Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- No Slido poll questions today
- Homework 7 is due Friday
- Homework 8 will be posted early next week
- Project report template and instructions will be posted soon

Reading

- Bishop 2.5 (density estimation)

Today's lecture

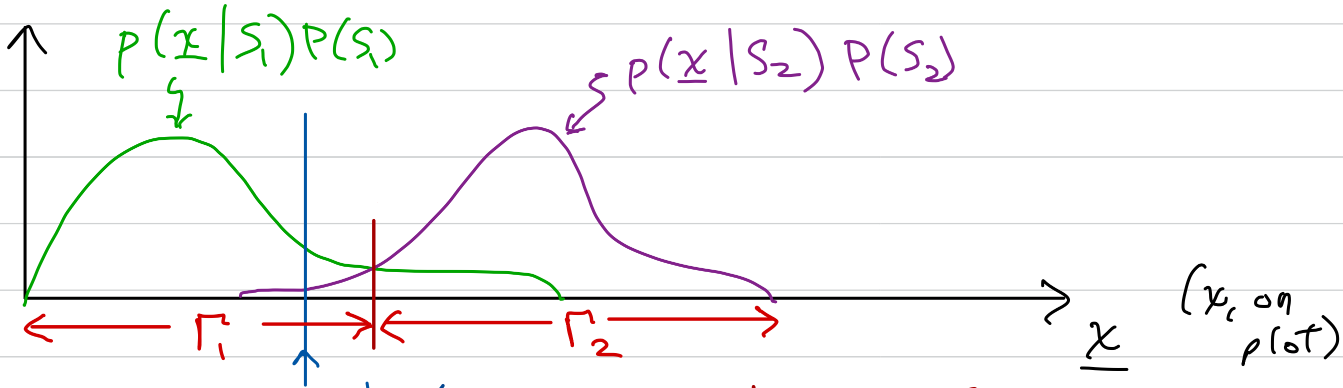
- Statistical classification and Bayes Decision Theory
  - Minimum-error classifiers and P_e (C=2)
  - Minimum-error classifiers and P_e (C>2)
  - Summary
  - Minimum-risk criterion
- Mahalanobis distance
- Classifiers: Gaussian density case
  - Linear Bayes (LDA)
  - Quadratic Bayes (QDA)
- Density estimation techniques for machine learning (time permitting)
  - Preliminaries

(1) $\left[P_e = P(S_1)\int_{\Gamma_2} p(\underline{x}|S_1)\,d\underline{x} + P(S_2)\int_{\Gamma_1} p(\underline{x}|S_2)\,d\underline{x} \right.$

→ P_e is our criterion fcn.

Goal: find Γ_1 and Γ_2 that minimize P_e.

Requirements: Γ_1 and Γ_2 must be non-overlapping.
$\Gamma_1 \cup \Gamma_2$ covers all of feature space.
(except possibly boundaries that have area (or volume) = 0)



⇒ assign to Γ_1 (for min. contribution to P_e).

In P_e, each pt. $\underline{x}$ must be included in 1 of the 2 terms; pick the smaller term.
⇒ minimizes P_e.

∴ Assign $\underline{x}$ to Γ_1 if $p(\underline{x}|S_2)P(S_2) < p(\underline{x}|S_1)P(S_1)$
Assign $\underline{x}$ to Γ_2 if $p(\underline{x}|S_2)P(S_2) > p(\underline{x}|S_1)P(S_1)$

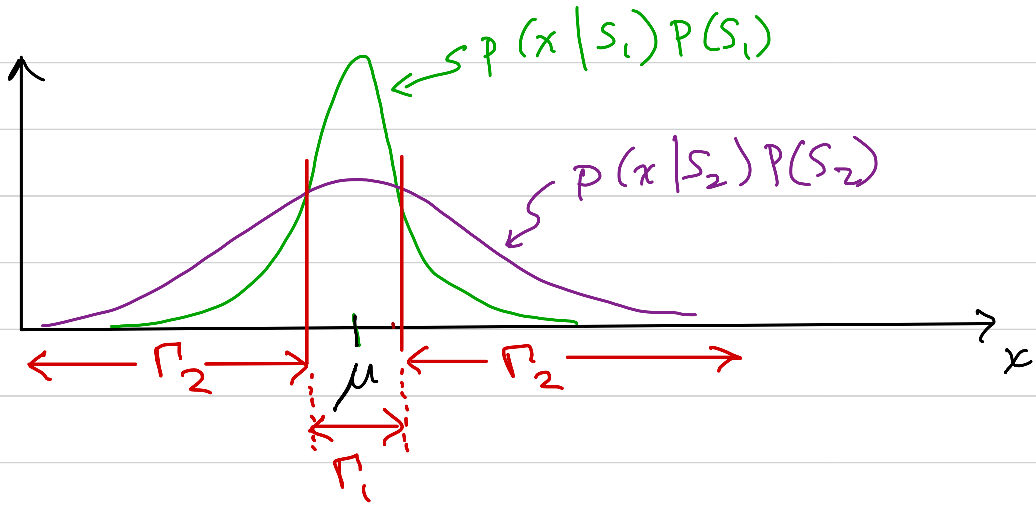
or

(2) $$\boxed{p(\underline{x}|S_1)P(S_1) \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} p(\underline{x}|S_2)P(S_2)}$$

