Practical Introduction to Machine Learning

Dăscălescu Dana, March 2023

Scope

Objectives

- Guide participants through the steps required to solve real-world problems using machine learning
- Provide hands-on experience working with real-world data sets and popular ML tools

Focus

- Not about defining the field of ML or providing a deep understanding of the mathematical foundation of ML algorithms
- Focuses on guiding participants through the steps required to solve real-world problems using machine learning

Overview

Data preprocessing

Bag-of-Words:

- *A method of representing text data based on the frequency of occurrence words within documents*

Standardization

Model selection and evaluation

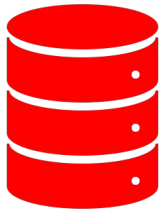
Support Vector Machines

Optimization

Finding the best performing model:

- Grid search

Problem definition



Data

MOROCO –The Moldavian and Romanian Dialectal Corpus

Text collected from the news domain

The samples belong to one of the following six topics: (0) culture, (1) finance, (2) politics, (3) science, (4) sports, (5) tech

Process

Prepare the data: transform the text data into a numerical representation

Pre-process the data: e.g. normalization

Train a classifier

Optimize hyperparameters

Output

For an unseen text document, predict the main category to which it belongs

Data preprocessing – Bag-of-words model

- This is the first sentence
- This sentence is the second sentence
- And this is the third sentence
- Is this the first sentence

Data preprocessing – Bag-of-words model

- This is the first sentence
- This sentence is the second sentence
- And this is the third sentence
- Is this the first sentence

this is the first sentence second and third

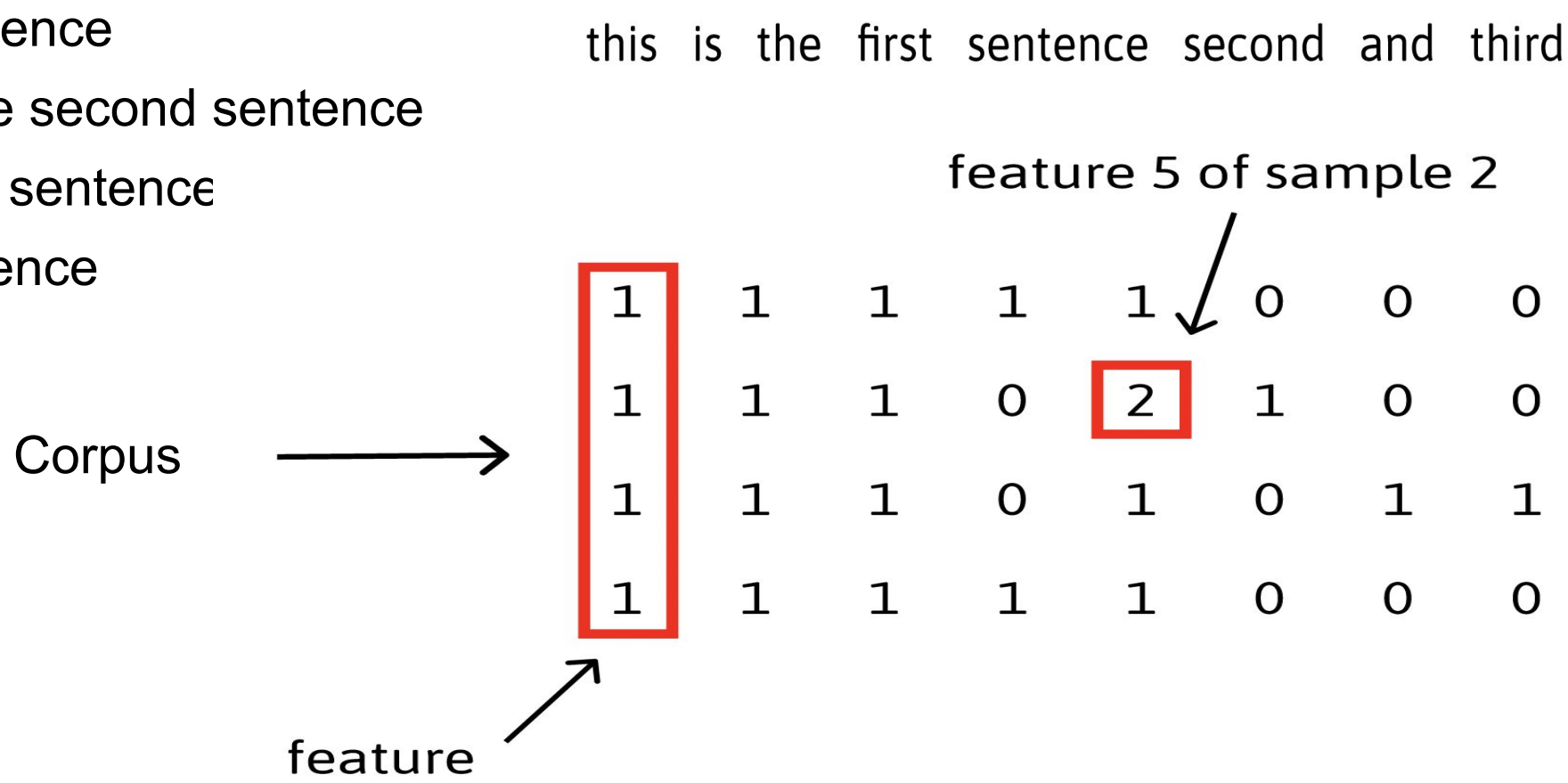
Data preprocessing – Bag-of-words model

- This is the first sentence
- This sentence is the second sentence
- And this is the third sentence
- Is this the first sentence

this	is	the	first	sentence	second	and	third
1	1	1	1	1	0	0	0
1	1	1	0	2	1	0	0
1	1	1	0	1	0	1	1
1	1	1	1	1	0	0	0

