The Gaussian classifier

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)
- Bayes decision rule is the rule that minimizes the risk

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

- given x, it consists of picking the prediction of minimum conditional risk

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

MAP rule

- for the "0-1" loss

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

- the optimal decision rule is the maximum a-posteriori probability rule

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

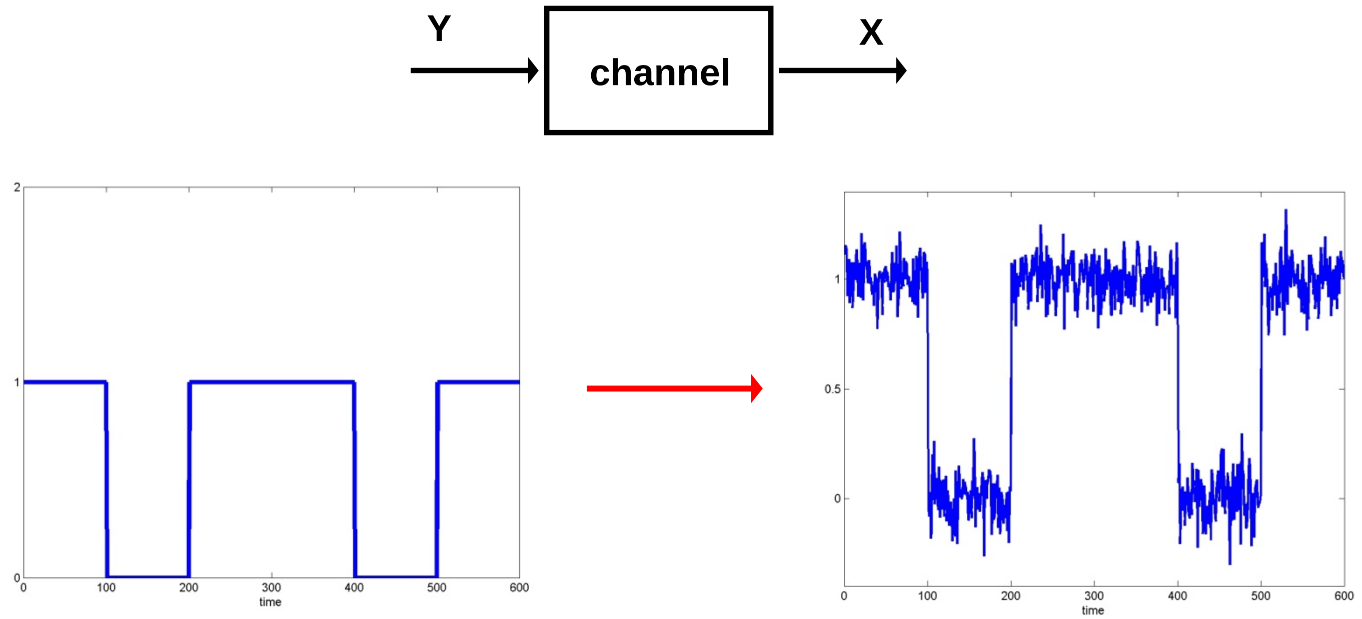
- the associated risk is the probability of error of this rule (Bayes error)
- there is no other decision function with lower error

MAP rule

- by application of simple mathematical laws (Bayes rule, monotonicity of the log)
- we have shown that the following three decision rules are optimal and equivalent
  - 1) $$\boxed{i^*(x) = \arg\max_i P_{Y|X}(i|x)}$$
  - 2) $$\boxed{i^*(x) = \arg\max_i \left[P_{X|Y}(x|i)P_Y(i)\right]}$$
  - 3) $$\boxed{i^*(x) = \arg\max_i \left[\log P_{X|Y}(x|i) + \log P_Y(i)\right]}$$
  - 1) is usually hard to use, 3) is frequently easier than 2)

Example

- the Bayes decision rule is usually highly intuitive
- we have used an example from 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

