^*(x)}^{M} P_{Y|X}(i|x) dx \\
&= \int P_X(x) P_{Y|X}(y \neq g^*(x)|x) dx \\
&= \int P_{Y,X}(y \neq g^*(x), x) dx
\end{aligned}$$

BDR for 0-1 loss

- but

$$\boxed{R^* = \int P_{Y,X}(y \neq g^*(x), x)dx}$$

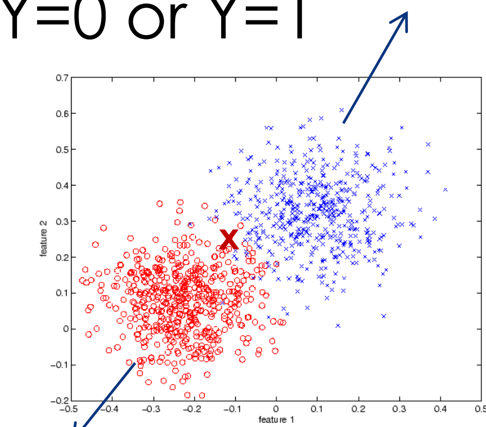
  - is just the probability of error of the decision rule g*(x)
  - note that the same result would hold for any g(x), i.e. R would be the probability of error of g(x)
  - this implies the following
- for the "0-1" loss
  - the Bayes decision rule is the MAP rule

$$\boxed{g^*(x) = \arg\max_i P_{Y|X}(i \mid x)}$$

  - the risk is the probability of error of this rule (Bayes error)
  - there is no other decision function with lower error

MAP rule

- usually can be written in a simple form given a probabilistic model for X and Y
- consider the two-class problem, i.e. Y=0 or Y=1
  - the BDR is



$$\boxed{\begin{aligned} i^*(x) &= \arg\max_i P_{Y|X}(i|x) \\ &= \begin{cases} 0, & \text{if } P_{Y|X}(0|x) \geq P_{Y|X}(1|x) \\ 1, & \text{if } P_{Y|X}(0|x) < P_{Y|X}(1|x) \end{cases} \end{aligned}}$$

  - pick "0" when $P_{Y|X}(0|x) \geq P_{Y|X}(1|x)$ and "1" otherwise
  - using Bayes rule $P_{Y|X}(0|x) \geq P_{Y|X}(1|x) \Leftrightarrow$

$$\frac{P_{X|Y}(x|0)P_Y(0)}{P_X(x)} \geq \frac{P_{X|Y}(x|1)P_Y(1)}{P_X(x)}$$

MAP rule

- noting that $P_X(x)$ is a non-negative quantity this is the same as
- pick "0" when

$$P_{X|Y}(x|0)P_Y(0) \geq P_{X|Y}(x|1)P_Y(1)$$

- by using the same reasoning, this can be easily generalized to

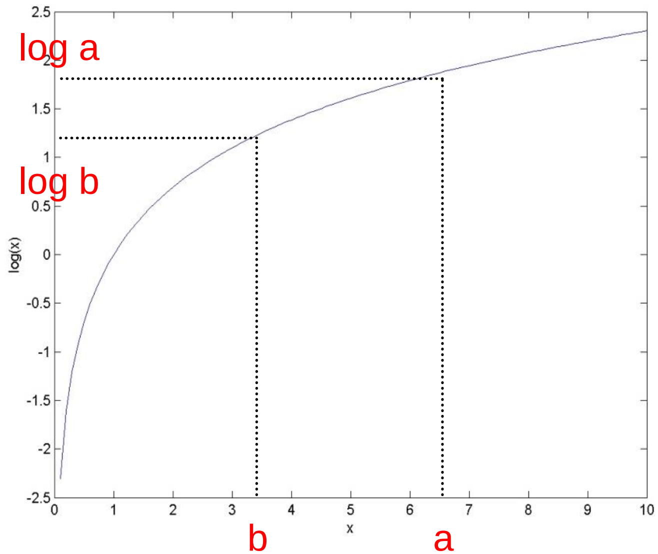
$$\boxed{i^*(x) = \arg\max_i P_{X|Y}(x\,|\,i)P_Y(i)}$$

- note that:
  - many class-conditional distributions are exponential (e.g. the Gaussian)
  - this product can be tricky to compute (e.g. the tail probabilities are quite small)
  - we can take advantage of the fact that we only care about the order of the terms on the right-hand side