↑ Bayes decision rule for min. error (c=2)

or $\ln[\cdot] \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} \ln[\cdot]$

Example: suppose $\mu_1 = \mu_2 = \mu$, $\sigma_1 < \sigma_2$: (Normal densities $p(\underline{x} \mid S_i)$)



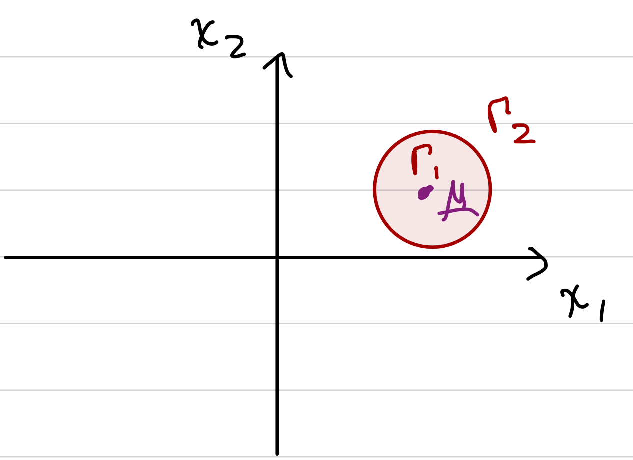
> Decision boundaries and regions? ↑

> Is this classifier linear? → No.

2D version: $p(\underline{x} \mid S_i) = N(\underline{x} \mid \underline{\mu}, \sigma_i^2 \underline{\underline{I}})$

$P(S_1) = P(S_2)$

$\sigma_1^2 < \sigma_2^2$



Use Bayes theorem:

$$P(S_k|\underline{x})\,p(\underline{x}) = p(\underline{x}|S_k)\,P(S_k)$$

$$p(\underline{x}) = \sum_{i=1}^{c} p(\underline{x}|S_i)\,P(S_i)$$

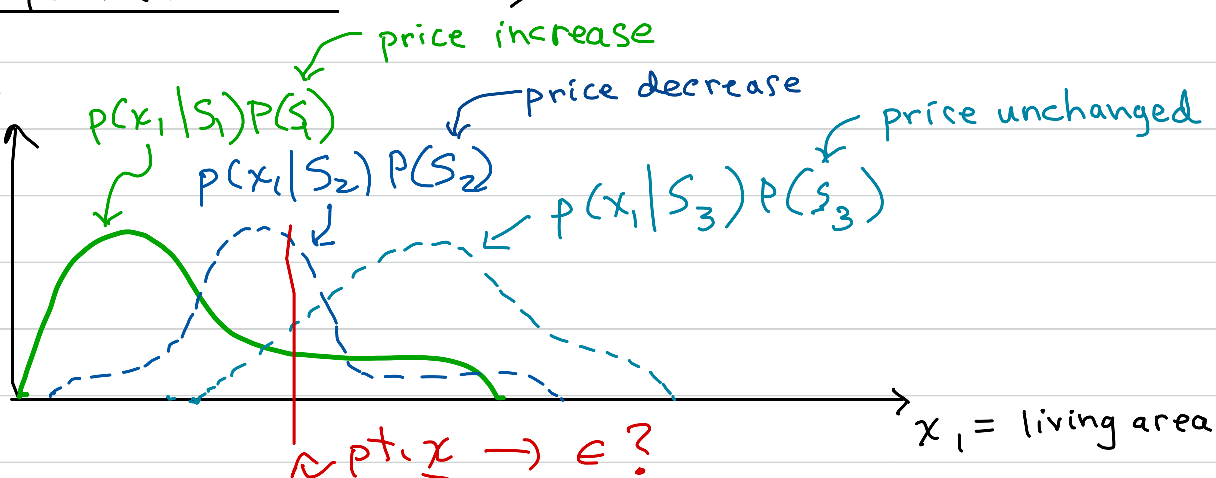
$$\Rightarrow \quad P(S_1|\underline{x})\,\cancel{p(\underline{x})} \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} P(S_2|\underline{x})\,\cancel{p(\underline{x})}$$

(3) $$\boxed{P(S_1|\underline{x}) \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} P(S_2|\underline{x})}$$

Bayes min. error dec. rule, in terms of posterior probabilities.

