9.913 Pattern Recognition for Vision

Class IV

Part I – Bayesian Decision Theory

Yuri Ivanov

TOC

Notation

x - scalar variable $\mathbf{x}$ - vector variable, sometimes x when clear

$p(x)$ - probability density function (continuous variable)

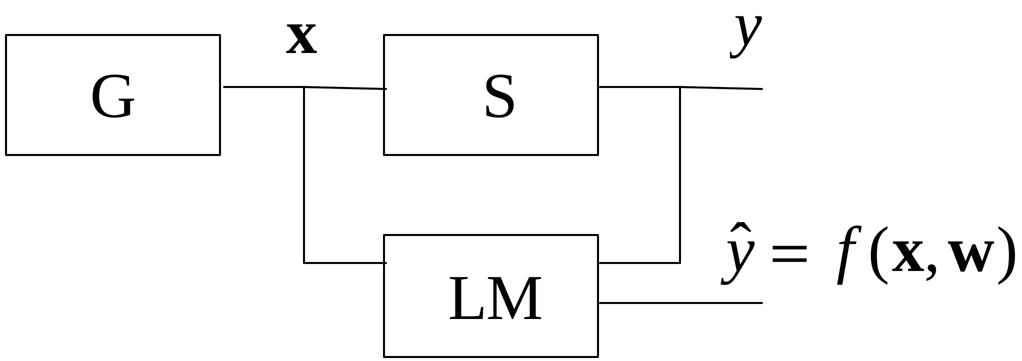
$P(x)$ - probability mass function (discrete variable)

$$p(\underbrace{a,b,c,d}_{\text{density of these}} \mid \underbrace{e,f,g,h}_{\text{function of these}})$$ - conditional density

$$\int f(\mathbf{x})d\mathbf{x} \equiv \int f(x_1,x_2,\ldots x_n)dx_1 dx_2 \ldots dx_n$$
$$\equiv \int\int\ldots\int f(x_1,x_2,\ldots x_n)dx_1 dx_2 \ldots dx_n$$

$$A \underset{w_2}{\overset{w_1}{>}} B$$ - if A > B then the answer is ω_1, otherwise ω_2

Learning machine:



G – Generator, or Nature: implements $p(\mathbf{x})$

S – Supervisor: implements $p(y|\mathbf{x})$

LM - Learning Machine: implements $f(\mathbf{x}, \mathbf{w})$

Loss and Risk

$L(y, f(\mathbf{x}, \mathbf{w}))$ - Loss Function – how much penalty we get for deviations from true *y*

$R(\mathbf{w}) = \int L(y, f(\mathbf{x}, \mathbf{w}))\, p(\mathbf{x}, y)\, d\mathbf{x} dy$ - Expected Risk – how much penalty we get on average

Goal of learning is to find $f(\boldsymbol{x},\boldsymbol{w}^*)$ such that $R(\boldsymbol{w}^*)$ is minimal.

Quick Illustration

What does it mean?

$$R(w) = \int\int L(y, f(x,w))\, \boxed{p(x,y)}\, dxdy$$

From basic probability: $p(x,y) = \boxed{p(y \mid x)}\, p(x)$

If no noise: $p(y \mid x) = d(y, g(x))$

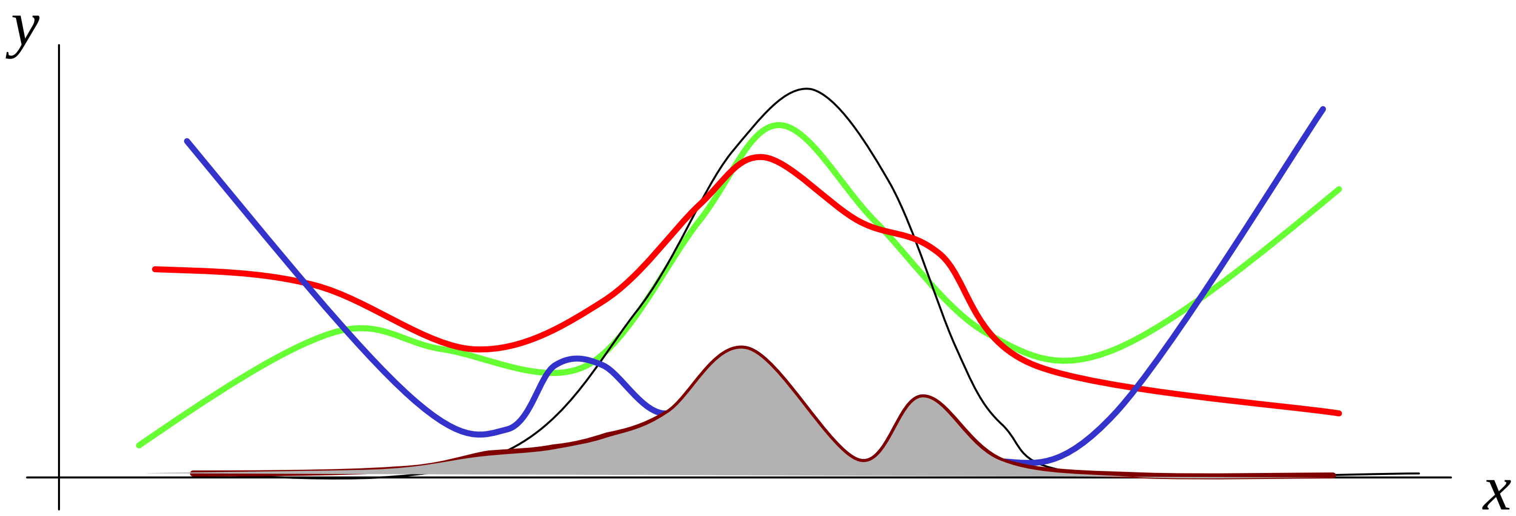
$$\Downarrow$$

$$R(w) = \int L(g(x), f(x,w))\, p(x)\, dx$$

Illustration cont.

So,

$$R(w) = \int L(g(x), f(x, w))\, p(x)\, dx$$

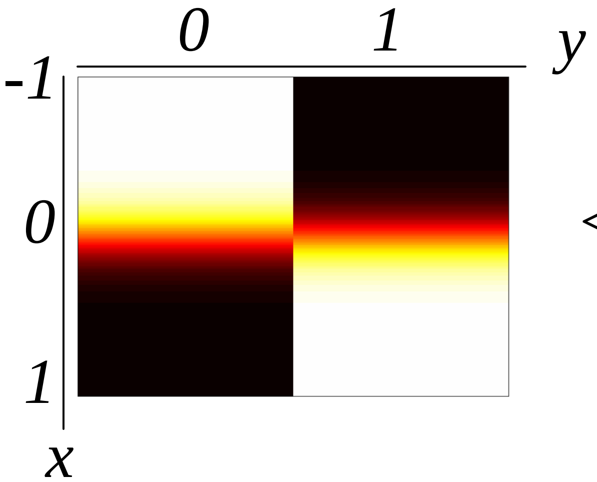


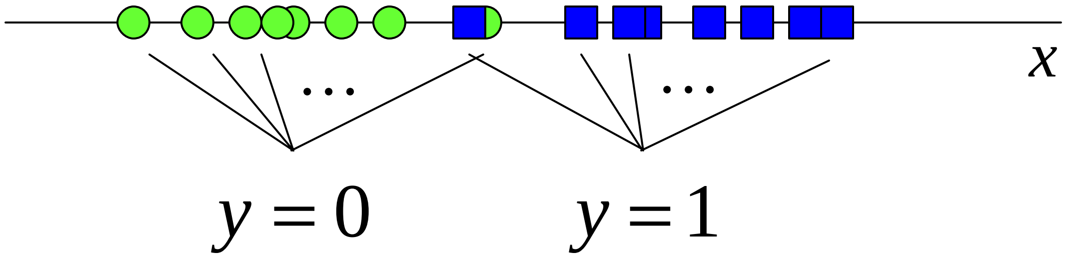
Example

Classification:

Measurements: $x \in [-1, \quad 1]$

Labels: $y = \{0,1\}$





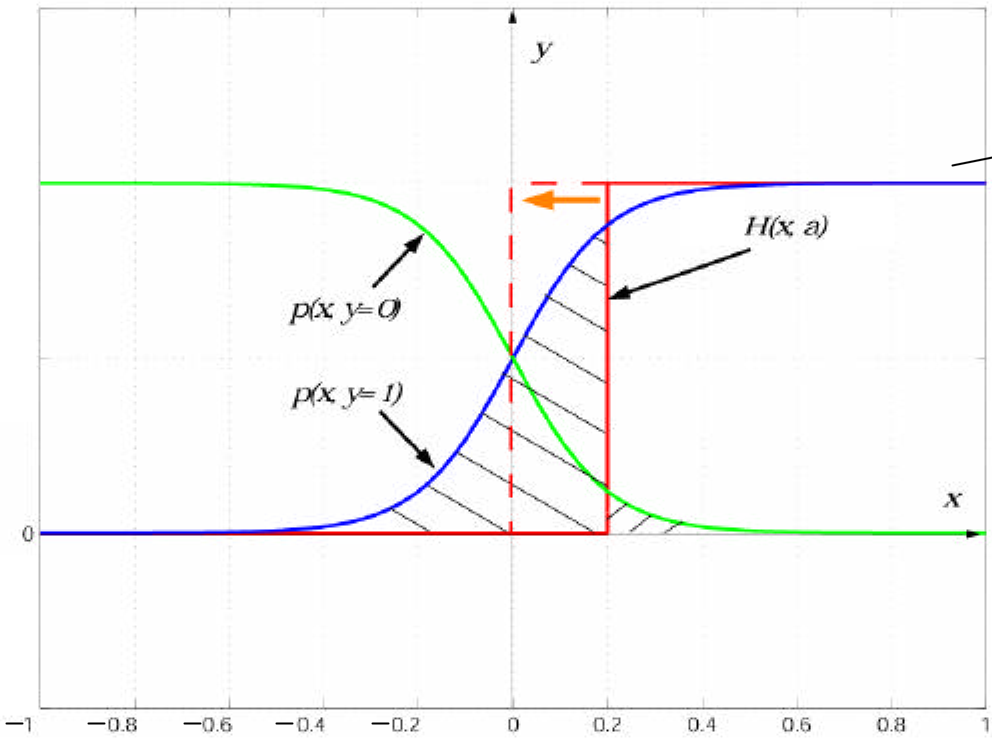
Find $f(\mathbf{x},\mathbf{w}^*)$ such that $R(\mathbf{w}^*)$ is minimal.

Example

Let's choose $f(x,a) = H(x,a)$ - step function

$$L(y, f(x,a)) = 1 - d(y - H(x,a))$$ - +*1* for every mistake

(superimposed)



$$R(a) = \sum_{Y=0}^{1} \int [1 - d(Y - H(x,a))] \, p(x, y = Y) dx$$

Learning in Reality

Fundamental problem: where do we get $p(x, y)$???

What we want: $$R(\mathbf{w}) = \int L(y, f(\mathbf{x}, \mathbf{w}))\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy$$

Approximate: estimate risk functional by averaging loss over ***observed*** (training) data.

What we get: $$R(\mathbf{w}) \leftarrow R_e(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w}))$$

Replace expected risk with empirical risk

What We Call Learning

Taxonomies of machine learning:

- by source of evaluation – Supervised, Transductive, Unsupervised, Reinforcement
- by inductive principle – ERM, SRM, MDL, Bayesian estimation
- by objective – classification, regression, vector quantization and density estimation

Taxonomy by Evaluation Source

- Supervised (*classification*, *regression*)

  Evaluation source - immediate error vector, that is, we get to see the true *y*
- Transductive

  Evaluation source – immediate error vector for SOME of the data
- Unsupervised (*clustering*, *density estimation*)

  Evaluation source - internal metric – we don't get to see true *y*
- Reinforcement

  Evaluation source - environment – we get to see some scalar value (possibly delayed) that in some way related to whether the label we chose was correct…

Taxonomy by Inductive Principle

- Empirical Risk Minimization (ERM)

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w})), \quad f(\mathbf{x}_i, \mathbf{w}) \in \Im$$

- Structural Risk Minimization (SRM)

"*complexity*"

$$\min_{\mathbf{w},h}\left(\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w})) + \Phi(h)\right),$$

$$f(\mathbf{x}_i, \mathbf{w}) \in \Im_k, \quad \Im_1 \subset \Im_2 \subset \ldots \subset \Im_k \subset \ldots, \quad k \propto h$$

- Minimum Description Length

$$\ell(D,H) = \ell(D|H) + \ell(H) \quad \Rightarrow \quad \min_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w})) + \Phi(\mathbf{w})\right)$$

- Bayesian Estimation

$$P(\mathbf{x}|\mathrm{X}) = \int P(\mathbf{x}|q)\, p(q|\mathrm{X})\, dq \quad \Rightarrow \quad \min\left(\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i, \mathbf{w})) + \Phi(P(\mathbf{w}))\right)$$

Taxonomy by Learning Objective

- Classification

$$L(y, f(\mathbf{x},\mathbf{w})) = 1 - \delta[(y, f(\mathbf{x},\mathbf{w}))$$

- Regression

$$L(y, f(\mathbf{x},\mathbf{w})) = (y - f(\mathbf{x},\mathbf{w}))^2$$

- Density Estimation

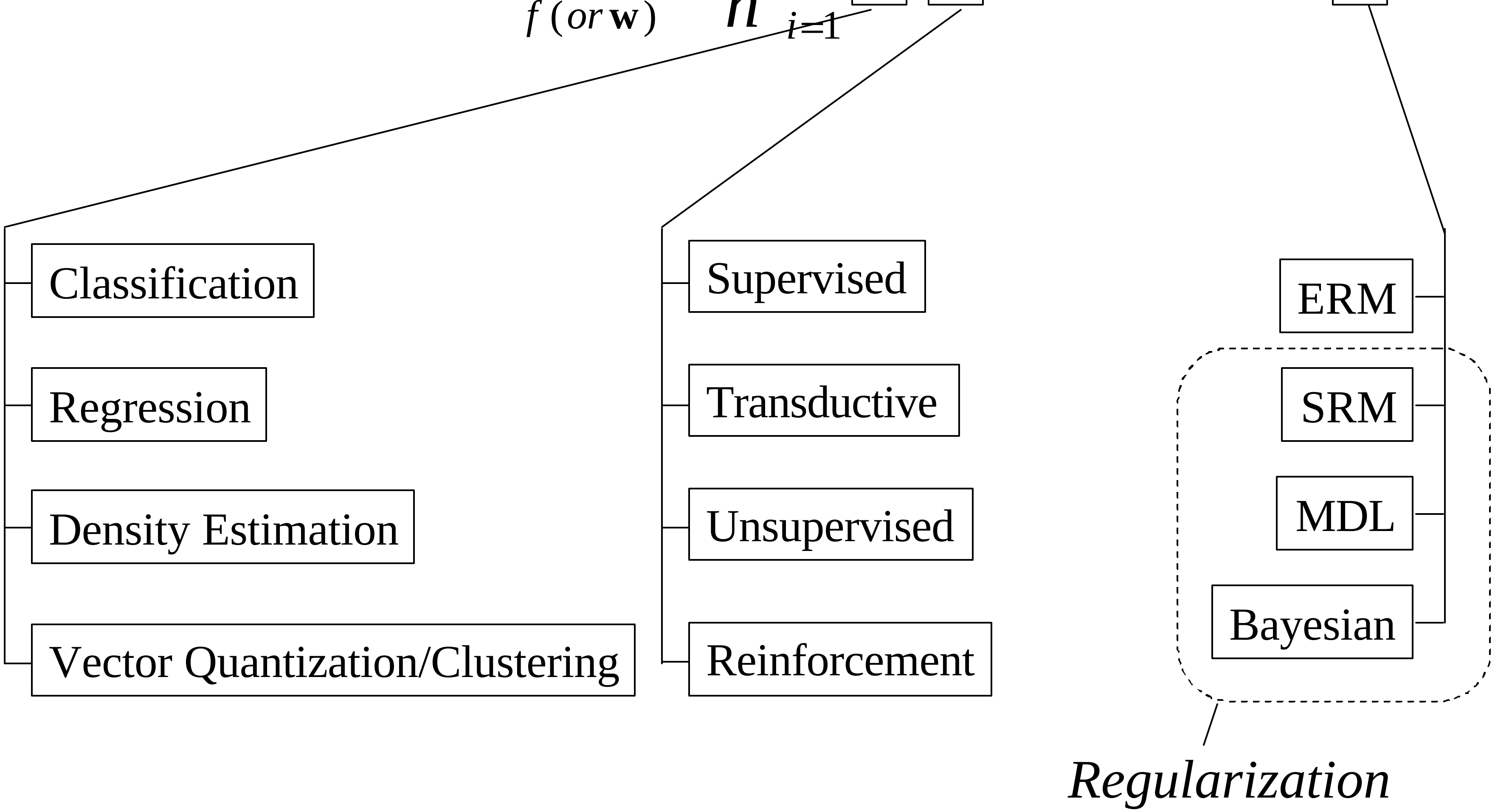
$$L(f(\mathbf{x},\mathbf{w})) = -\log(f(\mathbf{x},\mathbf{w}))$$

