Machine Learning I: Supervised Methods

B. Keith Jenkins

Announcements

- Homework 1 is due Friday
- Slido poll questions in lecture will begin soon - initially just a trial run

Today's lecture

- Discriminant functions and approaches to multiclass problems ($C > 2$) (part 2)
    - One vs. Rest (OvR) (part 2)
    - One vs. One (OvO)
    - Maximal Value Method (MVM)
    - Summary
- Computational complexity (part 1)

deferred to lecture 5.

Copyright 2021-2024 B. Keith Jenkins. All lecture notes contained herein are for use only by students and instructors of EE 559, Spring 2024.

From Lecture 3:

2. Multiclass ($C > 2$) problems

→ Can we pose a C-class problem ($C > 2$) as a set of 2-class problems?
Yes.

(i) Use C discriminant fcns: $g_k(\underline{x})$, $k = 1, 2, \cdots, C$.

One vs rest (OvR) (also called One vs-all)
to define C 2-class problems.
Each 2-class problem: S_k' vs. $\overline{S_k'}$ (e.g., hs. price decr. vs. hs. price not decr.)
(cat vs. not cat)

for $k = 1, 2, \cdots, C$.

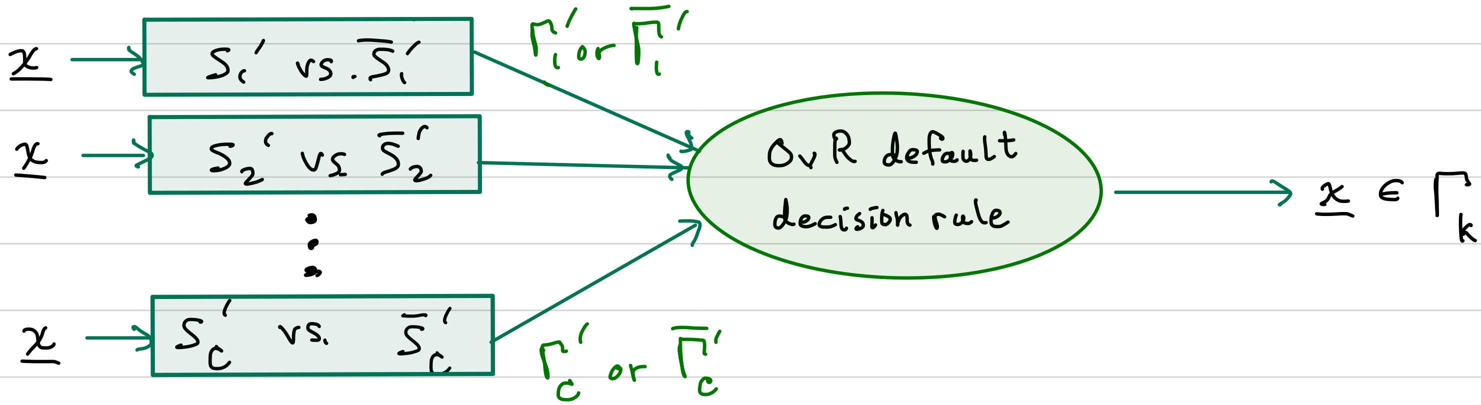
[see plot below]

Combine results:

$\boxed{\text{OvR } \underline{\text{Decision rule}}: \ \underline{x} \in \Gamma_k \text{ IFF } \underline{x} \in \Gamma_k' \text{ AND } \underline{x} \in \overline{\Gamma_j'} \ \forall j \neq k.}$

our default OvR decision rule. Other OvR decision rules are possible.