The log trick

- this is the log trick
  - which is to take logs
  - note that the log is a monotonically increasing function

$$\boxed{a > b \Leftrightarrow \log a > \log b}$$



  - from which

$$\boxed{\begin{aligned} i^*(x) &= \arg\max_i P_{X|Y}(x\,|\,i)P_Y(i) \\ &= \arg\max_i \log\left(P_{X|Y}(x\,|\,i)P_Y(i)\right) \\ &= \arg\max_i \log P_{X|Y}(x\,|\,i) + \log P_Y(i) \end{aligned}}$$

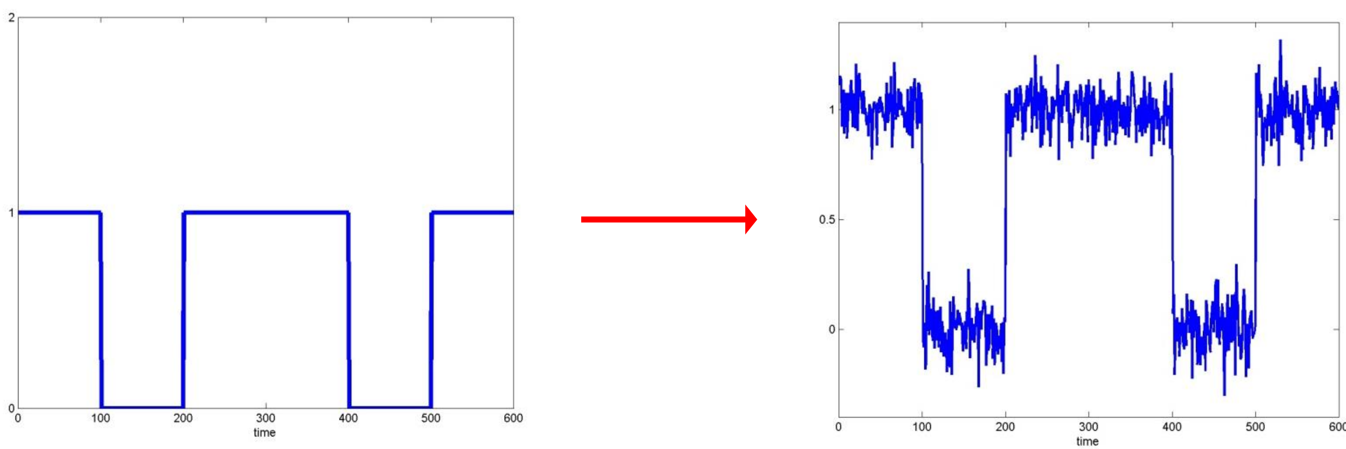
  - the order is preserved

MAP rule

- in summary
  - for the zero/one loss, the following three decision rules are
  - optimal and equivalent
  - 1) $\boxed{i^*(x) = \arg\max_i P_{Y|X}(i \mid x)}$
  - 2) $\boxed{i^*(x) = \arg\max_i \left[P_{X|Y}(x \mid i) P_Y(i)\right]}$
  - 3) $\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x \mid i) + \log P_Y(i)\right]}$
  - 1) is usually hard to use, 3) is frequently easier than 2)

Example

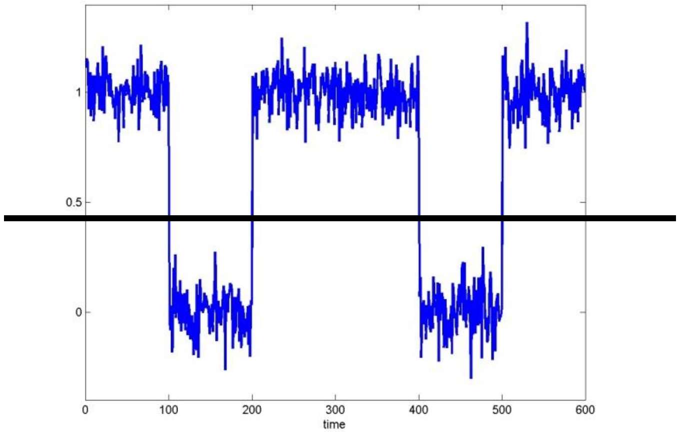
- the Bayes decision rule is usually highly intuitive
- example: communications
  - a bit is transmitted by a source, corrupted by noise, and received by a decoder