- Clustering/Vector Quantization

$$L(f(\mathbf{x},\mathbf{w})) = (\mathbf{x} - f(\mathbf{x},\mathbf{w})) \cdot (\mathbf{x} - f(\mathbf{x},\mathbf{w}))$$

The Lay of the Land

$$\text{find} \quad f\,(or\,\mathbf{w}) = \arg\min_{f\,(or\,\mathbf{w})}\left(\frac{1}{n}\sum_{i=1}^{n} L(y_i,\, f\,(\mathbf{x}_i\,,\mathbf{w})) + \Phi\right)$$



Making a decision about observation x is finding a rule that says:
If x is in region A, decide a, if x is in region B, decide b…

w - state of nature $w = \{w_i\}_{i=1}^{C}$

$P(w)$ - *Prior* probability Shorthand $P(w_i) \equiv P(w = w_i)$

$$\sum_{i=1}^{C} P(w_i) = 1$$

Poor man's decision rule:

Decide w_1 if $P(w_1) > P(w_2)$ otherwise w_2

Class-Conditional Density

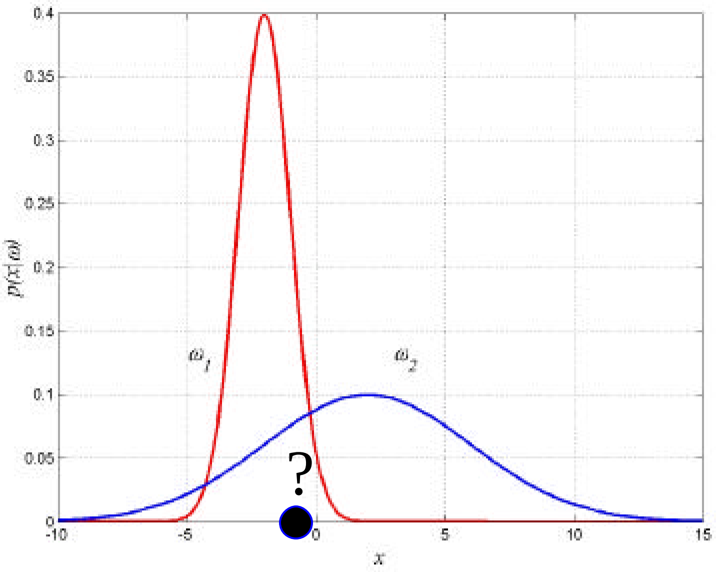
$p(x|w)$ - class-conditional density function

$$\int_{-\infty}^{\infty} p(x|w_i)dx = 1$$

$p(x|w)$ - density for *x* given that the nature is in the state w

How do we decide which class *x* came from?

Deciding on this is not fair

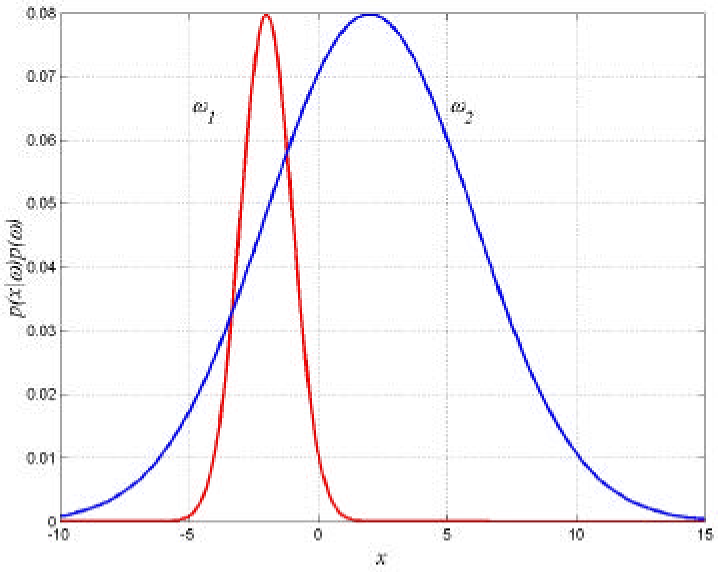


$$p(x,w) = p(x|w)P(w)$$ - joint density function

Good – It is fair

Bad – not very convenient.

$$P(w_1) = 0.2 \quad P(w_2) = 0.8$$



This relates to the measurement probabilistically, we need a function!

Bayes Rule and Posterior

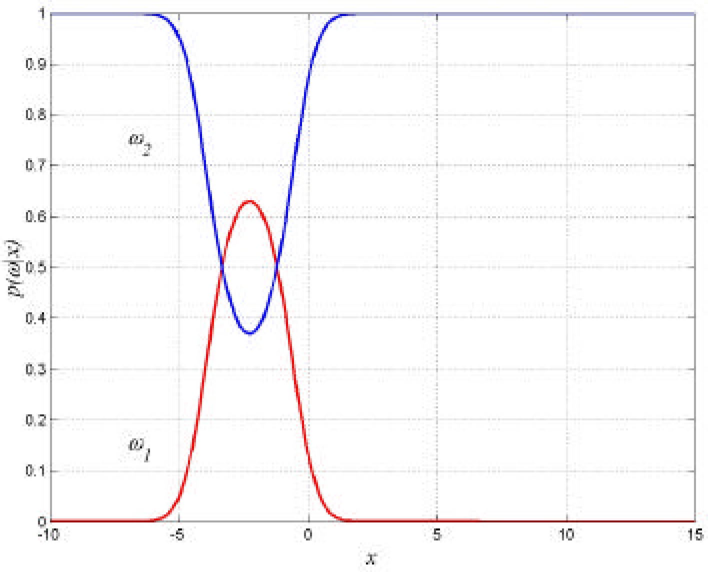
We want a probability of ω for each value of x: $P(w \mid x)$

Note that $$p(x,w) = p(x \mid w)P(w) = P(w \mid x)\,p(x) \quad \Rightarrow$$

$$\Rightarrow \boxed{P(w \mid x) = \frac{p(x \mid w)P(w)}{p(x)}}$$

likelihood — $p(x \mid w)$; *prior* — $P(w)$; *posterior* — $P(w \mid x)$; *evidence* — $p(x)$

Bayes rule



Continuum of binary distributions

Marginalization

Bayes Rule: how to convert prior to posterior by using measurements:

$$P(w \mid x) = \frac{p(x \mid w)P(w)}{p(x)}$$

What is $p(x)$?

$$p(x) = \sum_{i=1}^{C} p(x, w_i) = \sum_{i=1}^{C} p(x \mid w_i)P(w_i)$$

Or, more generally:

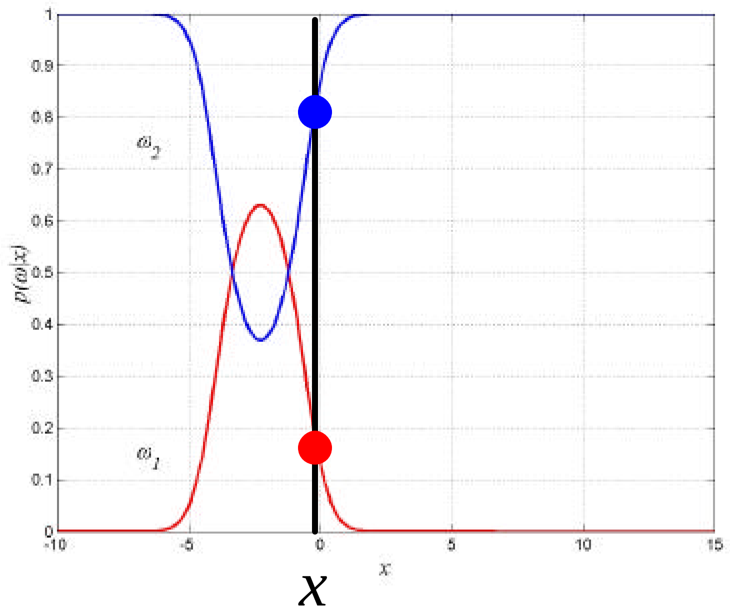
$$p(x) = \int p(x, y)\,dy \quad \text{- } \textit{marginalization}$$

Making Decisions

With posteriors we can compare class probabilities

Intuitively:

$$w = \arg\max\left(p(w_i \mid x) \right)$$



What is the probability of error?