Bayes min error ($c > 2$)

Ex: $c=3$:



$$P_e = 1 - P_{correct}$$

$$P_{correct} = \sum_{i=1}^{c} P(S_i) \int_{\Gamma_i} p(\underline{x}|S_i)\,d\underline{x}$$

$$P_e = 1 - P_{correct} = 1 - \Big\{ P(S_1)\int_{\Gamma_1} p(\underline{x}|S_1)\,d\underline{x} + P(S_2)\int_{\Gamma_2} p(\underline{x}|S_2)\,d\underline{x} + P(S_3)\int_{\Gamma_3} p(\underline{x}|S_3)\,dx_1 \Big\}$$

$$\Rightarrow \quad \boxed{\begin{array}{l} \overbrace{p(\underline{x} \mid S_k)}^{g_k(\underline{x})} P(S_k) > \overbrace{p(\underline{x} \mid S_j)}^{g_j(\underline{x})} P(S_j) \quad \forall j \neq k \\ \Rightarrow \underline{x} \in \Gamma_k \end{array}}$$

Bayes min. error decision rule, $C > 2$.

Is this OvR, OvO, MVM, or something different?

This is a maximal value method.

Can define:

$$\boxed{\begin{aligned} & g_i(\underline{x}) = p(\underline{x} \mid S_i) P(S_i) \\ \text{or} & \\ \text{could choose } & \tilde{g}_i(\underline{x}) = \ln\left[p(\underline{x} \mid S_i) P(S_i)\right] \\ & \qquad = \ln p(\underline{x} \mid S_i) + \ln P(S_i) \end{aligned}}$$

Summary of Bayes Decision Theory so far

1. Bayes minimum error classifier

   Decision rule:

$$p(\underline{x}|S_i)P(S_i) > p(\underline{x}|S_j)P(S_j) \quad \forall j \neq i \Rightarrow \underline{x} \in \Gamma_i$$

   2-class case:

$$p(\underline{x}|S_1)P(S_1) \underset{\Gamma_2}{\overset{\Gamma_1}{\gtrless}} p(\underline{x}|S_2)P(S_2)$$

2. Probability of error

$$P_\epsilon = 1 - P_{correct} = 1 - \sum_{i=1}^{c} \int_{\Gamma_i} p(\underline{x}|S_i)P(S_i)\,d\underline{x}$$

   2-class case:

$$P_e = \int_{\Gamma_2} p(\underline{x}|S_1)P(S_1)\,d\underline{x} + \int_{\Gamma_1} p(\underline{x}|S_2)P(S_2)\,d\underline{x}$$

3. Note: for discrete-valued features x_k,

$$\int dx_k \text{ becomes } \sum_{x_k}$$

   and same decision rules apply.

MINIMUM RISK CRITERION [Bishop 1.5.2]

For cases in which minimizing P_e is not optimal; e.g. misclassifying S_1 data pt. as S_2 is significantly more costly than vice versa.

Ex: medical test that screens for cancer. S_1 = positive ($\Rightarrow$ cancer)
False negative is worse than false positive. S_2 = negative ($\Rightarrow$ no cancer)
↳ (miss the cancer)

→ Allow different costs for different kinds of error.

Let L_{ji} = loss of assigning $\underline{x}$ to Γ_i when it actually belongs to S_j.

Total expected loss is then:

$$E\{L\} = \underbrace{\sum_{j=1}^{C}\sum_{i=1}^{C}}_{\text{average over all classes}} \underbrace{\left[\int_{\Gamma_i} L_{ji}\, p(S_j|\underline{x})\,d\underline{x}\right]}_{\text{Expected loss of assigning } \underline{x} \text{ to } \Gamma_i.}$$

⇒ Instead of $p(S_i|\underline{x}) > p(S_k|\underline{x}) \;\forall k\neq i \Rightarrow \underline{x}\in\Gamma_i$

we have:

Decision rule

$$\boxed{\underbrace{\sum_{j=1}^{C} L_{ji}\, p(S_j|\underline{x})}_{R(\alpha_i|\underline{x})} < \sum_{j=1}^{C} L_{jk}\, p(S_j|\underline{x}) \quad \forall k\neq i \Rightarrow \underline{x}\in\Gamma_i}$$

$R(\alpha_i|\underline{x}) \triangleq$ conditional risk of taking action α_i ($\underline{x}\in\Gamma_i$) given $\underline{x}\in S_j$.

$$\underline{\underline{L}} = \begin{bmatrix} L_{11} & L_{12} & & \\ L_{21} & L_{22} & & \\ & & \ddots & \\ & & & L_{CC} \end{bmatrix}, \quad \text{typically } L_{ii}=0 \;\forall i.$$

Mahalanobis Distance [Bishop 2.3.0]

$$d_M^2(\underline{x}, \underline{m}) = (\underline{x} - \underline{m})^T \underline{\underline{\Sigma}}^{-1} (\underline{x} - \underline{m})$$