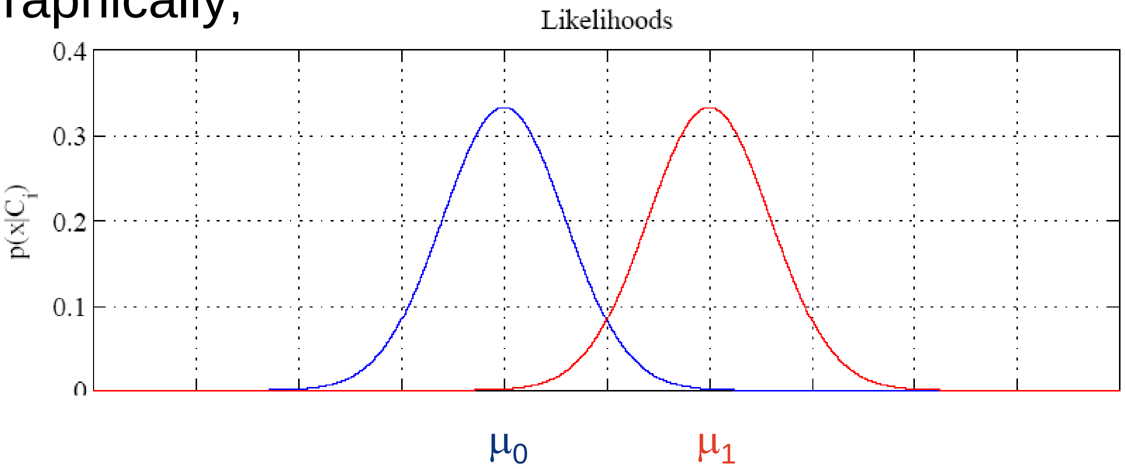
- this was modeled as a classification problem with Gaussian classes