OvR



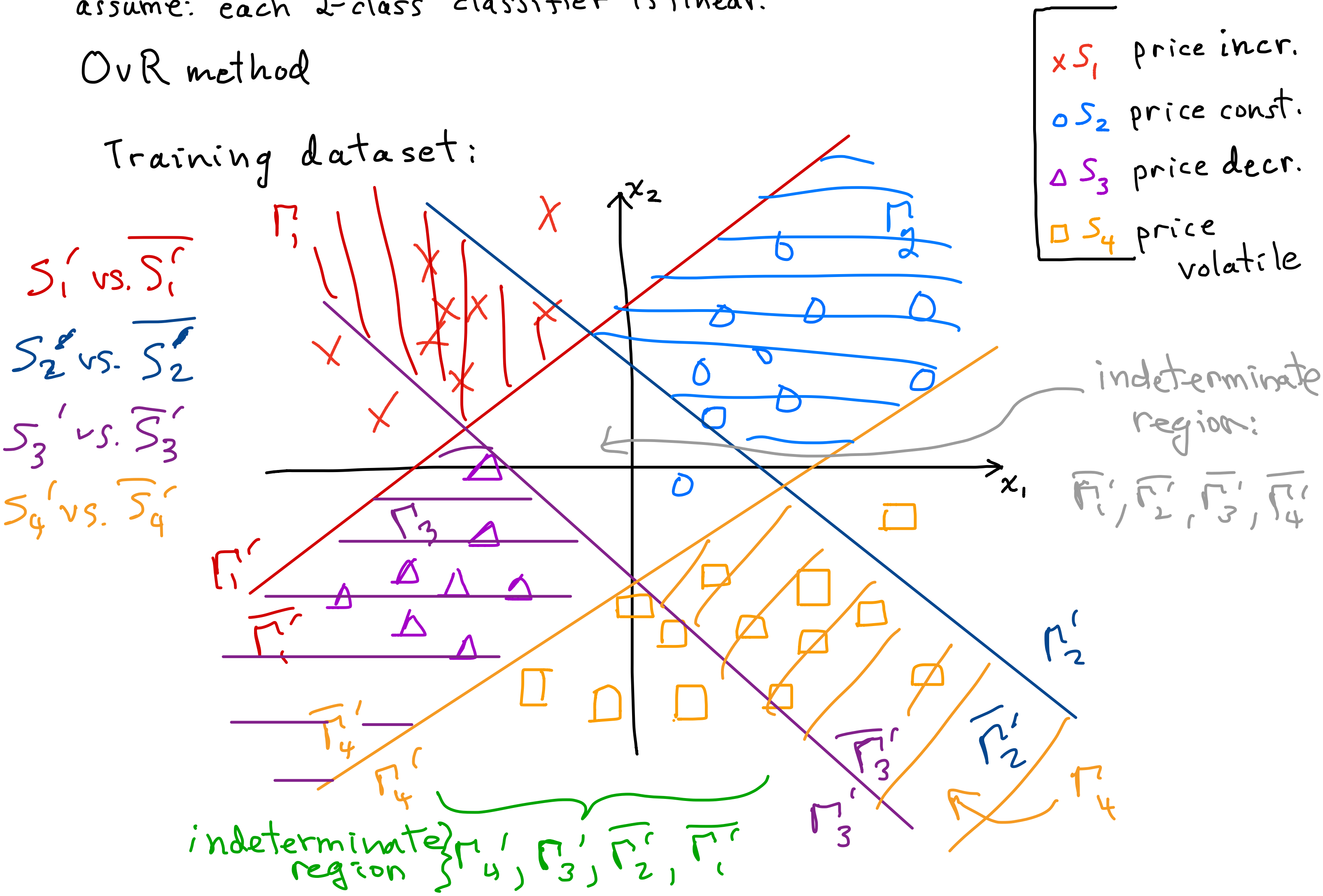
x
S_1' vs. $\bar{S}_1'$
Γ_1' or $\bar{\Gamma}_1'$
x
S_2' vs $\bar{S}_2'$
x
S_c' vs. $\bar{S}_c'$
Γ_c' or $\bar{\Gamma}_c'$
OvR default decision rule
$\underline{x} \in \Gamma_k$

Example: Consider a $C=4$-class problem with $D=2$ features

assume: each 2-class classifier is linear.

OvR method

Training dataset:



Apply 4 2-class problems: each S_k' vs $\overline{S_k'}$

Γ_1 is defined by: all $\underline{x} \in \left\{\Gamma_1' \text{ and } \overline{\Gamma_2'} \text{ and } \overline{\Gamma_3'} \text{ and } \overline{\Gamma_4'}\right\}$

(for our default OvR decision rule.)