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \\ \sigma_{21} & \ddots & & \\ \vdots & & & \\ & & & \sigma_{DD} \end{bmatrix}$$

Case 1:

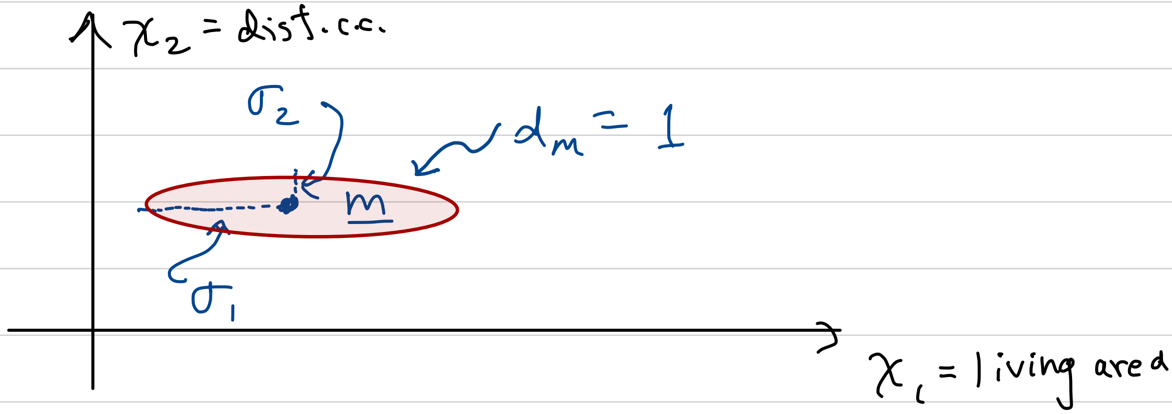
$$\underline{\underline{\Sigma}} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_D^2 \end{bmatrix}$$

$$\underline{\underline{\Sigma}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & & 0 \\ & \frac{1}{\sigma_2^2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sigma_D^2} \end{bmatrix}$$

$$d_M^2(\underline{x}, \underline{m}) = \sum_{i=1}^{D} \frac{1}{\sigma_i^2} (x_i - m_i)^2$$

$$d_M^2(\underline{x}, \underline{m}) = \text{const.} \Rightarrow ?$$

2-space: $$d_M^{\,2} = \frac{(x_1 - m_1)^2}{\sigma_1^{\,2}} + \frac{(x_2 - m_2)^2}{\sigma_2^{\,2}} = \text{const.}$$



→ Ellipse (2D)

→ Hyperellipsoid (D dim.)

CASE 2:

$$\underline{\underline{\Sigma}} = \text{general}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \\ \sigma_{21} & \ddots & & \\ \vdots & & \ddots & \\ \vdots & & & \sigma_{DD} \end{bmatrix}$$

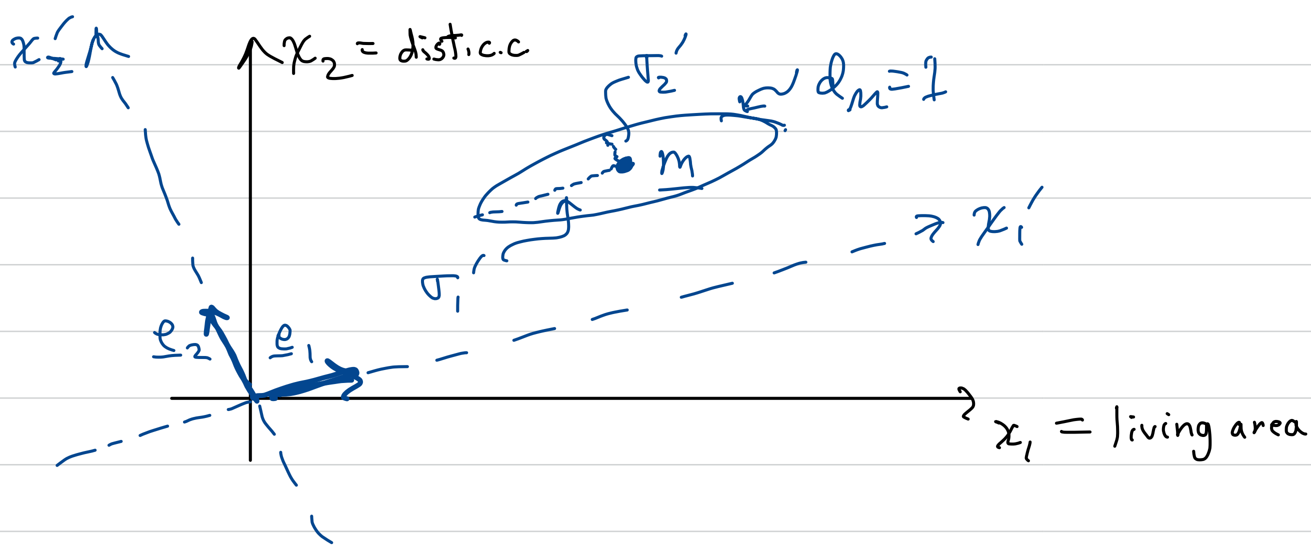
$$d_M^{\,2}(\underline{x}, \underline{m}) = (\underline{x} - \underline{m})^T \underline{\underline{\Sigma}}^{-1} (\underline{x} - \underline{m})$$