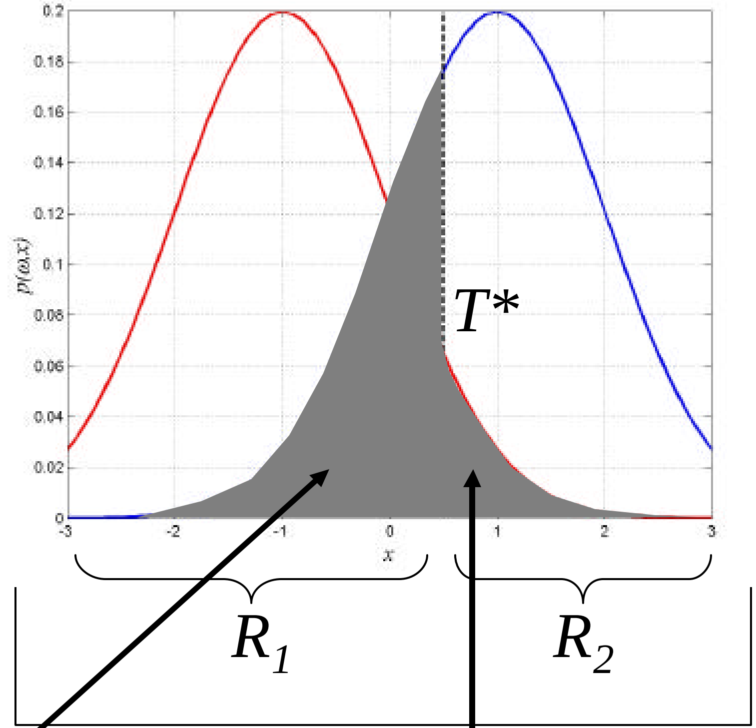
$$p(error \mid x) = \begin{cases} p(w_2 \mid x) & \text{if we choose } w_1 \\ p(w_1 \mid x) & \text{if we choose } w_2 \end{cases}$$

$$P(error) = \int P(error, x)dx = \int P(error \mid x)p(x)dx$$

Errors

How bad is a decision threshold?

$$P(e) = P(x \in R_1, w_2) + P(x \in R_2, w_1)$$



Recall that $P(a \in A) = \int\limits_{a \in A} p(a)\,da \Rightarrow$

$$P(e) = \int\limits_{x \in R_1} p(x, w_2)\,dx + \int\limits_{x \in R_2} p(x, w_1)\,dx =$$

$$= \int\limits_{x \in R_1} p(x \mid w_2) P(w_2)\,dx + \int\limits_{x \in R_2} p(x \mid w_1) P(w_1)\,dx$$

Bayes Decision Rule

To minimize $P(error) = \int P(error \mid x) p(x) dx$

we need to make $P(error|x)$ as small as possible for all x

$$\Downarrow$$

$$\min\left[P(error)\right] = \int \min_{w}\left[P(error \mid x)\right] p(x) dx$$

$$= \int \min\left[P(w_1 \mid x), P(w_2 \mid x)\right] p(x) dx$$

Bayes error

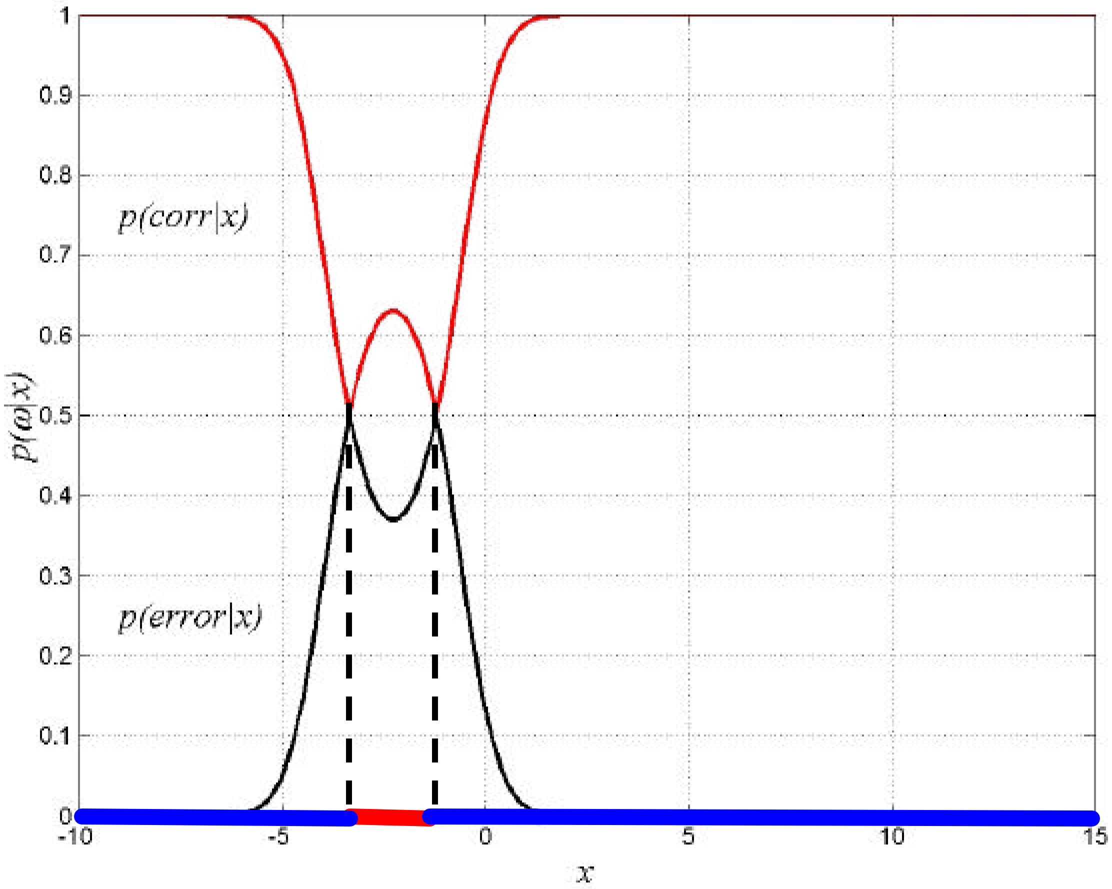
$$\Downarrow$$

Bayes decision rule:

Decide w_1 if $P(w_1 \mid x) > P(w_2 \mid x)$; otherwise w_2

Bayes Decision Rule

For Bayes decision rule (back to the 1st example):



R_1 – region where we always choose ω_1

R_2 – region where we always choose ω_2

Loss Function

$\{w_1, w_2, \ldots, w_c\}$ - set of *classes*

$\{a_1, a_2, \ldots, a_a\}$ - set of *actions*

$L(a_i \mid w_j)$ - *Loss function*, penalty sustained for taking an action α_i when the state of nature is ω_j

We make this up

For $x \in \mathbb{R}^d$ *conditional risk* for taking an action α_i in ω_j is the expected (average) loss for classifying ONE *x*:

$$R(a_i \mid x) = \sum_{j=1}^{c} L(a_i \mid w_j) P(w_j \mid x)$$

We will see a lot of x –es. To see how well we do, we average again:

$$R = \int R(a_i \mid x)\, p(x)\,dx = \int \left[\sum_{j=1}^{c} L(a_i \mid w_j) P(w_j \mid x) \right] p(x)\,dx$$

$$= \underbrace{\int \left[\sum_{j=1}^{c} L(a_i \mid w_j)\, p(w_j, x) \right] dx}$$