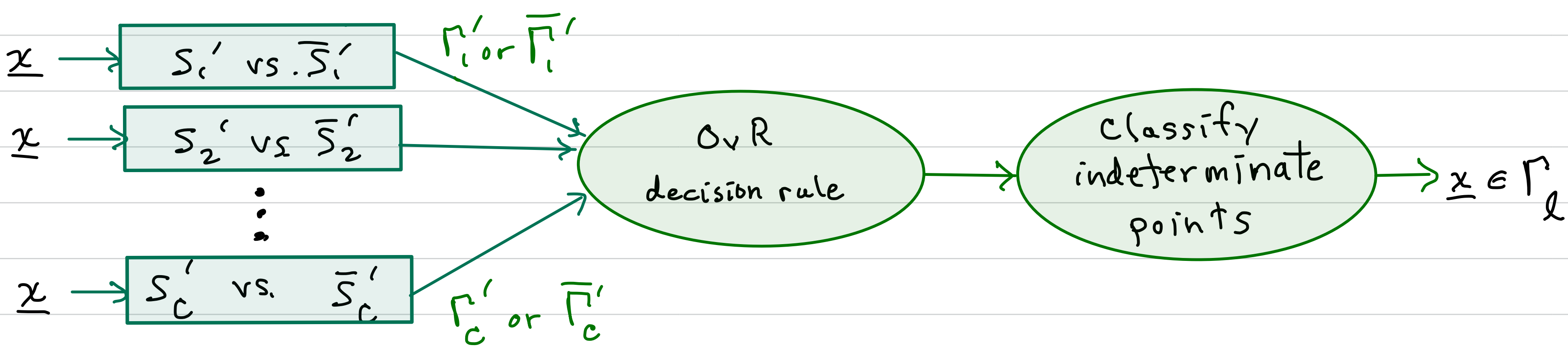
Comments on OvR method

1. **Def**: Given a dataset $\mathcal{D}$. If all data points $\underline{x}_i^{(k)}$ of class S_k be separated from all points of all other classes by a linear boundary, and this holds for all $k = 1, 2, \cdots, C$, then $\mathcal{D}$ is totally linearly separable.

2. In practice, for methods that can result in indeterminate regions, often an additional ad-hoc rule is used to classify the indeterminate points; for example:

For $\underline{x}$ in indeterminate region,

$$\underline{x} \in \Gamma_k \quad \text{iff} \quad k = \operatorname*{argmax}_k \left[g_k(\underline{x}) \right] \longrightarrow \underline{x} \in \Gamma_k$$



3. Alternate OvR decision rules can also be used. For example, if there is a confidence measure for each binary classification result for a given point $\underline{x}$:

$$\text{Confidence}(S_k' \text{ given } \underline{x})$$ (from S_k' vs $\overline{S_k'}$ classification of $\underline{x}$)

then final decision rule can be:

$$\boxed{\underline{x} \in \Gamma_{k^*} \text{ iff } \quad k^* = \underset{k}{\operatorname{argmax}} \left[\text{Confidence}(\Gamma_k' \text{ given } \underline{x})\right].}$$

Alternate OvR decision rule.

Examples of confidence measures:

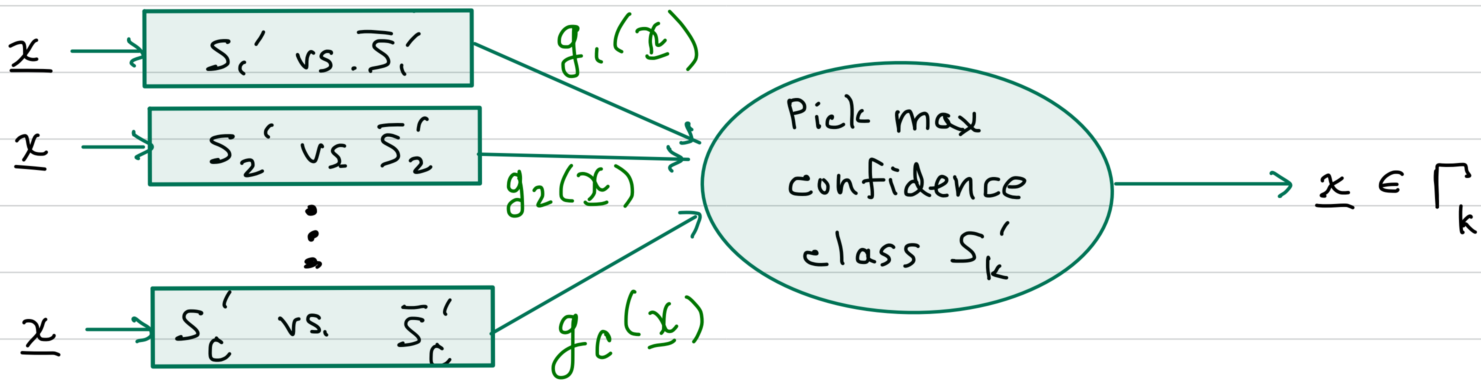
(i) $\text{Confidence}(\Gamma_k' \text{ given } \underline{x}) = g_k(\underline{x})^*$

(ii) $\text{Confidence}(\Gamma_k' \text{ given } \underline{x}) = P(S_k' \mid \underline{x})$

Confidence in S_k' vs. $\overline{S_k'}$ result, for input point $\underline{x}$.

* Assumes scale of $g_j(\underline{x})$ is comparable to $g_k(\underline{x})$.

In this case we have:



For case (i).

(ii) <u>One vs. one</u> (OvO) (or all vs. all) (or all pairs)

→ set of S_k vs. S_j decisions. (all possible pairs)

Use discr. fcns $g_{kj}(\underline{x})$. Each 2-class problem:

$$g_{kj} \underset{\Gamma_j}{\overset{\Gamma_k}{\gtrless}} 0 \quad ; \qquad g_{jk}(\underline{x}) = -g_{kj}(\underline{x})$$

Combine results:

<u>Decision rule</u> (OvO, default):

$$\underline{x} \in \Gamma_k \text{ iff } g_{kj}(\underline{x}) > 0 \quad \forall j \neq k.$$