Data preprocessing – Bag-of-words model

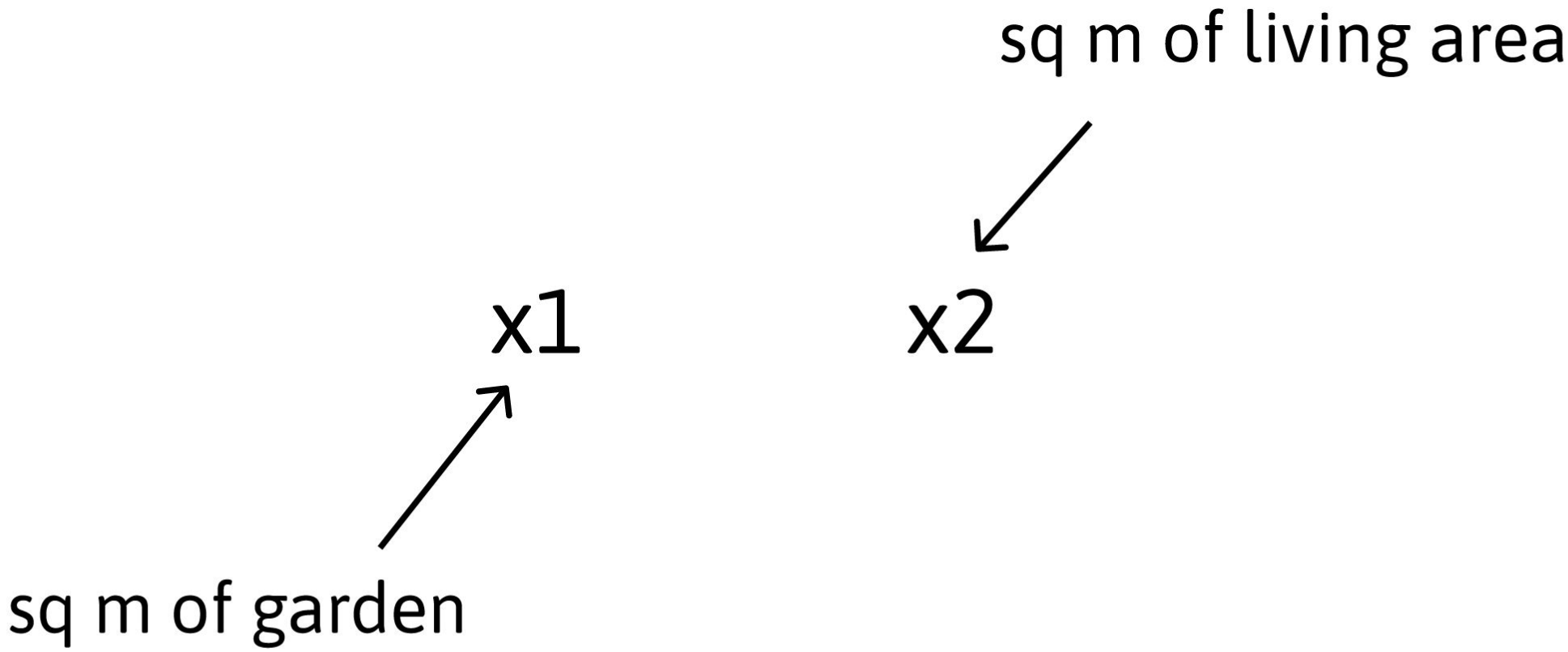
- This is the first sentence
- This sentence is the second sentence
- And this is the third sentence
- Is this the first sentence