This is exactly the expression for expected risk from before

Similarly to the earlier argument about $P(error)$:

$$\min[R] = \int \min\left[R(a_i \mid x) \right] p(x)\,dx = R^* \quad \text{- Bayes risk}$$

Quick Summary

$L(a_i \,|\, w_j)$ - loss

$R(a_i \,|\, x) = E_{w|x}\left[L(a_i \,|\, w_j) \right]$ - conditional risk (expected loss)

$R = E_x\left[R(a_i \,|\, x) \right]$ - total risk (expected cond. risk)

$R^* = \min\left[R \right]$ - Bayes risk (minimum risk)

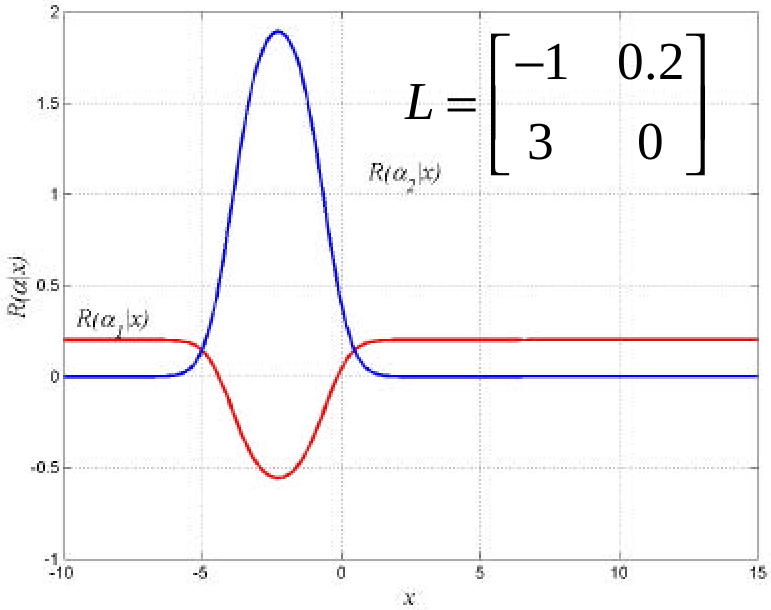
Minimum Risk Classification

Two classes and a simple action:

a_i - decide to choose ω_i

$$l_{ij} = L(a_i \mid w_j)$$

$$w = \{w_1, w_2\}$$



Then:

$$\begin{cases} R(a_1 \mid x) = l_{11} P(w_1 \mid x) + l_{12} P(w_2 \mid x) \\ R(a_2 \mid x) = l_{21} P(w_1 \mid x) + l_{22} P(w_2 \mid x) \end{cases}$$

Obvious decision – decide in favor of the class with minimal risk:

$$R(a_1 \mid x) \underset{w_2}{\overset{w_1}{<}} R(a_2 \mid x)$$

Likelihood Ratio Test

Rewriting $R(\alpha_i|x)$'s:

$$(l_{21}-l_{11})P(w_1\,|\,x)\underset{w_2}{\overset{w_1}{>}}(l_{12}-l_{22})P(w_2\,|\,x)\Rightarrow$$

$$(l_{21}-l_{11})\,p(x\,|\,w_1)P(w_1)\underset{w_2}{\overset{w_1}{>}}(l_{12}-l_{22})\,p(x\,|\,w_2)P(w_2)\Rightarrow$$

"Class model" *"Class prior"*

$$\frac{p(x\,|\,w_1)}{p(x\,|\,w_2)}\underset{w_2}{\overset{w_1}{>}}\frac{(l_{12}-l_{22})P(w_2)}{(l_{21}-l_{11})P(w_1)}$$

Likelihood Ratio Test

LRT Example

- You are driving to Blockbuster's to return a video due today
- It is 5 min to midnight
- You hit a red light
- You see a car that you 60% sure looks like a police car
- Traffic fine is $5 AND you are late
- Blockbuster's fine is $10

Should you run the red light?

Minimum Risk

$$P(police \mid x) = 0.6 \qquad P(\overline{police} \mid x) = 0.4$$

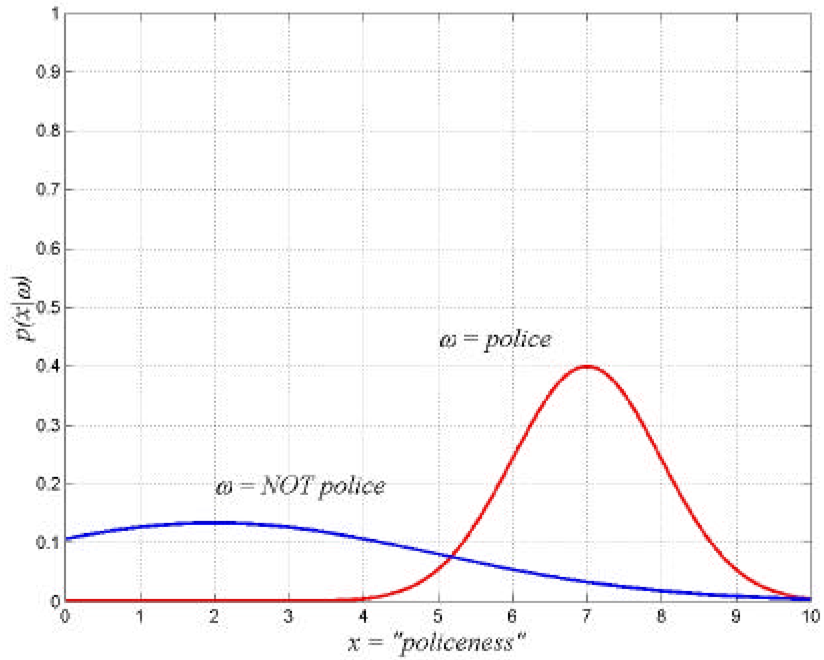
You pay

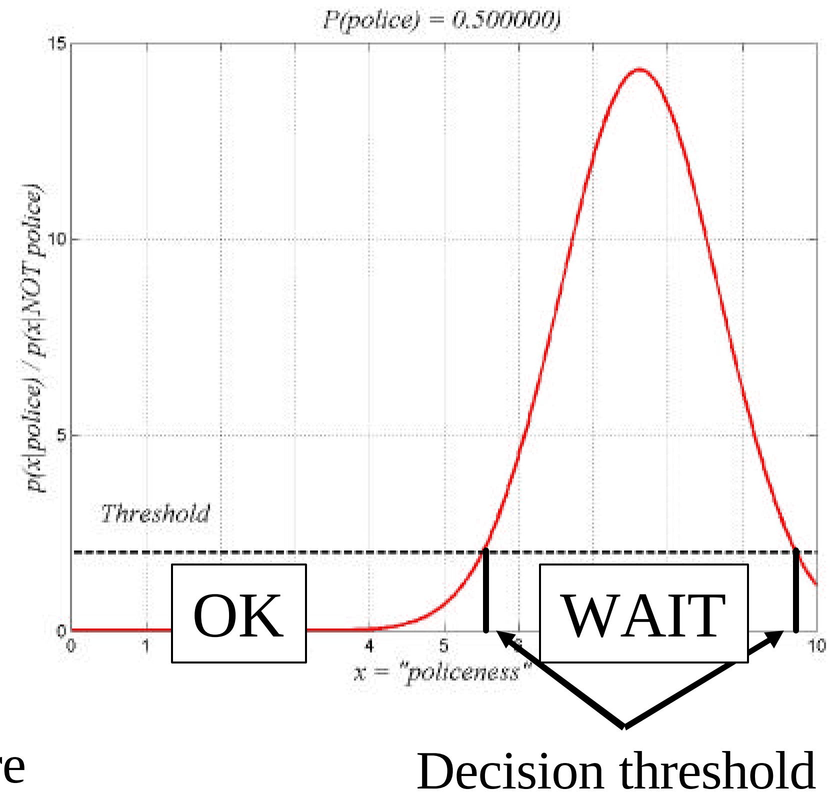
	police	not police
run	\$15	\$0
not run	\$10	\$10

$$L = \begin{pmatrix} 15 & 0 \\ 10 & 10 \end{pmatrix}$$

$$\begin{cases} R(run \mid x) = l_{11} P(police \mid x) + l_{12} P(\overline{police} \mid x) = \$9 \\ R(wait \mid x) = l_{21} P(police \mid x) + l_{22} P(\overline{police} \mid x) = \$10 \end{cases}$$