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

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

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

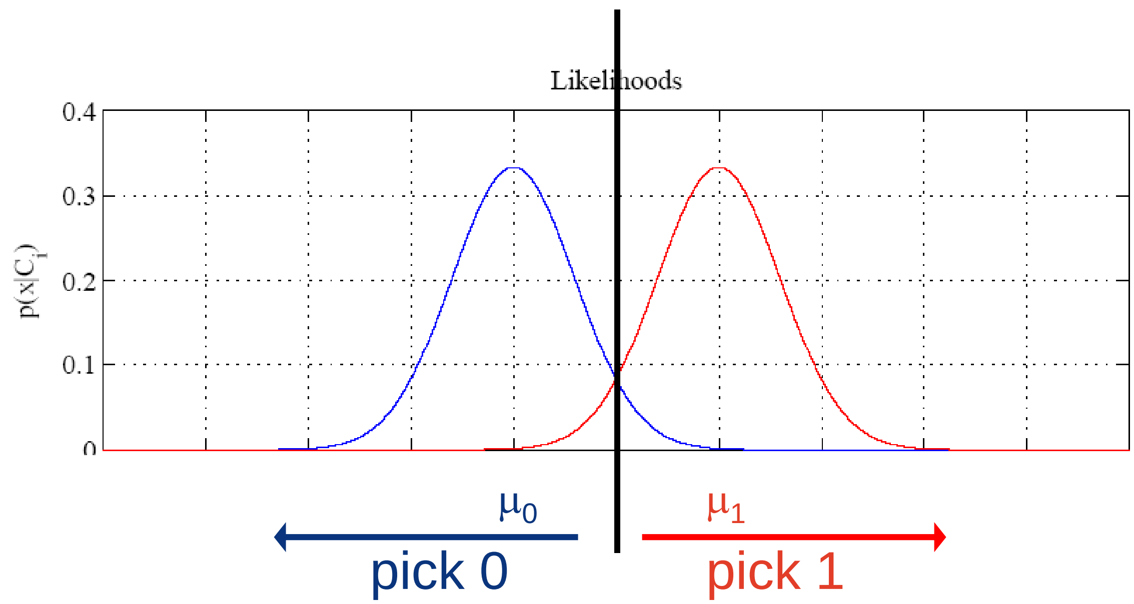
- or, graphically,



BDR

- for which the optimal decision boundary is a threshold
  - pick "0" if

$$\boxed{x < \frac{\mu_1 + \mu_0}{2}}$$



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

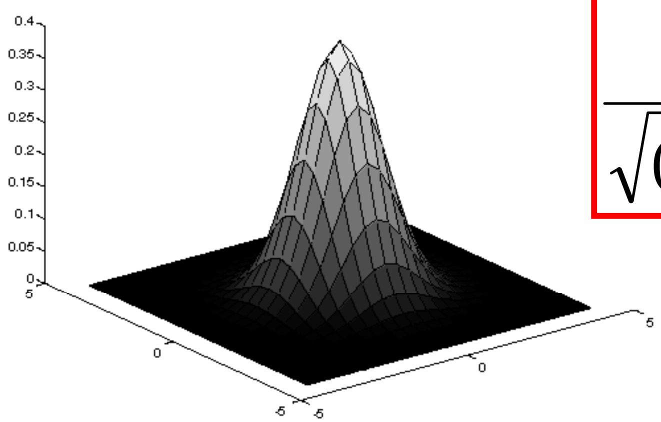
$$\boxed{1 \Big/ \frac{\mu_1 - \mu_0}{\sigma^2}}$$

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"

The Gaussian classifier

- this is one example of a Gaussian classifier
  - in practice we rarely have only one variable
  - typically X = (X_1, …, X_n) is a vector of observations
- the BDR for this case is equivalent, but more interesting
- the central different is the class-conditional distributions are multivariate Gaussian



$$P_{X|Y}(x\,|\,i) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma_i|}}\exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\}$$

The Gaussian classifier

- in this case

$$\boxed{P_{X|Y}(x|i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right\}}$$

- the BDR

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

- becomes

$$\boxed{\begin{aligned} i^*(x) = \arg\max_i \Bigg[& -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) \\ & -\frac{1}{2}\log(2\pi)^d |\Sigma_i| + \log P_Y(i) \Bigg] \end{aligned}}$$