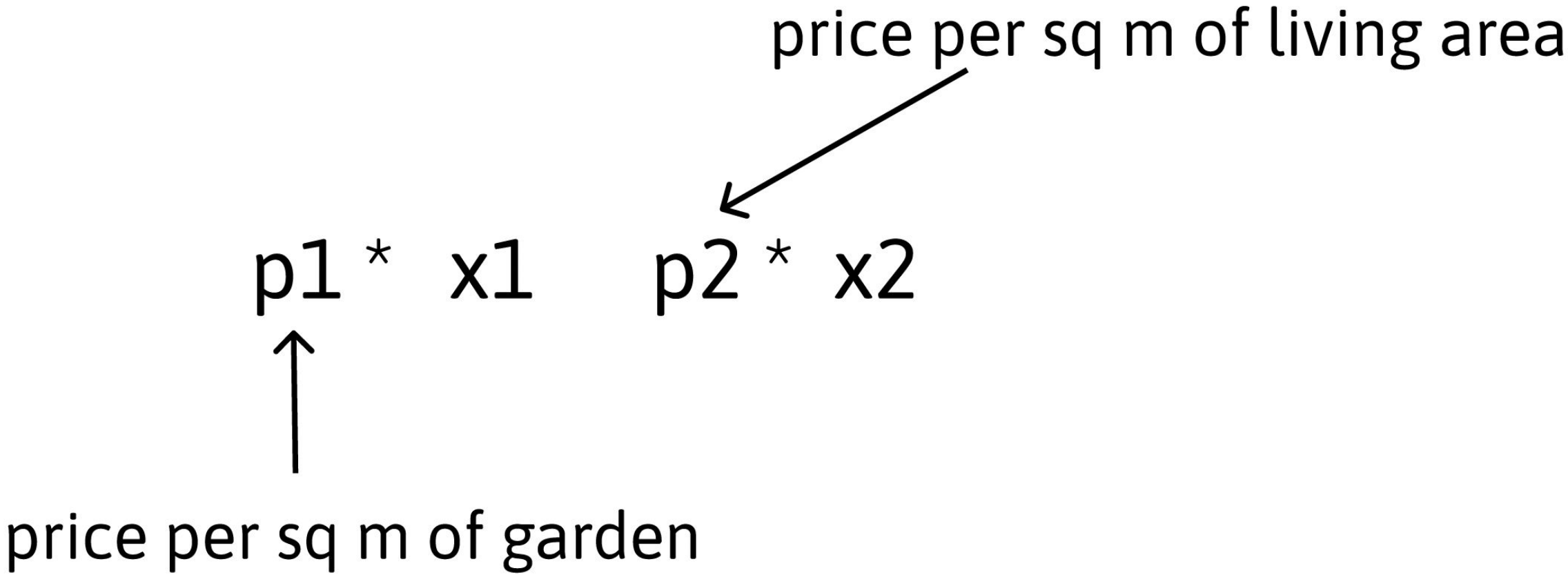
How do we perform classification?