The risk is higher if you wait

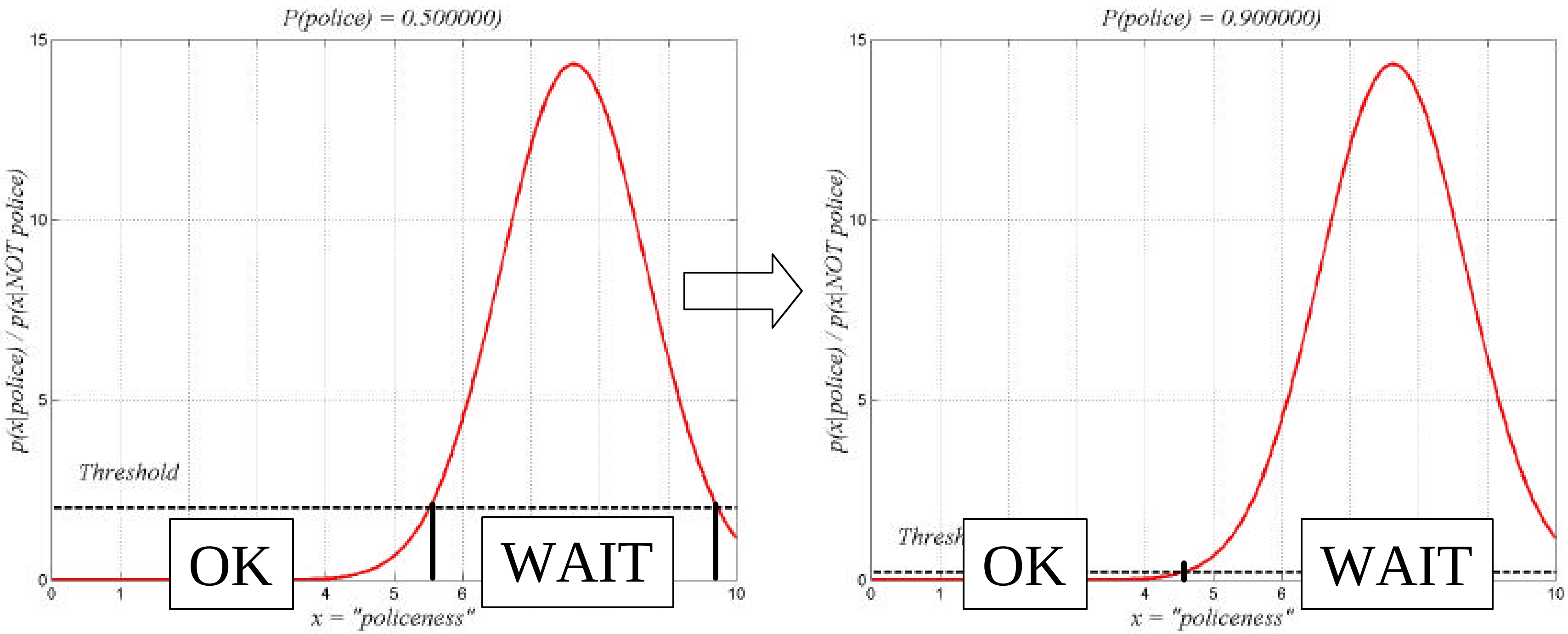




Let’s say we have this “policeness” feature

$$\frac{p(x|w_1)}{p(x|w_2)} \underset{w_2}{\overset{w_1}{>}} \frac{(l_{12}-l_{22})P(w_2)}{(l_{21}-l_{11})P(w_1)}$$

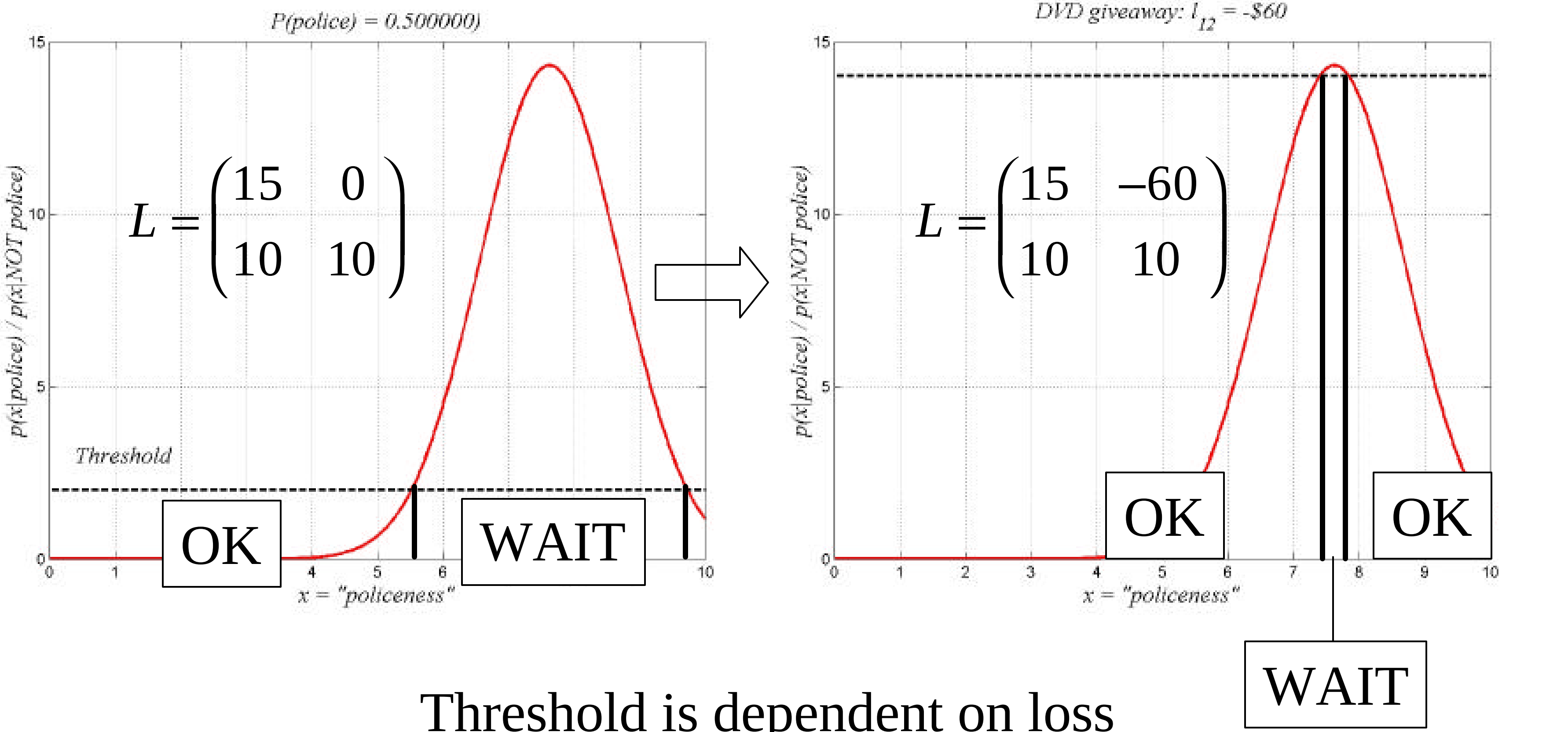
LRT Example



Threshold is dependent on priors

$$\frac{p(x \mid w_1)}{p(x \mid w_2)} \underset{w_2}{\overset{w_1}{>}} \frac{(l_{12} - l_{22})P(w_2)}{(l_{21} - l_{11})P(w_1)}$$

LRT Example



Threshold is dependent on loss

$$\frac{p(x|w_1)}{p(x|w_2)} \underset{w_2}{\overset{w_1}{>}} \frac{(l_{12}-l_{22})P(w_2)}{(l_{21}-l_{11})P(w_1)}$$

Minimum Error Rate Classification

Let‘s simplify the Min. Risk classification:

$$L = \begin{pmatrix} 0 & 1 & 1 \\ 1 & \ddots & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