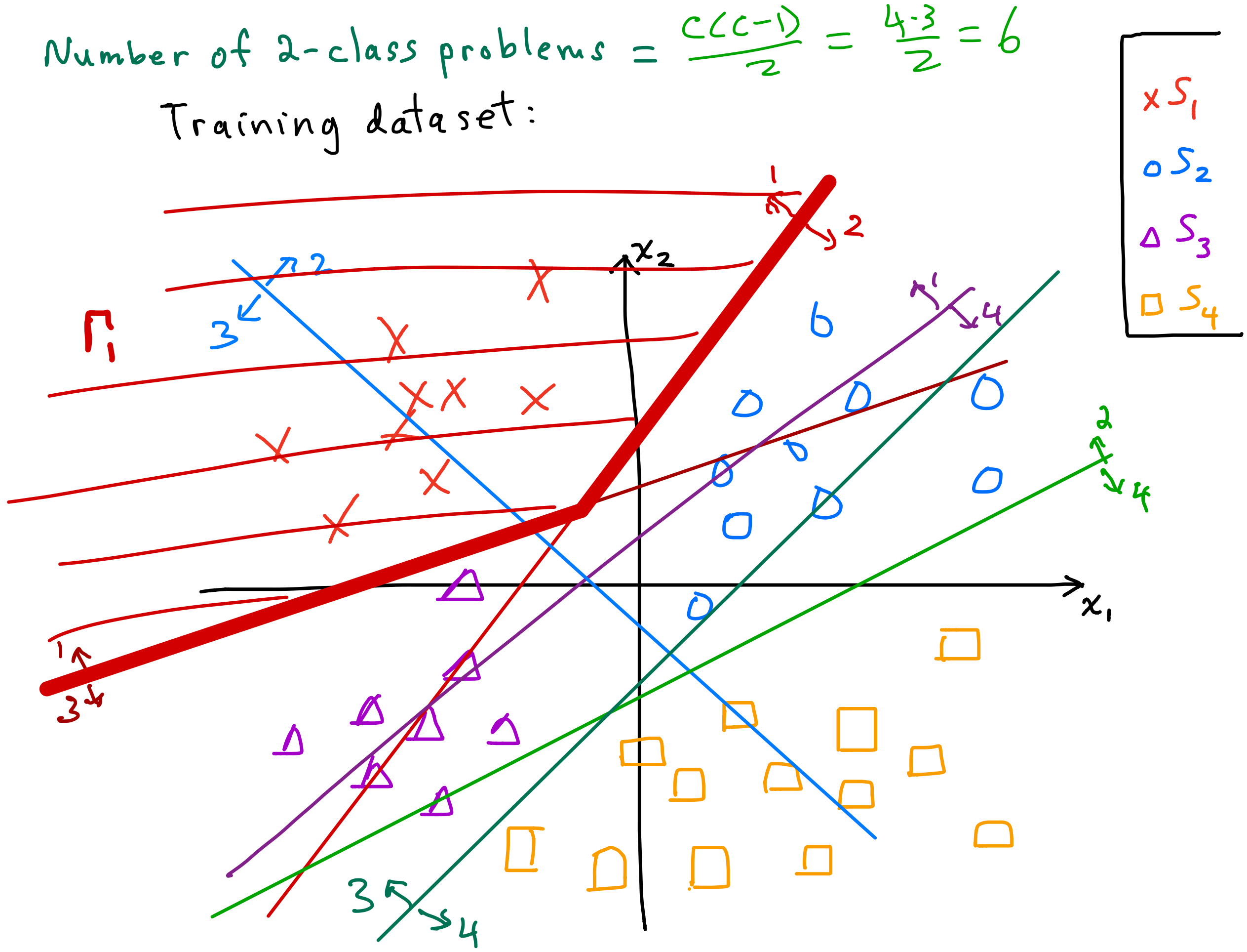
How many 2-class problems? $\binom{c}{2} = \dfrac{c(c-1)}{2}$

Example: Consider a $C=4$-class problem with $D=2$ features

Same data as previous example.

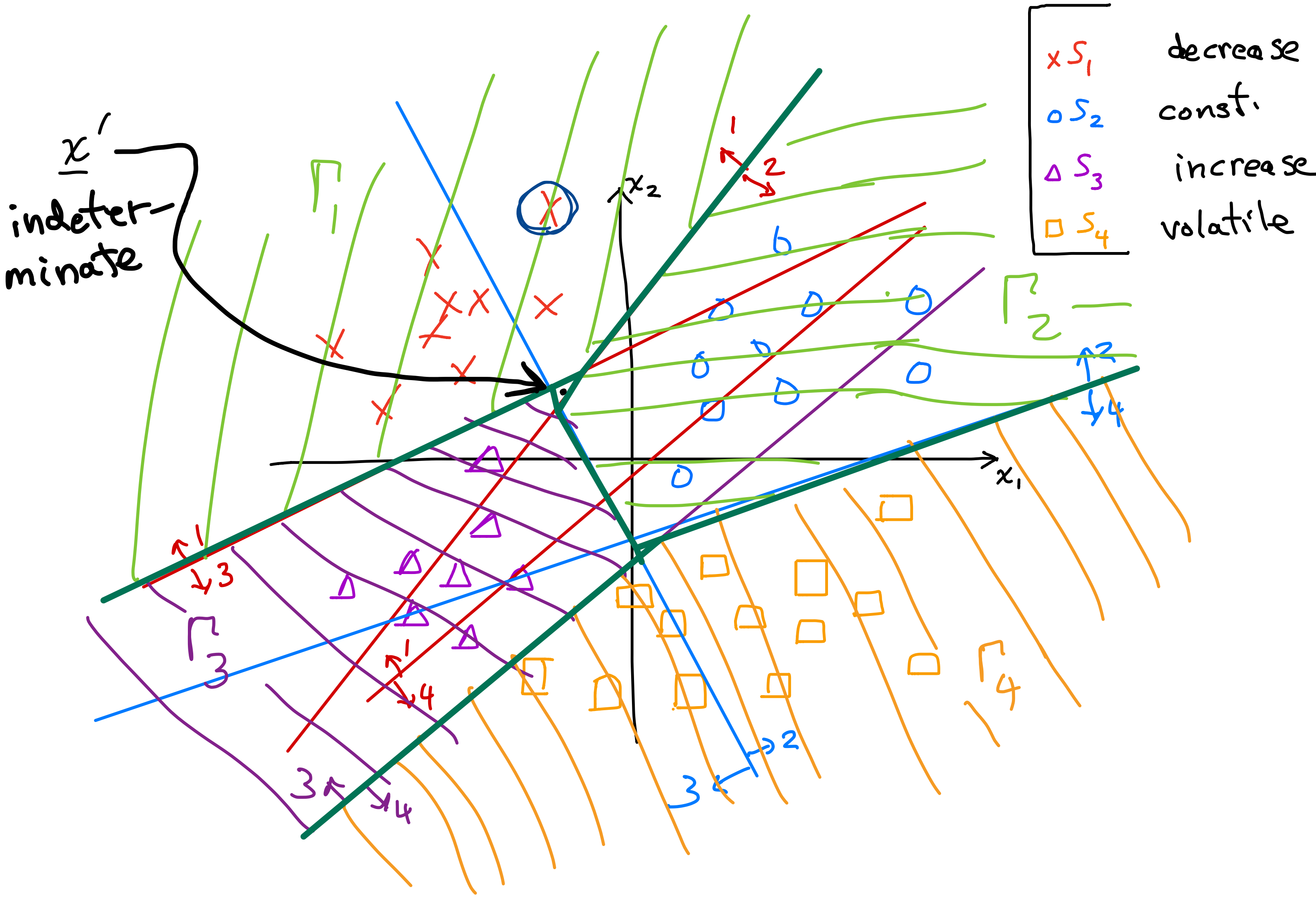
Number of 2-class problems $= \frac{C(C-1)}{2} = \frac{4\cdot 3}{2} = 6$