Let's take a more practical example

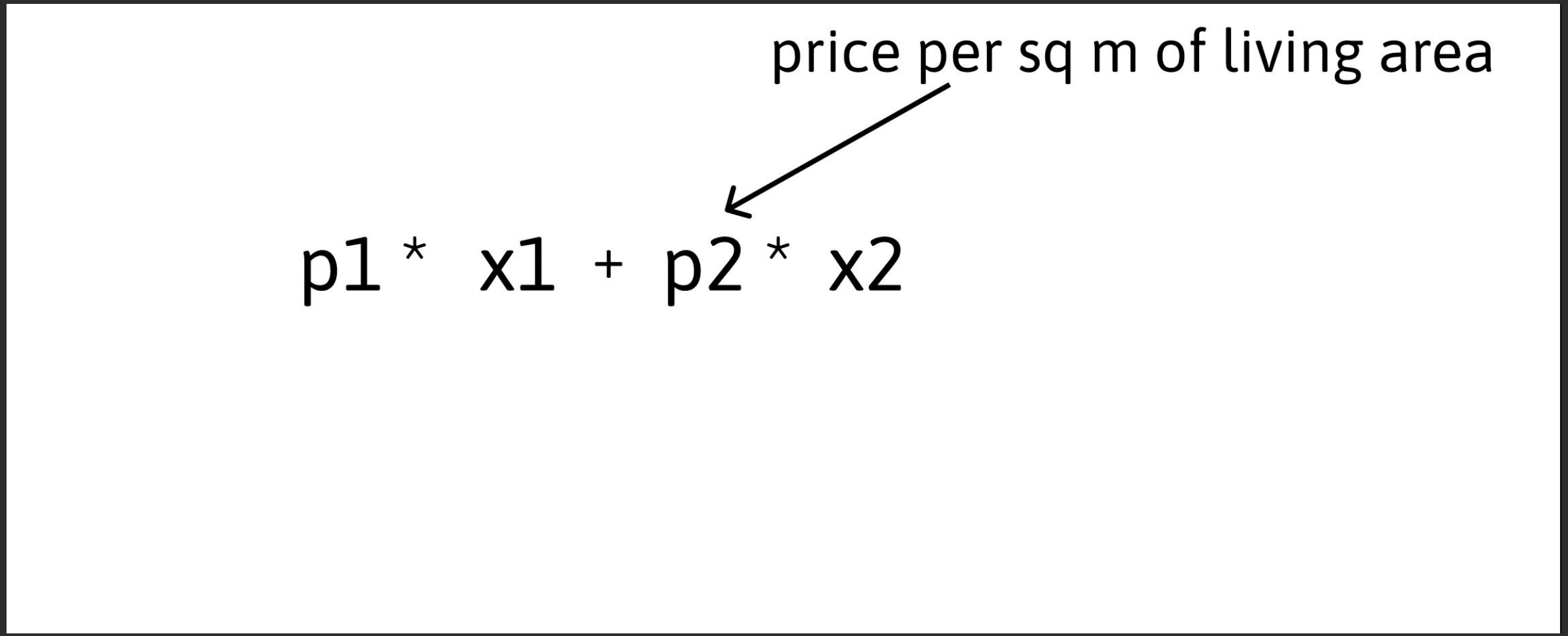
Price house prediction



Price house prediction

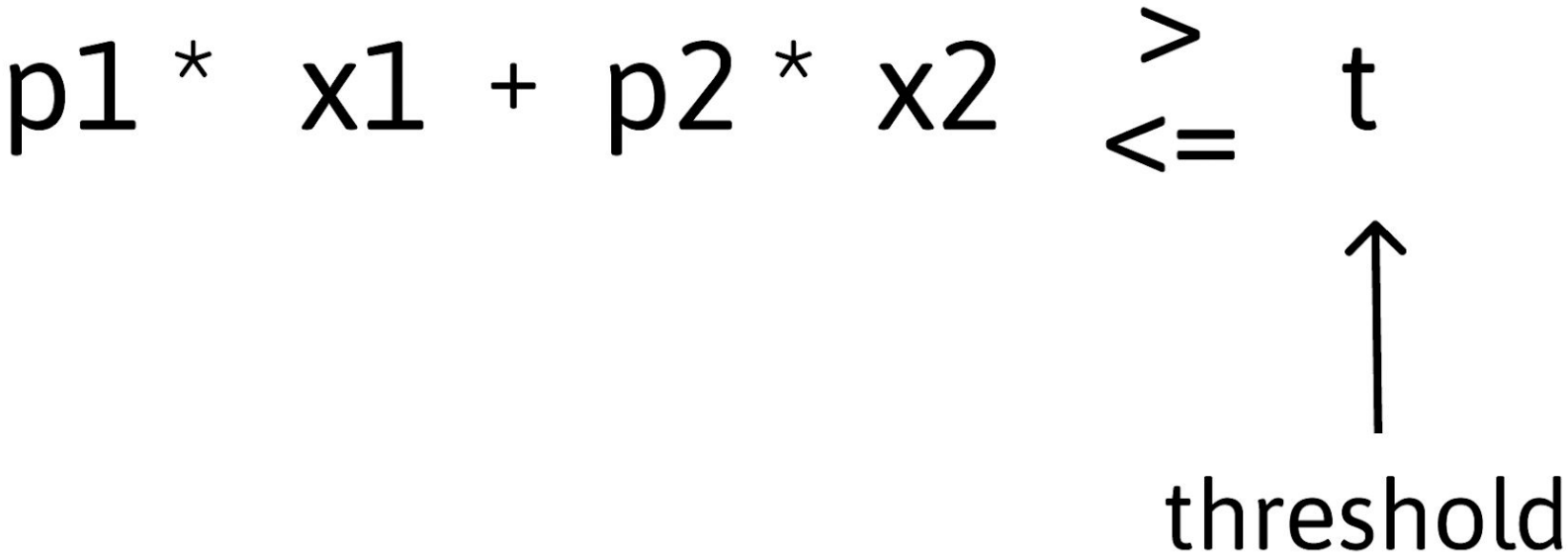


Price house prediction



Price house prediction

Decide whether is an expensive house or not?



Price house prediction

Decide whether is an expensive house or not?

$$p1 * x1 + p2 * x2 - t \begin{matrix} > \\ <= \end{matrix} 0$$

Price house prediction

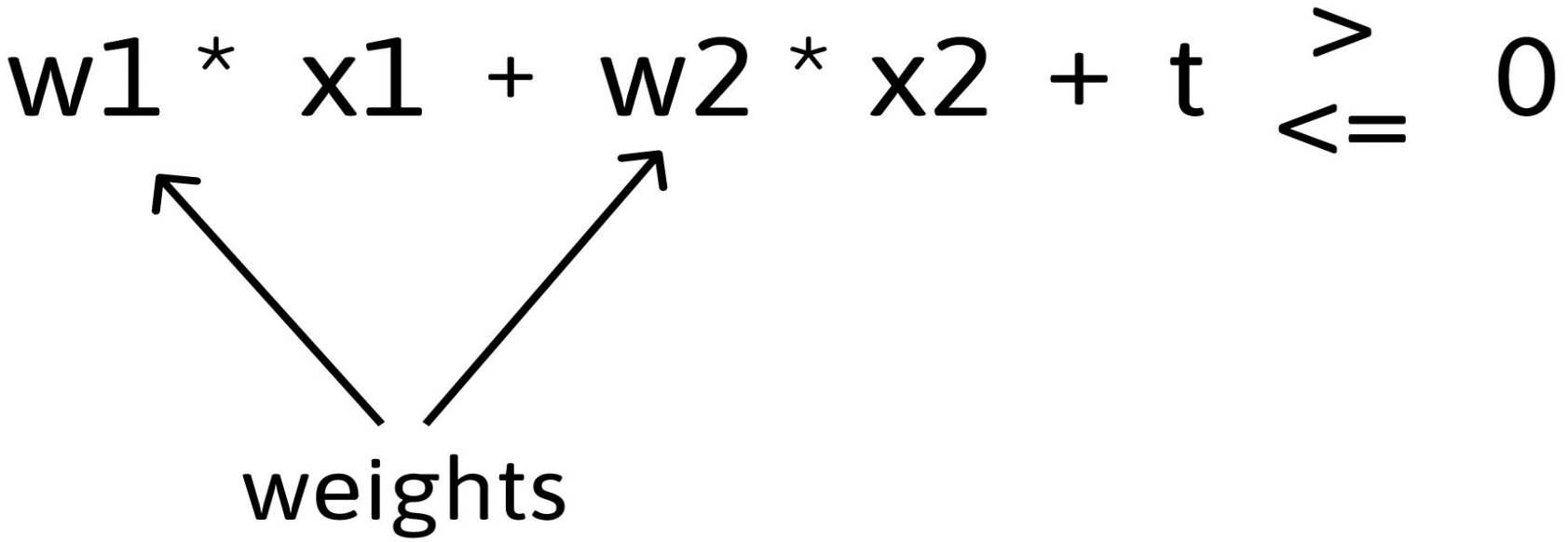
Decide whether is an expensive house or not?