- *zero-one* loss, just counts errors

Then the conditional risk becomes:

$$R(a_i \mid x) = \sum_{j=1}^{c} L(a_i \mid w_j) P(w_j \mid x)$$

$$= \sum_{\forall j \neq i} P(w_j \mid x) = 1 - P(w_i \mid x)$$

So, ω_i having the highest value of the posterior minimizes the risk:

$$P(w_i \mid x) \overset{w_i}{>} P(w_j \mid x) \quad \forall j \neq i$$

- good ol’ Bayes decision rule

Minimax Criterion

Is there a decision rule such that the risk is insensitive to priors?

$$R = \int_{R_1} \left[l_{11} P(w_1 \mid x) + l_{12} P(w_2 \mid x) \right] dx + \int_{R_2} \left[l_{21} P(w_1 \mid x) + l_{22} P(w_2 \mid x) \right] dx$$

After some algebra:

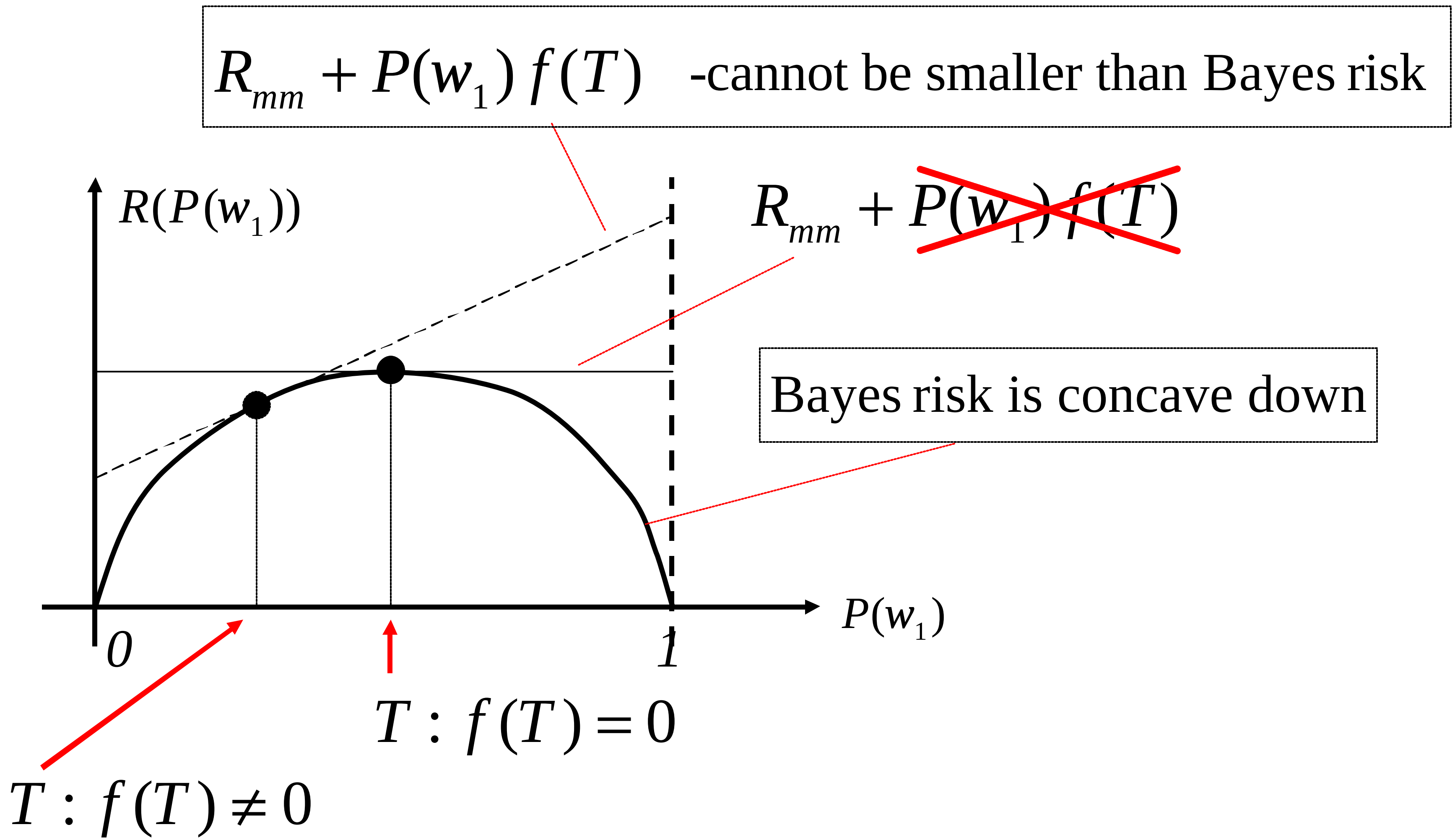
$$R(P(w_1)) = \boxed{l_{22} + (l_{12} - l_{22}) \int_{R_1} p(x \mid w_2) dx} + P(w_1) f(T)$$

Minimax risk

Goal – find the decision boundary T, such that $f(T)$ is 0.

At the boundary for which the minimum risk is maximal the risk is independent of priors.

Messy Illustration of the Minimax Solution



Worst possible Bayes risk is independent of priors

Discriminant Functions and Decision Surfaces

Discriminant functions conveniently represent classifiers:

$$C = \{g_1(x), g_2(x), \ldots g_n(x)\}$$

$$w_i : i = \arg\max\left(g_i(x)\right)$$

Eg:

$$g_i(x) = -R(a_i \mid x)$$

$$g_i(x) = P(w_i \mid x)$$

$$g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$$

Discriminants DO NOT have to relate to probabilities.

Discriminant for Binary Classification

For two-class problem:

$$g(x) = g_1(x) - g_2(x)$$

Then (assuming that classes are encoded as -1 and +1):

$$w_i = sign(g(x))$$

Eg:

$$g(x) = P(w_1 \mid x) - P(w_2 \mid x)$$

$$g(x) = \ln \frac{p(x \mid w_1)}{p(x \mid w_2)} + \ln \frac{P(w_1)}{P(w_2)}$$

Boundary for Two Normal Distributions

If we assume a Gaussian for a class model:

$$p(x \mid w_i) = \frac{1}{(2p)^{d/2} \mid \Sigma_i \mid^{1/2}} e^{-\frac{1}{2}\left((x-m_i)^T \Sigma_i^{-1} (x-m_i)\right)}$$

… and the minimum error rate classifier:

$$g_i(x) = \ln\left[p(x \mid w_i) P(w_i) \right] =$$

$$= -\frac{1}{2}\left((x-m_i)^T \Sigma_i^{-1} (x-m_i)\right) - \ln\left[(2p)^{d/2} \mid \Sigma_i \mid^{1/2} \right] + \ln P(w_i)$$

Boundary Between Two Normal Distributions

Discriminant (after some algebra):

$$g(x) = g_1(x) - g_2(x) == x^T W x + wx + w_0 \quad \textit{– quadratic}$$

where $W = -\frac{1}{2}\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)$ - a matrix