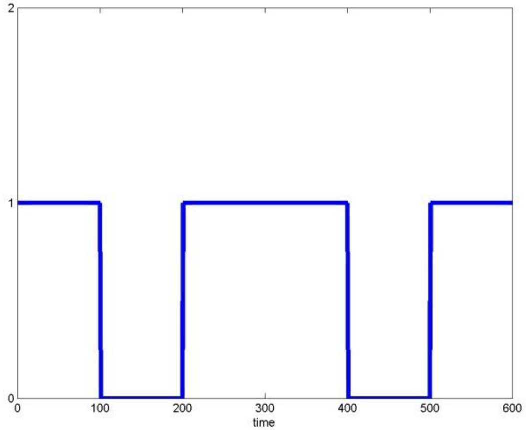


  - Q: what should the optimal decoder do to recover Y?

Example

- intuitively, it appears that it should just threshold X
  - pick T
  - decision rule $$\boxed{Y = \begin{cases} 0, & \text{if } x < \mathrm{T} \\ 1, & \text{if } x > \mathrm{T} \end{cases}}$$





  - what is the threshold value?
  - let's solve the problem with the BDR

Example

- we need
  - class probabilities:
    - in the absence of any other info let's say

$$\boxed{P_Y(0) = P_Y(1) = 1/2}$$

  - class-conditional densities:
    - noise results from thermal processes, electrons moving around and bumping each other
    - a lot of independent events that add up
    - by the central limit theorem it appears reasonable to assume that the noise is Gaussian
- we denote a Gaussian random variable of mean μ and variance σ^2 by

$$\boxed{X \sim N(\mu, \sigma^2)}$$

Example

- the Gaussian probability density function is

$$\boxed{P_X(x) = G(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

- since noise is Gaussian, and assuming it is just added to the signal we have

$$\boxed{X = Y + \varepsilon, \qquad \varepsilon \sim N(0,\sigma^2)}$$

  - in both cases, X corresponds to a constant (Y) plus zero-mean Gaussian noise
  - this simply adds Y to the mean of the Gaussian

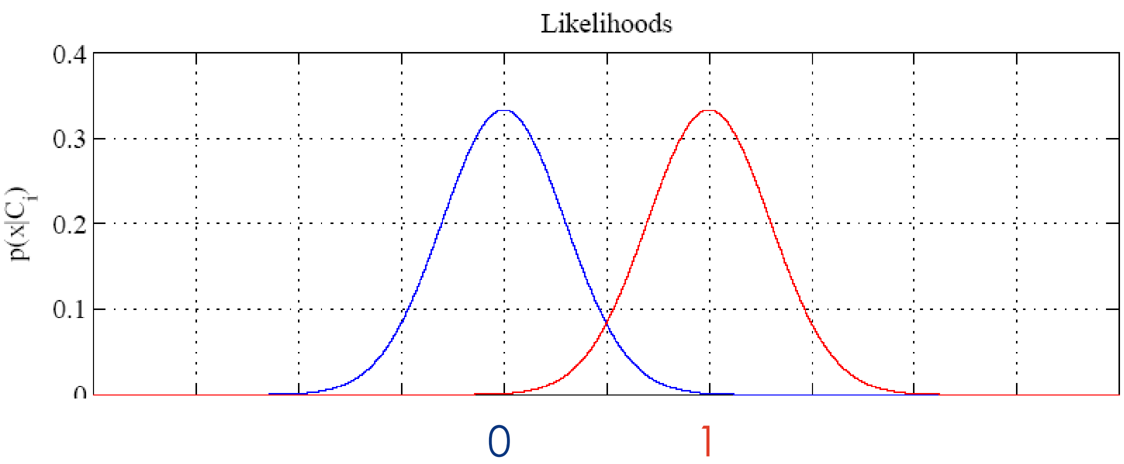
Example

- in summary

$$\boxed{\begin{aligned} P_{X|Y}(x\,|\,0) &= G(x,0,\sigma) \\ P_{X|Y}(x\,|\,1) &= G(x,1,\sigma) \end{aligned}}$$

$$\boxed{P_Y(0) = P_Y(1) = \tfrac{1}{2}}$$

  - or, graphically,



Example

- to compute the BDR, we recall that

$$\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x\,|\,i) + \log P_Y(i)\right]}$$