$$p1 * x1 + p2 * x2 + t \gtrless 0$$

sign doesn't matter

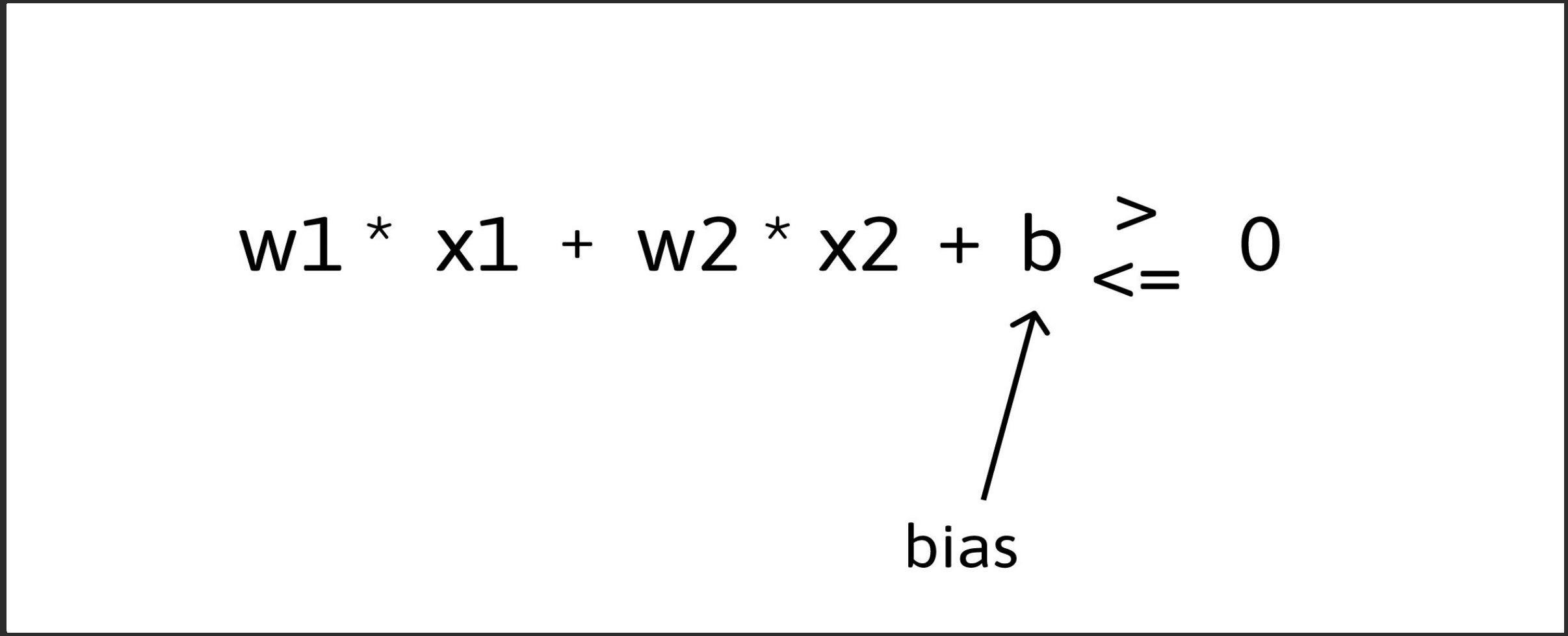
Price house prediction

Decide whether is an expensive house or not?