$w = \Sigma_1^{-1} m_1 - \Sigma_2^{-1} m_2$ - a vector

$w_o = \ldots$ well, the rest of it - a scalar

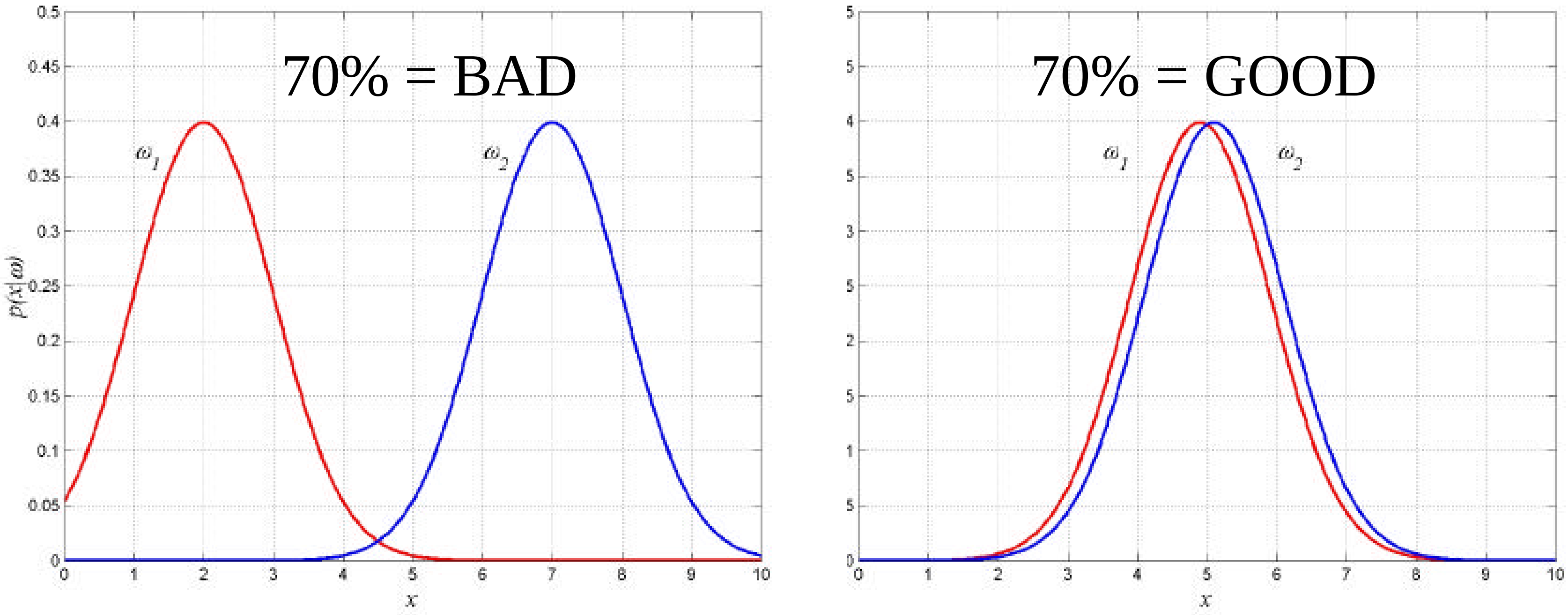
Special cases:

1) $\Sigma_1 = \Sigma_2 \Rightarrow W = 0 \Rightarrow g(x) - \textit{linear}$

2) $\Sigma_i = s_i I \Rightarrow W = \left[\frac{1}{2s_2} - \frac{1}{2s_1}\right] I = s\, I \Rightarrow g(x) - a\ \textit{circle}$

Evaluating Decisions

Is 70% classification rate good or bad?

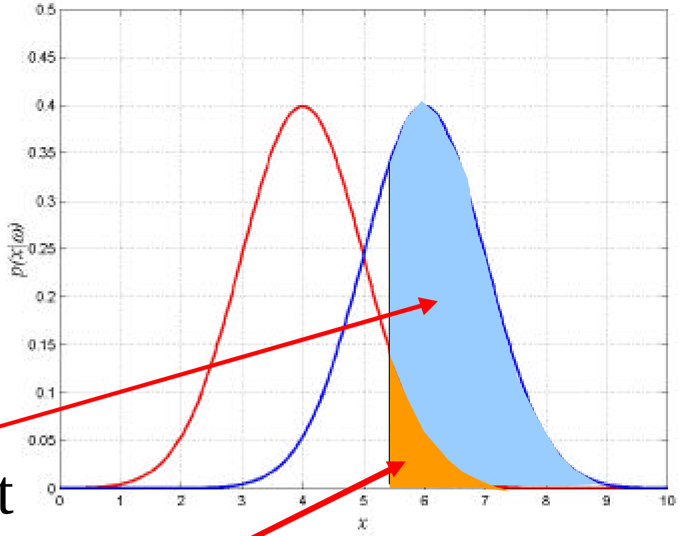


Discriminability:

$$d' = \frac{|m_1 - m_2|}{s}$$

High *d'* means that the classes are easy to discriminate.

We do not know m_1, m_2, s



but we can get:

$P(x > x^* \,|\, x \in w_2)$ - probability of hit

$P(x > x^* \,|\, x \in w_1)$ - probability of false alarm

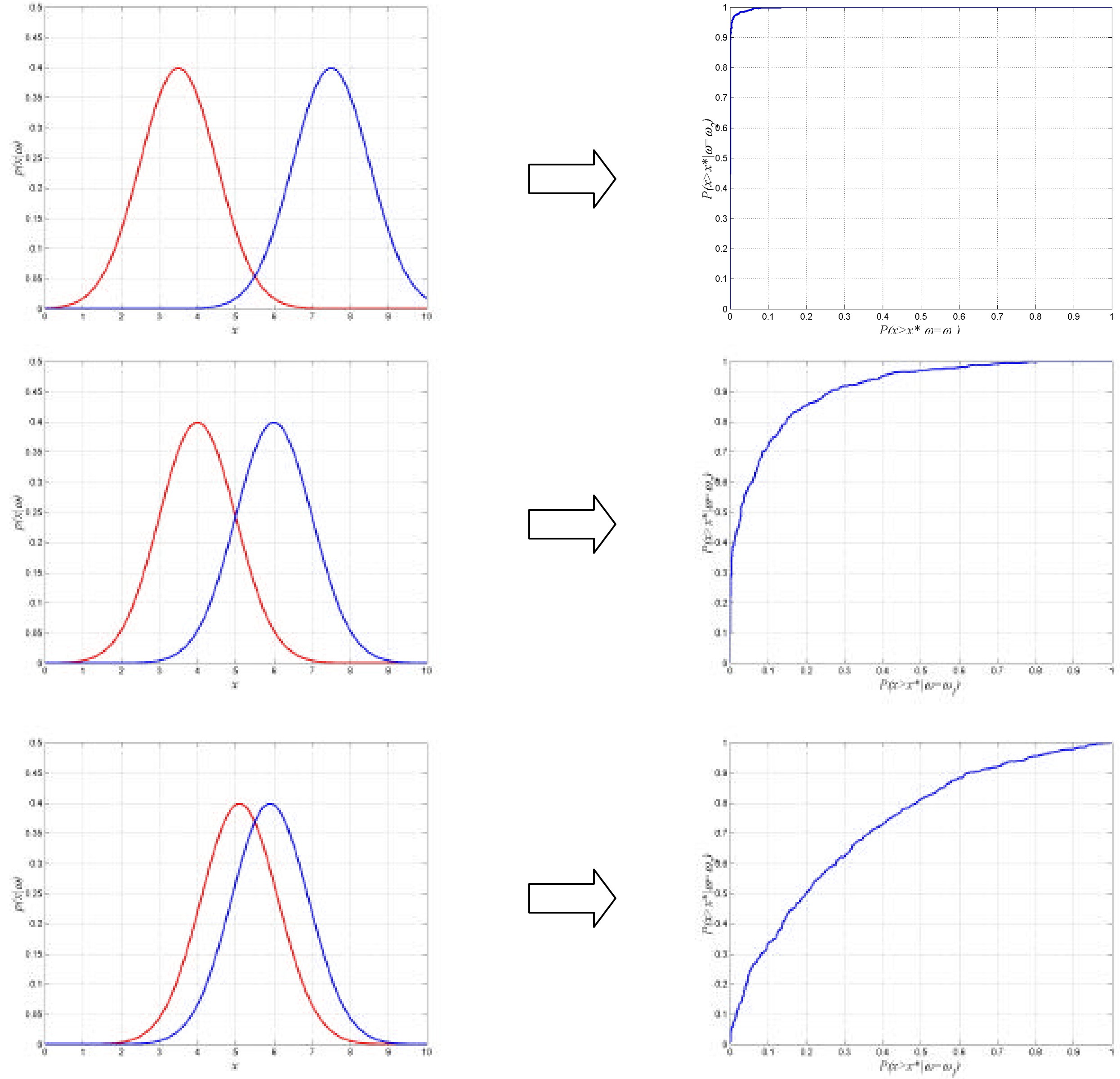
$P(x < x^* \,|\, x \in w_2)$ - probability of miss

$P(x < x^* \,|\, x \in w_1)$ - probability of correct rejection

Each x^* corresponds to a point on hit/false_alarm plane.

This is called an ROC curve

ROC Curve



Reality

In practice, it is done for a single parameter

Using the data for which true ω is known:

- Identify a parameter of interest
- Identify the parameter range
- Vary the parameter within the range
- Compute P(hit) and P(false_alarm) empirically for each value

It tells us how well the classifier can deal with the data set.

Homework – Part I

- A “chance” puzzle
  - Try to solve it
  - Understand the solution
  - Simulate in Matlab
- Build an ROC curve
  - Almost like in class
  - *Can you implement it efficiently?*