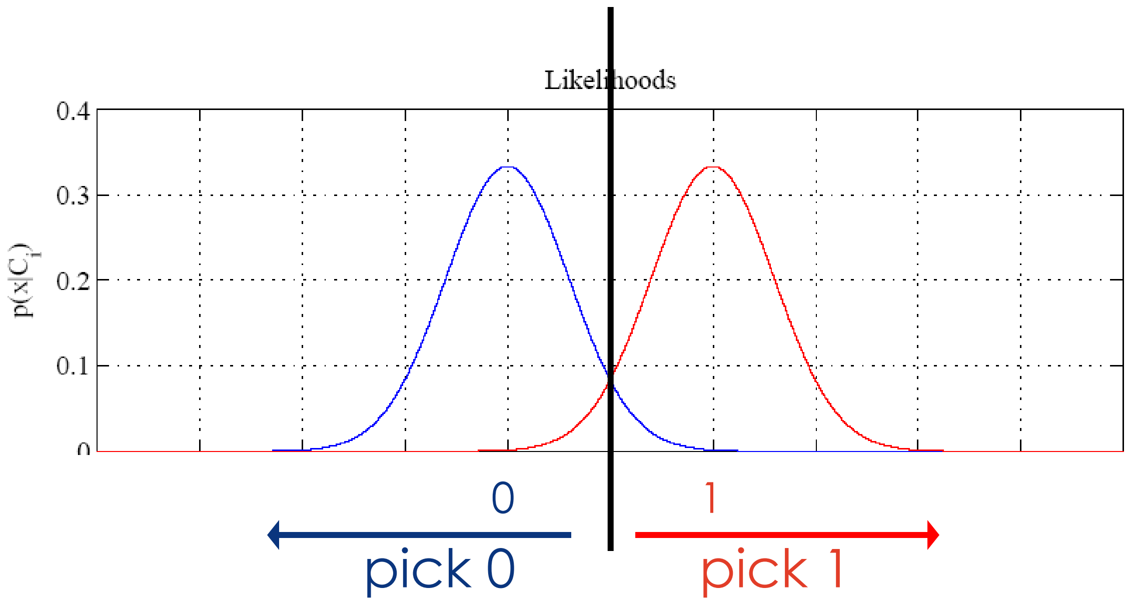
- and note that
  - terms which are constant (as a function of i) can be dropped
  - since we are just looking for the i that maximizes the function
  - since this is the case for the class-probabilities

$$P_Y(0) = P_Y(1) = \tfrac{1}{2}$$

  - we have $i^*(x) = \arg\max_i \log P_{X|Y}(x\,|\,i)$

BDR

- this is intuitive
  - we pick the class that "best explains" (gives higher probability) the observation
  - in this case, we can solve visually



  - but the mathematical solution is equally simple

BDR

- let's consider the more general case

$$\boxed{P_{X|Y}(x|0) = G(x, \mu_0, \sigma) \qquad P_{X|Y}(x|1) = G(x, \mu_1, \sigma)}$$

  - for which

$$\begin{aligned} i^*(x) &= \arg\max_i \log P_{X|Y}(x|i) \\ &= \arg\max_i \log\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}} \right\} \\ &= \arg\max_i \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu_i)^2}{2\sigma^2} \right\} \\ &= \arg\min_i \frac{(x-\mu_i)^2}{2\sigma^2} \end{aligned}$$