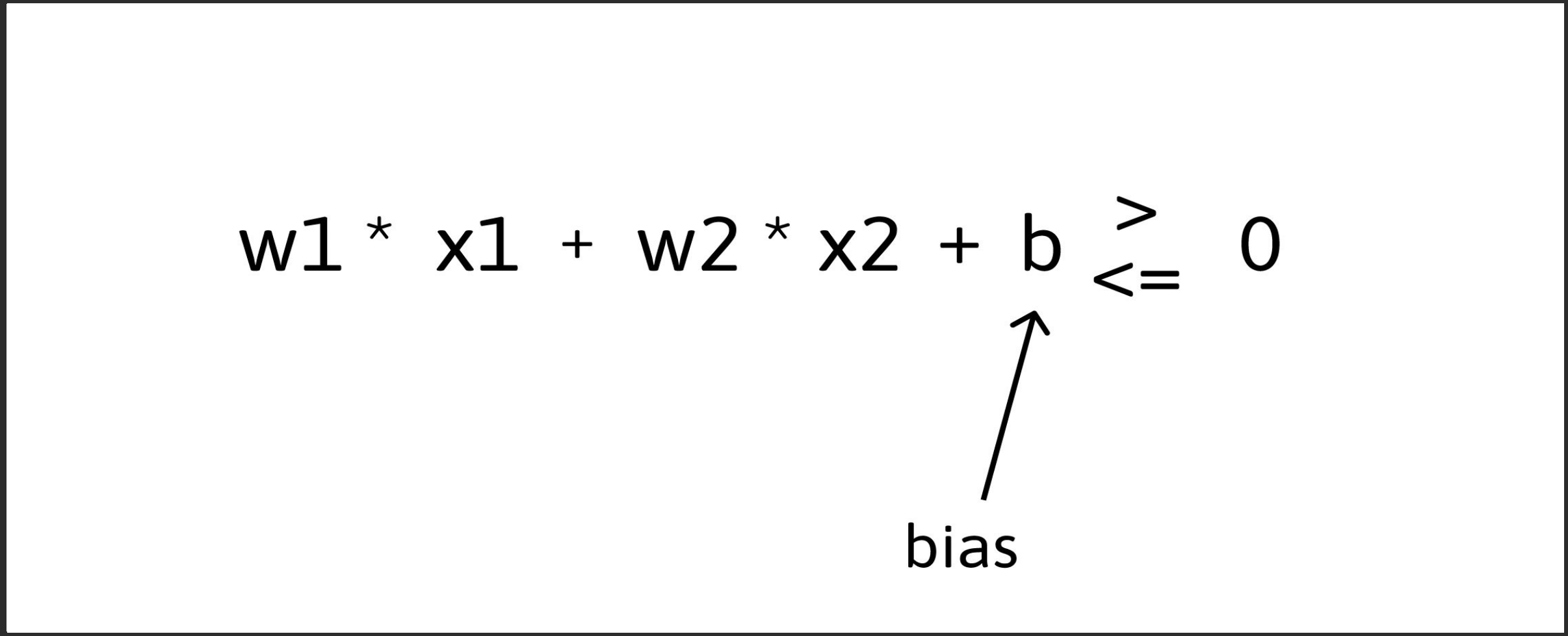
Price house prediction

Decide whether is an expensive house or not?



Price house prediction

Decide whether is an expensive house or not?



Price house prediction
Solving the problem

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Test set

(15, 7)

Price house prediction
Solving the problem

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Determine w1, w2 and b such that

w1 * 20 + w2 * 10 + b > 0 and
w1 * 10 + w2 * 10 + b > 0 and
w1 * 5 + w2 * 5 + b <= 0
at the same time

Test set

(15, 7)

Price house prediction
Solving the problem

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Determine w1, w2 and b such that

w1 * 20 + w2 * 10 + b > 0 and
w1 * 10 + w2 * 10 + b > 0 and
w1 * 5 + w2 * 5 + b <= 0
at the same time

Test set

(15, 7)

See whether
w1 * 15 + w2 * 7 + b $\begin{matrix} > \\ <= \end{matrix}$ 0

Price house prediction
Solving the problem

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Determine w1, w2 and b such that

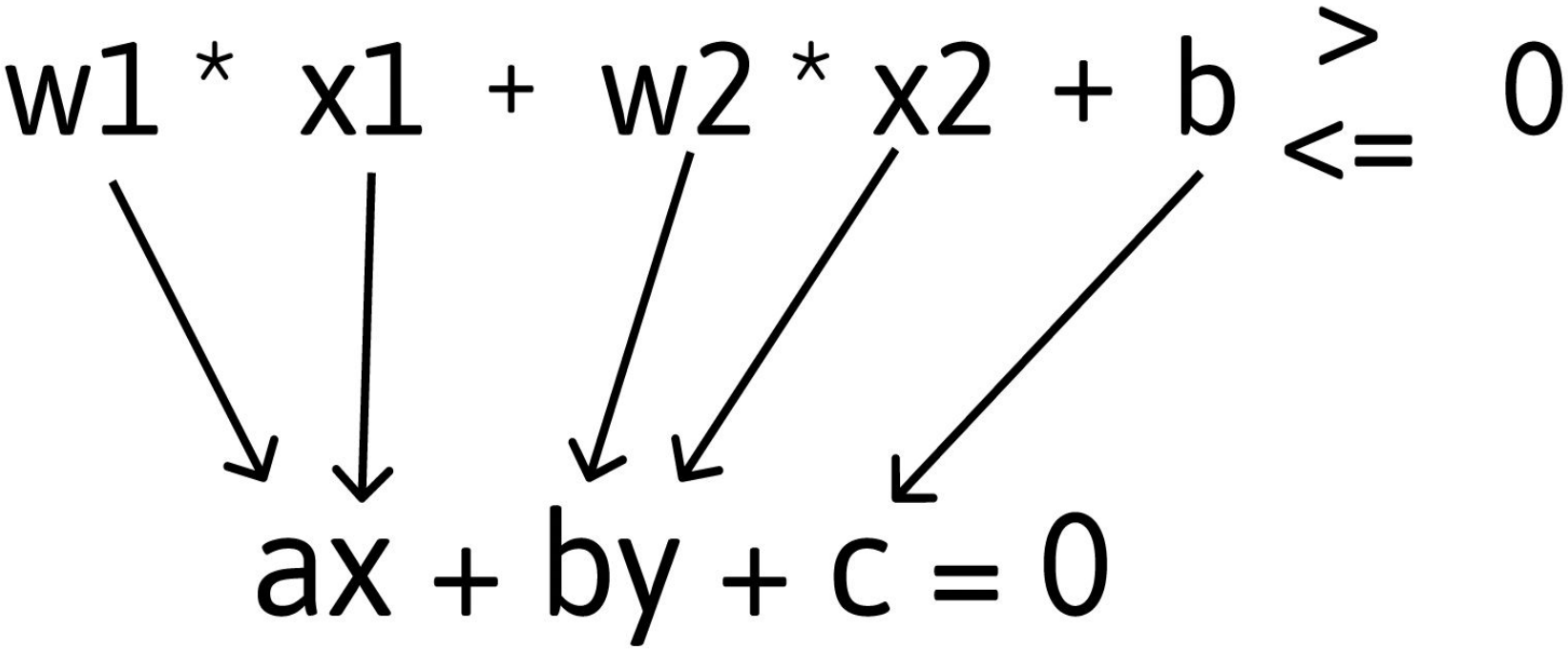
w1 * 20 + w2 * 10 + b > 0 and
w1 * 10 + w2 * 10 + b > 0 and
w1 * 5 + w2 * 5 + b <= 0
at the same time