The Gaussian classifier

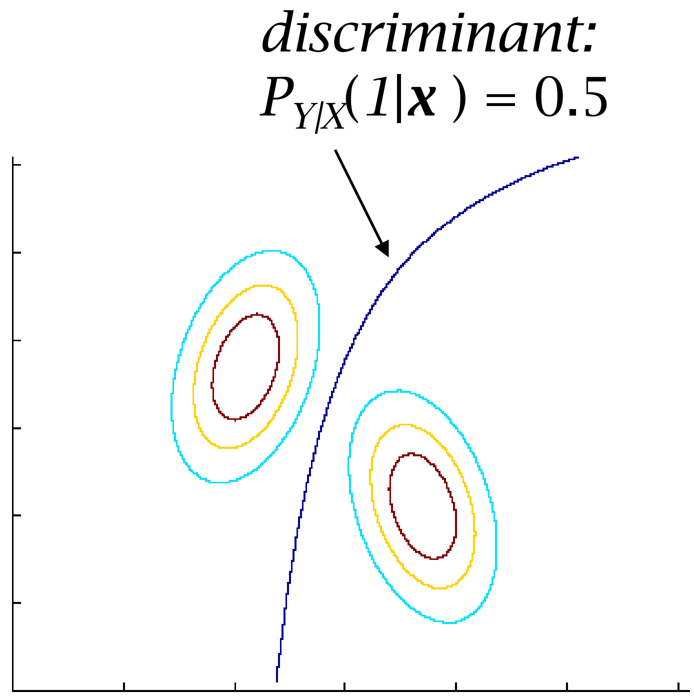
- this can be written as

$$\boxed{i^*(x) = \arg\min_i \left[d_i(x,\mu_i) + \alpha_i \right]}$$

with

$$\boxed{d_i(x,y) = (x-y)^T \Sigma_i^{-1} (x-y)}$$

$$\boxed{\alpha_i = \log(2\pi)^d \left|\Sigma_i\right| - 2\log P_Y(i)}$$

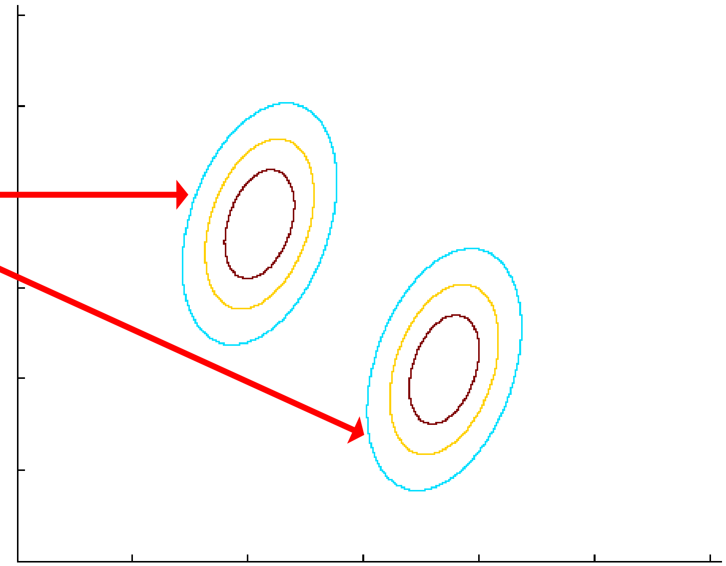


- the optimal rule is to assign x to the closest class
- closest is measured with the Mahalanobis distance $d_i(x,y)$
- to which the α constant is added to account for the class prior

The Gaussian classifier

- first special case of interest:
  - all classes have the same covariance,

$$\boxed{\Sigma_i = \Sigma, \quad \forall i}$$



- the BDR becomes

$$\boxed{i^*(x) = \arg\min_i \left[d(x, \mu_i) + \alpha_i \right]}$$

  - with

$$\boxed{d(x, y) = (x - y)^T \Sigma^{-1} (x - y)}$$

same metric for all classes

$$\boxed{\alpha_i = \log \cancel{(2\pi)^d} |\Sigma| - 2 \log P_Y(i)}$$

constant, not function of i, can be dropped

The Gaussian classifier

- in detail

$$i^*(x) = \arg\min_i \left[(x-\mu_i)^T \Sigma^{-1} (x-\mu_i) - 2\log P_Y(i) \right]$$

$$= \arg\min_i \left[x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_i - \mu_i^T \Sigma^{-1} x + \mu_i^T \Sigma^{-1} \mu_i - 2\log P_Y(i) \right]$$

$$= \arg\min_i \left[x^T \Sigma^{-1} x - 2\mu_i^T \Sigma^{-1} x + \mu_i^T \Sigma^{-1} \mu_i - 2\log P_Y(i) \right]$$

$$= \arg\max_i \left[\underbrace{\mu_i^T \Sigma^{-1}}_{w_i^T} x \underbrace{- \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P_Y(i)}_{w_{i0}} \right]$$

The Gaussian classifier

- in summary,

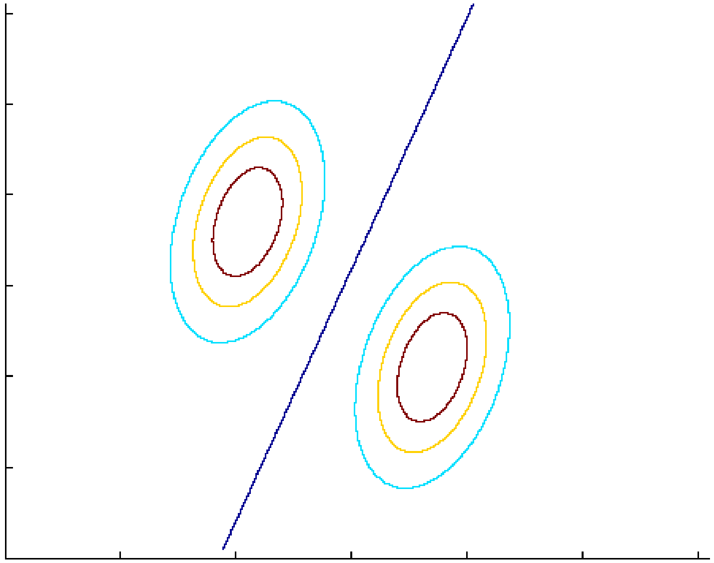
$$\boxed{i^*(x) = \arg\max_i g_i(x)}$$

- with

$$\boxed{\begin{aligned} g_i(x) &= w_i^T x + w_{i0} \\ w_i &= \Sigma^{-1}\mu_i \\ w_{i0} &= -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P_Y(i) \end{aligned}}$$

- the BDR is a linear function or a linear discriminant

discriminant:
$P_{Y|X}(1|\mathbf{x}) = 0.5$



Any questions?