Apply orthonormal transformation (rotate basis):

$$\underline{x}' = \underline{\underline{E}}^T \underline{x}$$

$$\underline{\underline{\Sigma}}' = \underline{E}^T \underline{\underline{\Sigma}} \underline{E} = \underline{\underline{\Lambda}} = \text{diagonal} \rightarrow \text{CASE 1.}$$

$\Rightarrow d_M^{\,2} = \text{const.} \Rightarrow$ hyperellipsoids (axes rotated)



BAYES MIN. ERROR CLASSIFIERS — GAUSSIAN DENSITY CASE

[Bishop 2.3.3 — optional reading]

$$p(\underline{x} \mid S_k) = N(\underline{x} \mid \underline{m}_k, \underline{\underline{\Sigma}}_k)$$

$$= \frac{1}{(2\pi)^{D/2} \left| \underline{\underline{\Sigma}}_k \right|^{1/2}} \exp\left\{ -\frac{1}{2}\left[(\underline{x}-\underline{m}_k)^T \underline{\underline{\Sigma}}_k^{-1} (\underline{x}-\underline{m}_k) \right] \right\}$$

$$\underset{i}{\text{Maximize}} \; p(\underline{x} \mid S_i) P(S_i)$$

$$g_i(\underline{x}) = \ln\{ p(\underline{x} \mid S_i) P(S_i) \} = \ln p(\underline{x} \mid S_i) + \ln P(S_i)$$

$$(1) \quad \boxed{g_i(\underline{x}) = -\frac{1}{2}\ln \left| \underline{\underline{\Sigma}}_i \right| - \frac{1}{2}(\underline{x}-\underline{m}_i)^T \underline{\underline{\Sigma}}_i^{-1} (\underline{x}-\underline{m}_i) + \ln P(S_i)}$$

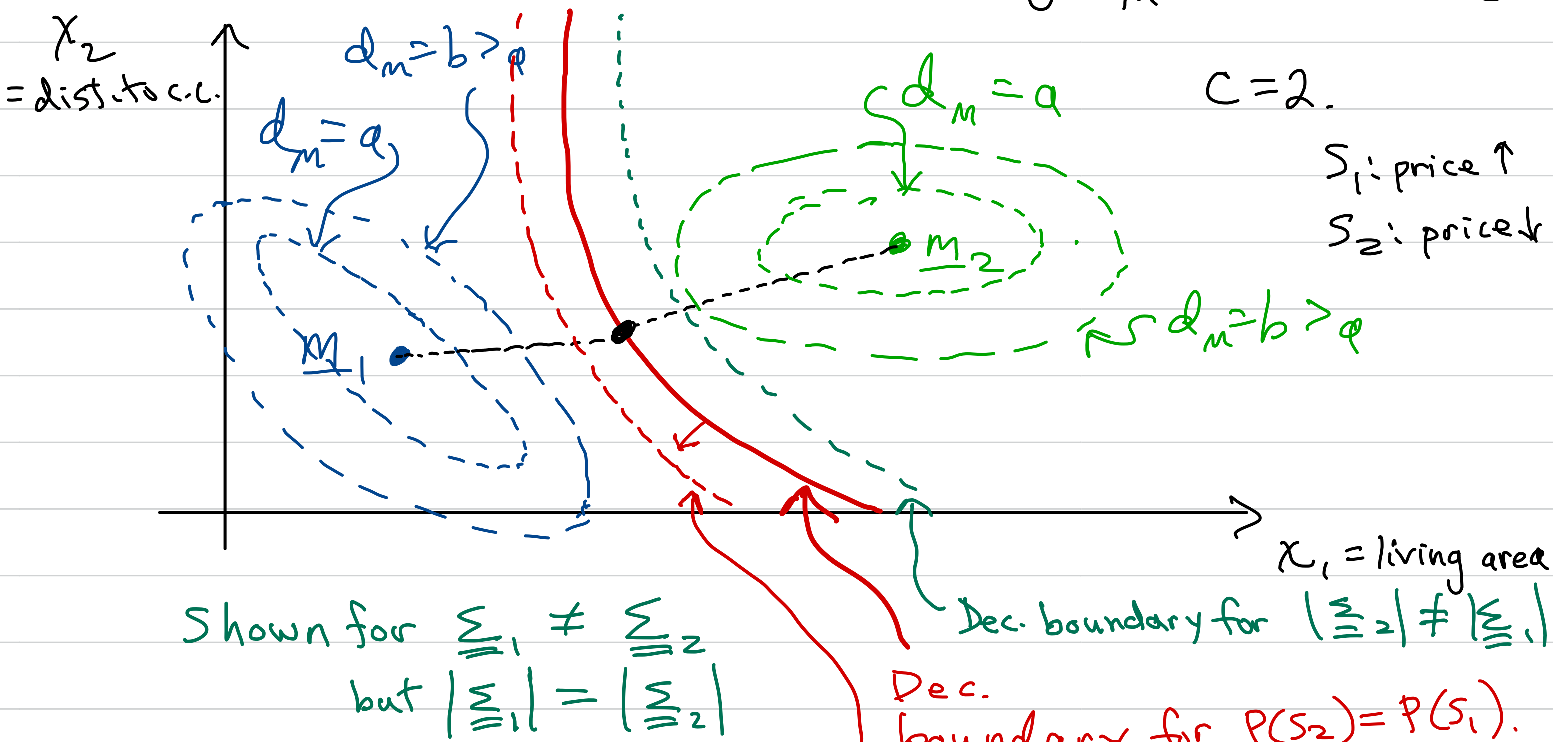
Let: $\left|\underline{\underline{\Sigma}}_1\right| = \left|\underline{\underline{\Sigma}}_2\right| = \cdots$