Test set

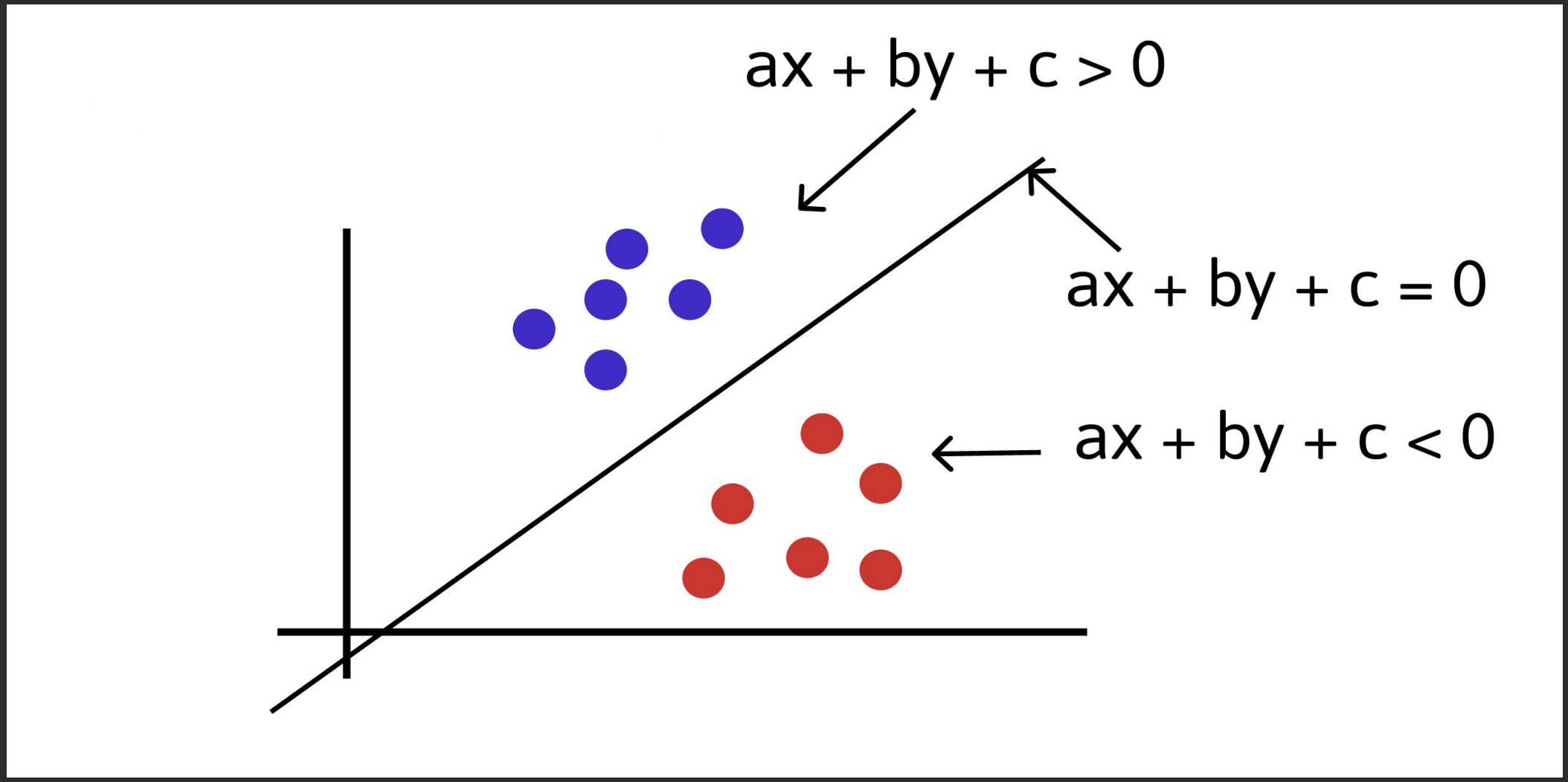
(15, 7)