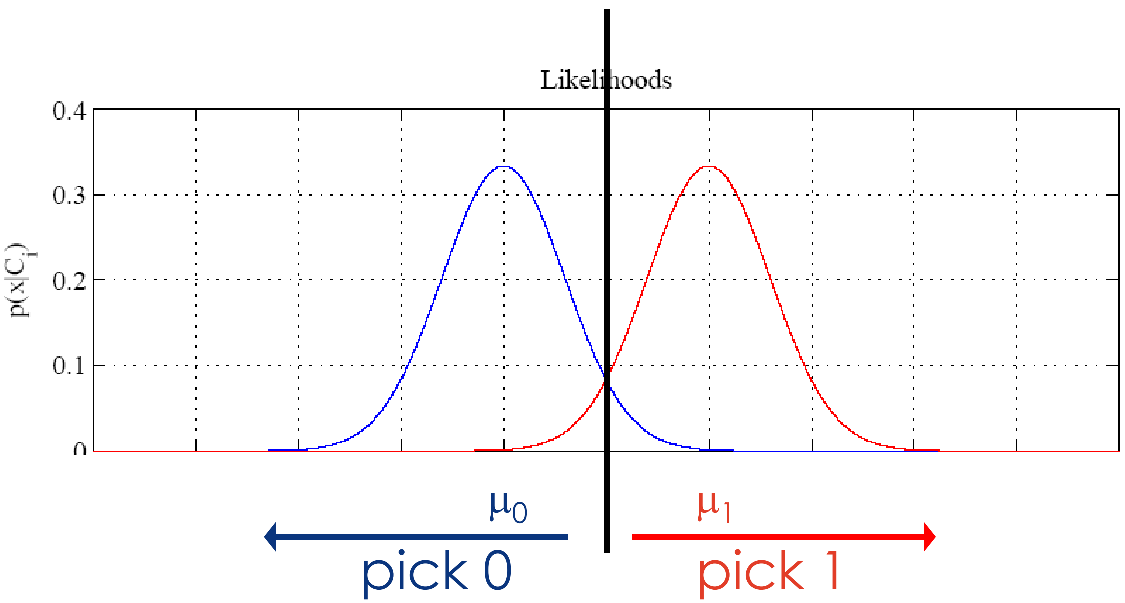
BDR

- or $$i^* = \arg\min_i \frac{(x-\mu_i)^2}{2\sigma^2}$$
$$= \arg\min_i (x^2 - 2x\mu_i + \mu_i^2)$$
$$= \arg\min_i (-2x\mu_i + \mu_i^2)$$

- the optimal decision is, therefore
  - pick 0 if
$$-2x\mu_0 + \mu_0^2 < -2x\mu_1 + \mu_1^2$$
$$2x(\mu_1 - \mu_0) < \mu_1^2 - \mu_0^2$$
  - or, pick 0 if
$$\boxed{x < \frac{\mu_1 + \mu_0}{2}}$$

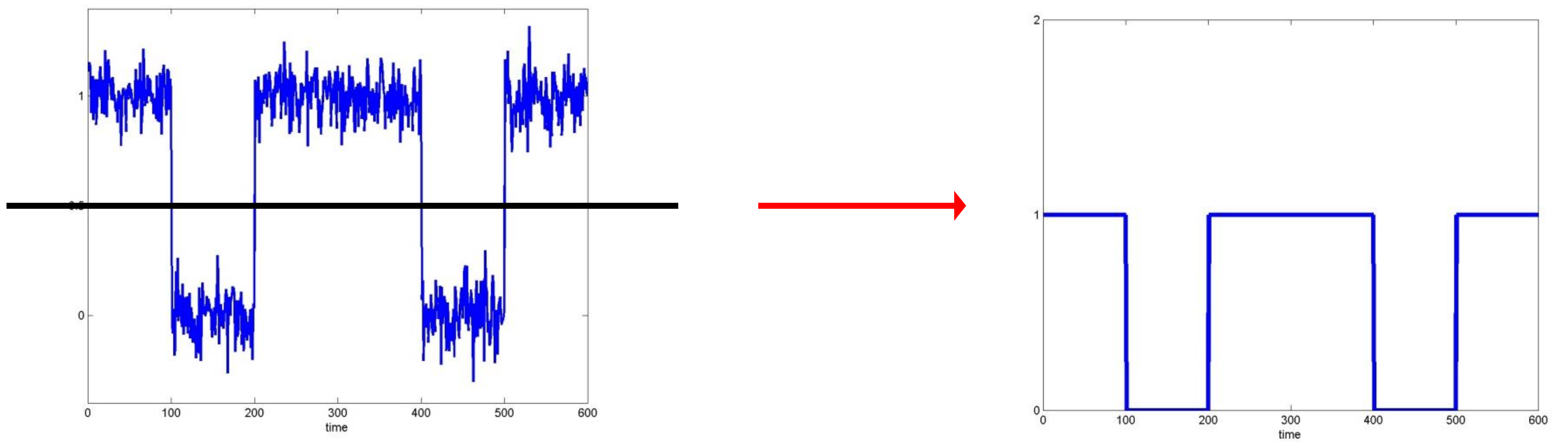
BDR

- for a problem with Gaussian classes, equal variances and equal class probabilities
  - optimal decision boundary is the threshold
  - at the mid-point between the two means