and: $P(S_1) = P(S_2) = \cdots$

Then:

$$g_i(\underline{x}) = -(\underline{x}-\underline{m}_i)^T \underline{\underline{\Sigma}}_i^{-1} (\underline{x}-\underline{m}_i) = -d_M^2(\underline{x}, \underline{m}_i)$$

$\Rightarrow$ Nearest-means classifier using d_M instead of d_E.



Comments

1. Include: $P(S_i) \neq P(S_j)$

   $\Rightarrow$ Boundary shifts
   e.g.: $P(S_2) > P(S_1)$ (see plot)

2. Include: $|\underline{\underline{\Sigma}}_i| \neq |\underline{\underline{\Sigma}}_j|$

   $\Rightarrow$ incorporates differences in ellipsoid volumes from class to class.

Case A $\quad \underline{\underline{\Sigma}}_i = \underline{\underline{\Sigma}}$

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{m}_i)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{m}_i) + \ln P(S_i)$$

$$= -\frac{1}{2}\left[\underline{x}^T \underline{\underline{\Sigma}}^{-1}\underline{x} - 2\underline{m}_i^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \underline{m}_i^T \underline{\underline{\Sigma}}^{-1}\underline{m}_i\right] + \ln P(S_i)$$

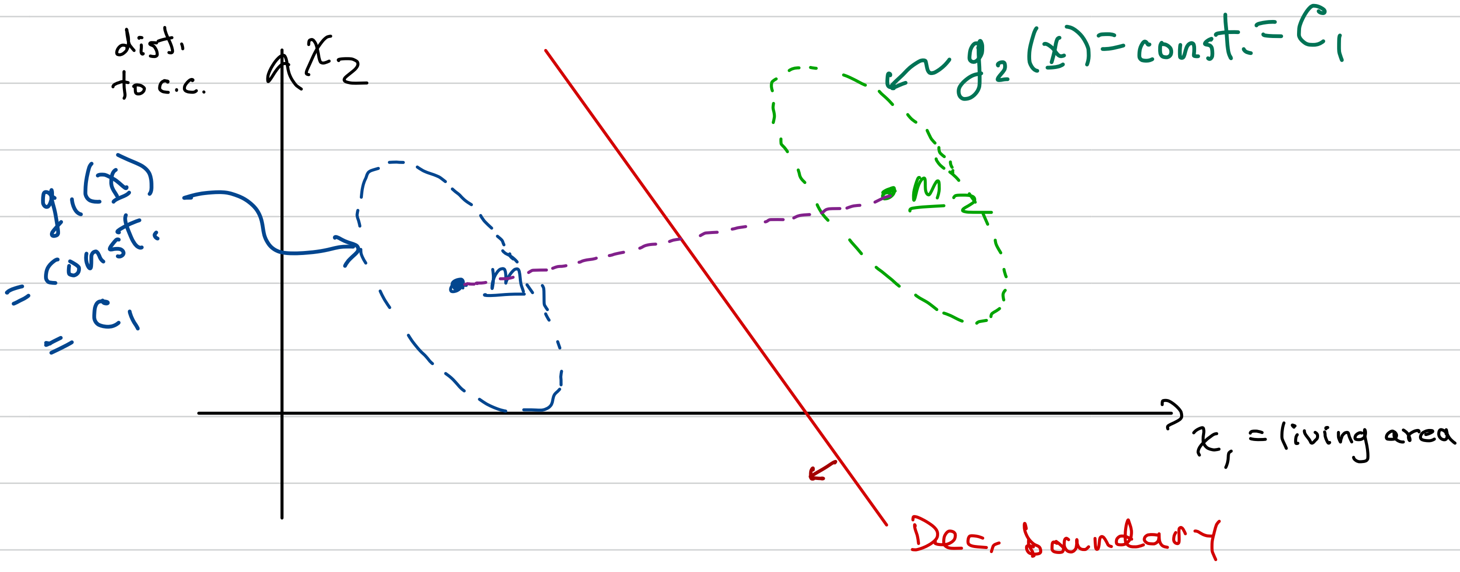
$$\boxed{\text{Let } g_i(\underline{x}) = \underline{m}_i^T \underline{\underline{\Sigma}}^{-1}\underline{x} - \frac{1}{2}\underline{m}_i^T \underline{\underline{\Sigma}}^{-1}\underline{m}_i + \ln P(S_i)}$$