Training dataset:



One vs. one

At $\underline{x}'$: can't be: Γ_1, Γ_2, Γ_3, Γ_4 $\Rightarrow$ indeterminate



Comments (this example):

1. has indeterm. regions (smaller than OvR).
2. No errors on training data

Comments on OvO method

1. **Def:** If $\exists$ $\frac{c(c-1)}{2}$ linear separating boundaries H_{kj}, such that H_{kj} separates all data points of S_k from all data points of S_j, $j \neq k$, then the data is <u>pairwise linearly separable.</u>

2. For any indeterminate points, an additional ad hoc rule can be used to classify them.

3. An alternate decision rule for OvO:

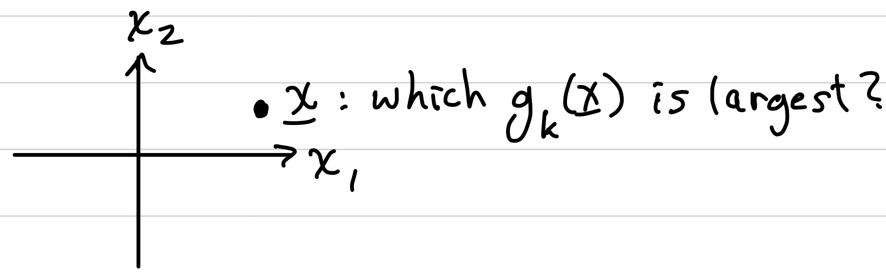
   Take vote: how many 2-class classifiers decided S_k' over S_j' ?
   $\underline{x} \in \Gamma_k$ if S_k' has the largest number of votes.

(iii) Maximal Value Method (MVM)

- for the linear case, called a linear machine.
- is not based on a set of 2-class classifiers.