BDR

- back to our signal decoding problem
  - in this case T = 0.5
  - decision rule $$\boxed{Y = \begin{cases} 0, & \text{if } x < 0.5 \\ 1, & \text{if } x > 0.5 \end{cases}}$$



  - this is, once again, intuitive
  - we place the threshold midway along the noise sources

BDR

- what is the point of going through all the math?
  - now we know that the intuitive threshold is actually optimal, and in which sense it is optimal (minimum probability or error)
  - the Bayesian solution keeps us honest.
  - it forces us to make all our assumptions explicit
  - assumptions we have made
    - uniform class probabilities $\boxed{P_Y(0) = P_Y(1) = \tfrac{1}{2}}$
    - Gaussianity $\boxed{P_{X|Y}(x|i) = G(x, \mu_i, \sigma_i)}$
    - the variance is the same under the two states $\boxed{\sigma_i = \sigma, \forall i}$
    - noise is additive $\boxed{X = Y + \varepsilon}$
  - even for a trivial problem, we have made lots of assumptions

BDR

- what if the class probabilities are not the same?
  - e.g. coding scheme 7 = 11111110
  - in this case $P_Y(1) >> P_Y(0)$
  - how does this change the optimal decision rule?

$$i^*(x) = \arg\max_i \left\{ \log P_{X|Y}(x|i) + \log P_Y(i) \right\}$$

$$= \arg\max_i \left[\log\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}} \right\} + \log P_Y(i) \right]$$

$$= \arg\max_i \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu_i)^2}{2\sigma^2} + \log P_Y(i) \right\}$$

$$= \arg\min_i \left\{ \frac{(x-\mu_i)^2}{2\sigma^2} - \log P_Y(i) \right\}$$

BDR

- or $$i^* = \arg\min_i \left\{ \frac{(x-\mu_i)^2}{2\sigma^2} - \log P_Y(i) \right\}$$
$$= \arg\min_i (x^2 - 2x\mu_i + \mu_i^{\,2} - 2\sigma^2 \log P_Y(i))$$
$$= \arg\min_i (-2x\mu_i + \mu_i^{\,2} - 2\sigma^2 \log P_Y(i))$$
- the optimal decision is, therefore
  - pick 0 if
$$-2x\mu_0 + \mu_0^{\,2} - 2\sigma^2 \log P_Y(0) < -2x\mu_1 + \mu_1^{\,2} - 2\sigma^2 \log P_Y(1)$$
$$2x(\mu_1 - \mu_0) < \mu_1^{\,2} - \mu_0^{\,2} + 2\sigma^2 \log \frac{P_Y(0)}{P_Y(1)}$$
  - or, pick 0 if
$$\boxed{x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{P_Y(0)}{P_Y(1)}}$$

BDR

- what is the role of the prior for class probabilities?

$$\boxed{x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{P_Y(0)}{P_Y(1)}}$$

- the prior moves the threshold up or down, in an intuitive way
  - $P_Y(0)>P_Y(1)$: threshold increases
  - since 0 has higher probability, we care more about errors on the 0 side
  - by using a higher threshold we are making it more likely to pick 0
  - if $P_Y(0)=1$, all we care about is Y=0, the threshold becomes infinite
  - we never say 1
- how relevant is the prior?
  - it is weighed by

$$\boxed{1 \Big/ \frac{\mu_1 - \mu_0}{\sigma^2}}$$

BDR

- how relevant is the prior?
  - it is weighed by the inverse of the normalized distance between the means

$$\boxed{1 \Big/ \frac{\mu_1 - \mu_0}{\sigma^2}}$$

distance between the means in units of variance

  - if the classes are very far apart, the prior makes no difference
    - this is the easy situation, the observations are very clear, Bayes says "forget the prior knowledge"
  - if the classes are exactly equal (same mean) the prior gets infinite weight
    - in this case the observations do not say anything about the class, Bayes says "forget about the data, just use the knowledge that you started with"
    - even if that means "always say 0" or "always say 1"

Any questions?