$$= \underline{w}^T \underline{x} + w_0$$

is this classifier linear? → Yes!

⇒ Classifier is linear.

Often called Linear Bayes.



Also called: LDA: linear discriminant analysis.
(if parameters are estimated from the data)

Case B: $\underline{\underline{\Sigma}}_i$ = arbitrary

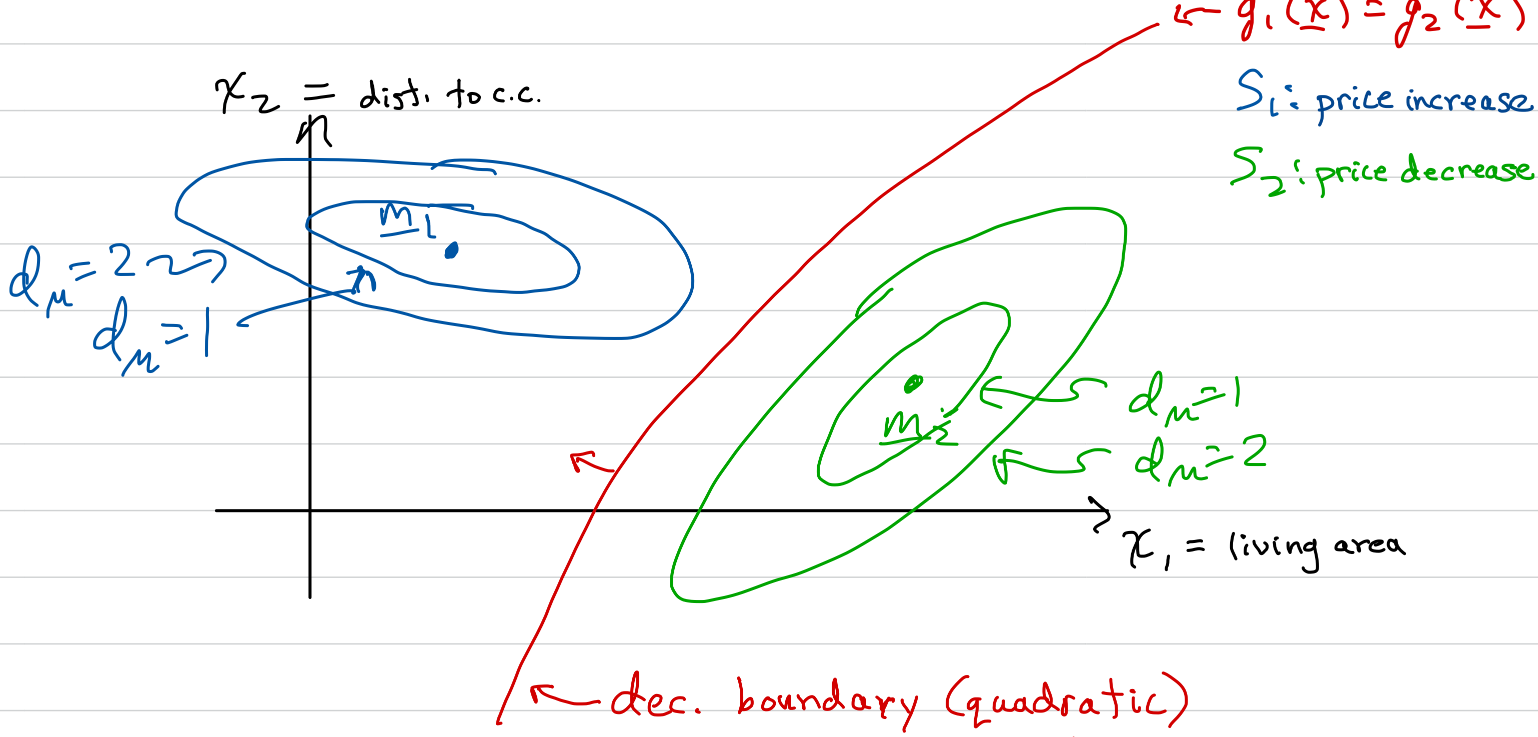
$$g_i(\underline{x}) = \frac{-1}{2}\ln|\underline{\underline{\Sigma}}_i| - \frac{1}{2}\underbrace{d_M^2(\underline{x}, \underline{m}_i)}_{(\underline{x}-\underline{m}_i)^T \underline{\underline{\Sigma}}_i^{-1}(\underline{x}-\underline{m}_i)} + \ln P(S_i)$$