Let each class S_k have 1 discriminant function: $g_k(\underline{x})$

Decision rule: $\quad g_k(\underline{x}) > g_j(\underline{x}) \;\; \forall j \neq k \;\; \Rightarrow \;\; \underline{x} \in \Gamma_k$



Decision rule can also be expressed as:

$$\boxed{\underline{x} \in \Gamma_k \;\; \text{iff} \;\; k = \arg\max_m \{ g_m(\underline{x}) \}}$$ MVM decision rule

$\Rightarrow$ Decision boundary between Γ_k and Γ_j is: $\quad g_k(\underline{x}) = g_j(\underline{x})$

(some boundaries may be redundant)

MVM

Learn:

Train $g_k(\underline{x})$, $k=1,2,\cdots,C$ together

$\downarrow$

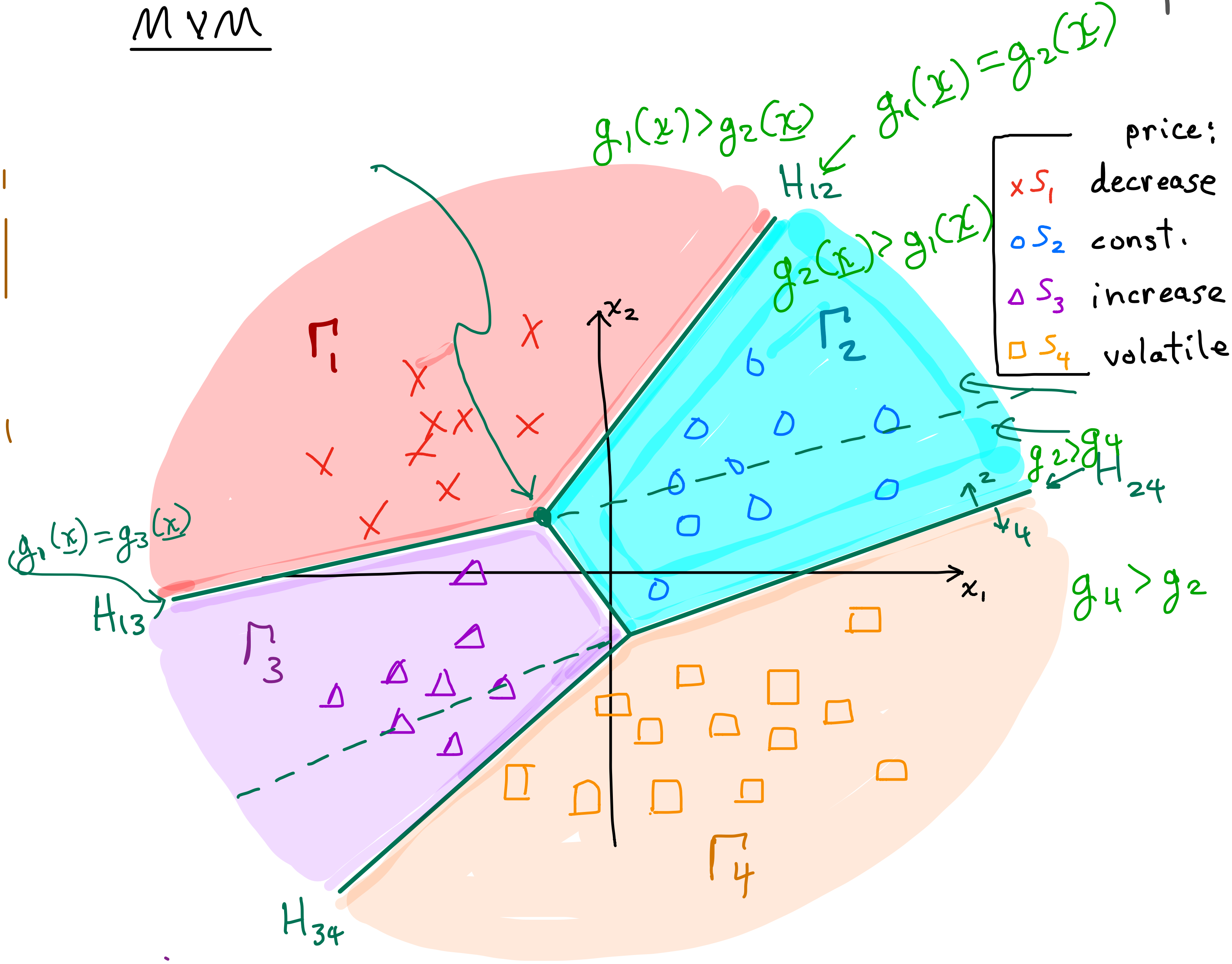
Predict:

Compute $g_k(\underline{x})$, $k=1,2,\cdots,C$

$g_1(\underline{x})$, $\vdots$, $g_C(\underline{x})$

$$k^* = \arg\max_k g_k(\underline{x}) \rightarrow \underline{x} \in \Gamma_{k^*}$$

MVM



$*$ $H_{jk}: \ g_j(\underline{x})=g_k(\underline{x})$

Comments on MVM

1. <u>Def</u>: If there exists $g_i(\underline{x})$, $i = 1, 2, \cdots, C$, such that

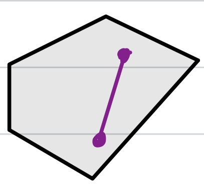
$$g_k\left[\underline{x}_m^{(k)}\right] > g_j\left[\underline{x}_m^{(k)}\right]$$

$$\forall\, m = 1, 2, \cdots, N_k, \quad \forall j \neq k, \quad \forall k$$

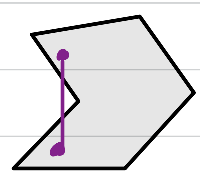
with all $g_i(\underline{x})$ expressible as linear functions, then the dataset is <u>linearly separable</u>.

2. No indeterminate regions (if unlikely special cases are avoided, (like $g_k(\underline{x}) = g_j(\underline{x})$ over some region).

3. If $g_k(\underline{x})$ are linear $\forall k$, then decision regions Γ_k are convex.

Convex  ?

Yes

Convex  ?

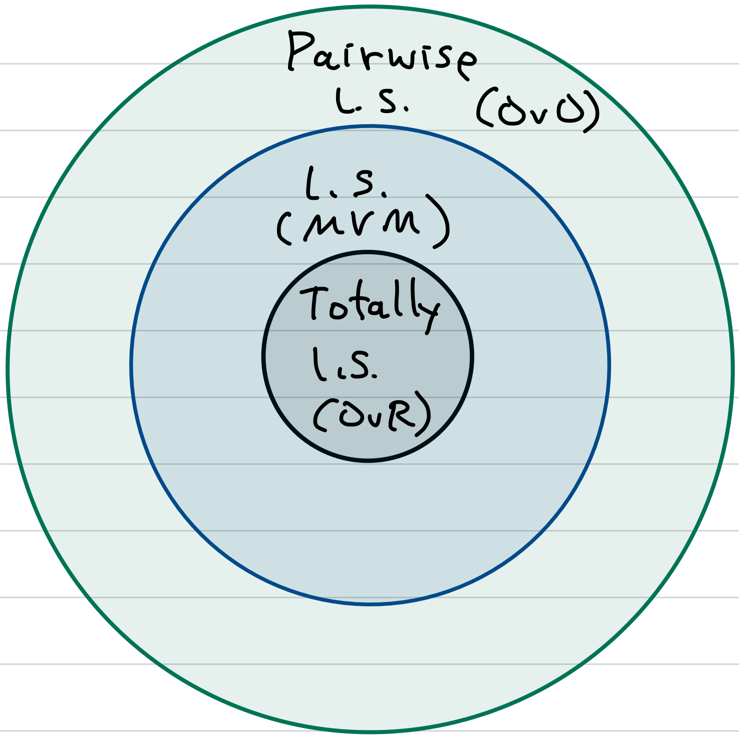
No.

Summary of multiclass classification approaches

Method	Linear $g(\underline{x})$ defines type of linear separability
OvR	Totally linearly separable
OvO	Pairwise linearly separable
MVM	Linearly separable

How powerful is each method?
→ How large is the class of datasets that each method can fully separate using linear $g(\underline{x})$?

→ Venn diagram:



Nonlinear case

Can use any of the above 3 methods (OvR, OvO, MVM), by letting each $g_k(\underline{x})$ or each $g_{kj}(\underline{x})$ be a nonlinear function of $\underline{x}$.

(More later)