See whether
w1 * 15 + w2 * 7 + b $\begin{matrix} > \\ <= \end{matrix}$ 0

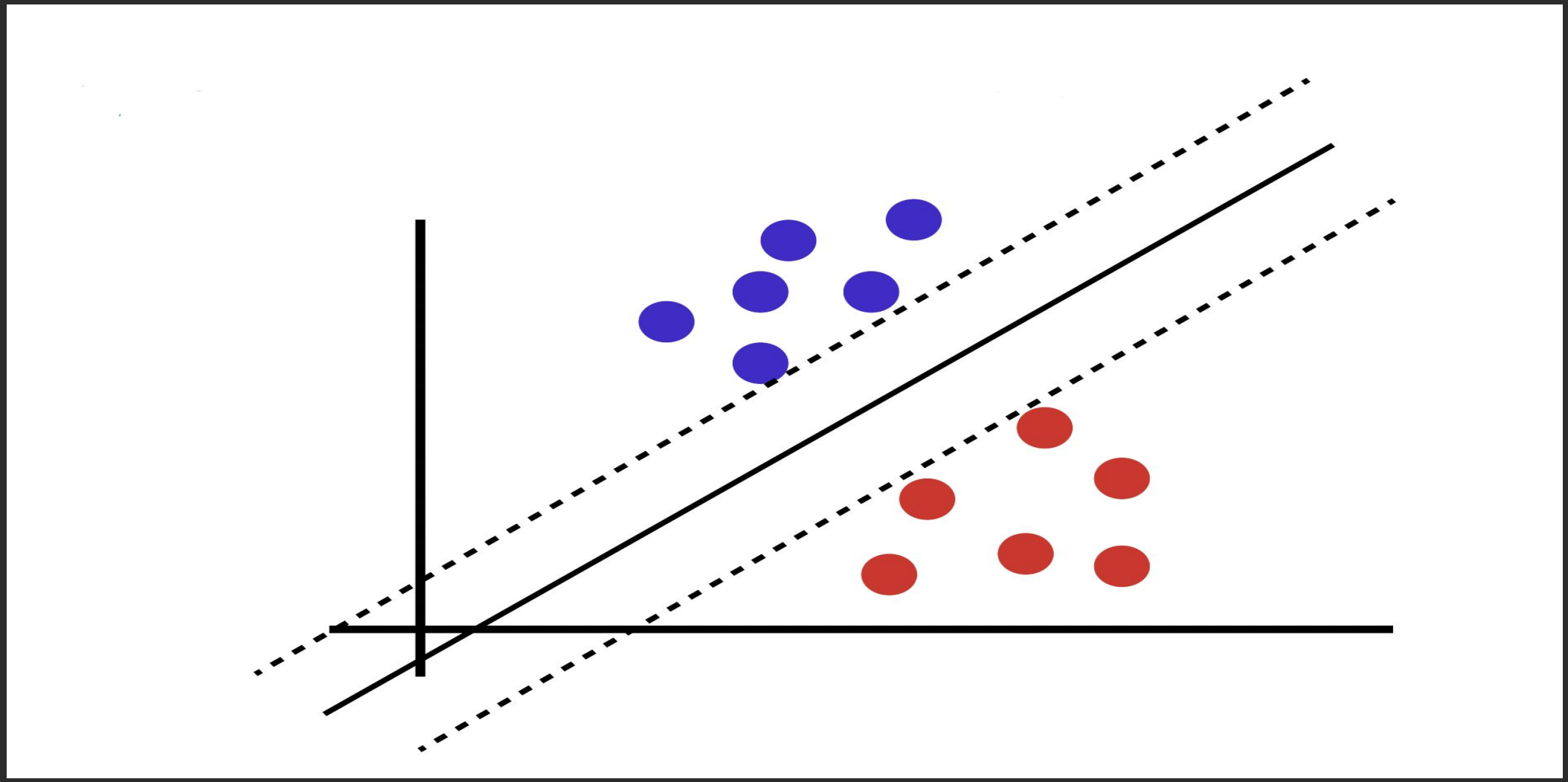
Price house prediction
Solving the problem



Price house prediction
Solving the problem



Support Vector Machines



Support Vector Machines

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Determine w1, w2 and b such that

$w1 * 20 + w2 * 10 + b > 0$ and
$w1 * 10 + w2 * 10 + b > 0$ and
$w1 * 5 + w2 * 5 + b <= 0$
at the same time

While minimizing $||w||$, where $w = (w1, w2, b)$

Support Vector Machines

Training set

(20, 10) 1
(10, 10) 1
(5, 5) 0

Determine w1, w2 and b such that

$w1 * 20 + w2 * 10 + b >= 1$ and
$w1 * 10 + w2 * 10 + b >= 1$ and
$w1 * 5 + w2 * 5 + b <= -1$
at the same time

While minimizing $||w||$, where $w = (w1, w2, b)$

How do we know if our model is good? (Validation)

- Split the available annotated data, e.g. 90% for training, 10% for validation
- Train the model on the training data and test the current configuration on the validation data
- Choose the model with the best performance on the validation data

Tips for competition: You can also train the best performing model on all the data (train + val) before submitting

Q&A

Thank you!