$g_i(\underline{x})$ is quadratic fcn. of $\underline{x}$.

$\Rightarrow$ Dec. boundaries are quadratic

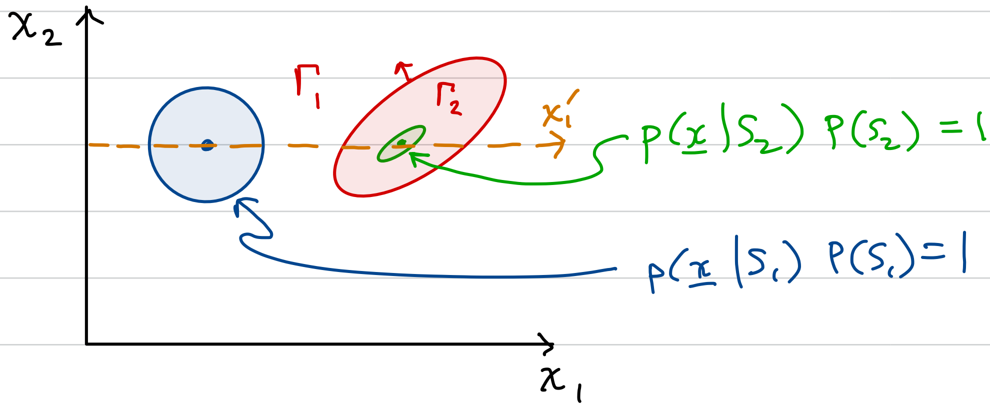
$\rightarrow$ Quadratic Bayes

Also called QDA: quadratic discriminant analysis
(if parameters are estimated from the data)

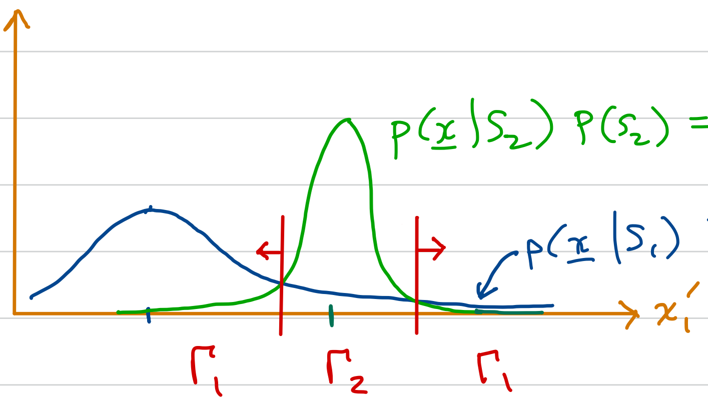


dec. boundary (quadratic)
ex: hyperparabola
hyperhyperbola
hyperellipsoid
hypersphere

Ex 1:

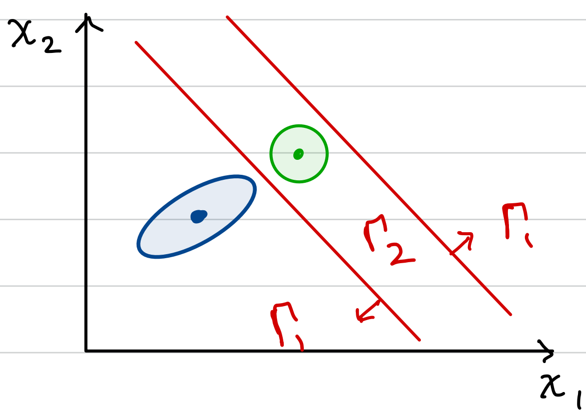


x_2
Γ_1
Γ_2
x_1'
$p(\underline{x}|S_2)\,P(S_2) = 1$
$p(\underline{x}|S_1)\,P(S_1) = 1$
x_1



$p(\underline{x}|S_2)\,P(S_2) = 1$
$p(\underline{x}|S_1)\,P(S_1) = 1$
x_1'
Γ_1
Γ_2
Γ_1

Ex 2:



x_2
Γ_2
Γ_1
Γ_1
x_1