the fastest possible intro to machine learning

Machine Learning develops and uses algorithms to make predictions from data algorithms

predictions data

data

Think data in a relational table or spreadsheet

- Rows are “Data Instances” (person, for example)
- Columns represent “Features” (height, weight, gender)
- The set of features for a data instance is a “Feature Vector”

$$v1 = (0,0,254,180,180)$$

$$v2 = (0,0,254,135,215)$$

data

Think data in a relational table or spreadsheet

- Rows are “Data Instances” (person, for example)
- Columns represent “Features” (height, weight, gender)
- The set of features for a data instance is a “Feature Vector”

$$v1 = (M, 72, 180, (120,80), 50000, S)$$

$$v2 = (F, 70, 130, (120,75), 75000, M)$$

data

Think data in a relational table or spreadsheet

- Rows are “Data Instances” (person, for example)
- Columns represent “Features” (height, weight, gender)
- The set of features for a data instance is a “Feature Vector”
- You may have hundreds of millions of Data Instances
- You may have many features
- Training data has the answers marked
- Test dataset has known answers but is unmarked

$$v1 = (1, 72, 180, (120{,}80), 50000, 0)$$

$$v2 = (2, 70, 130, (120{,}75), 75000, 1)$$

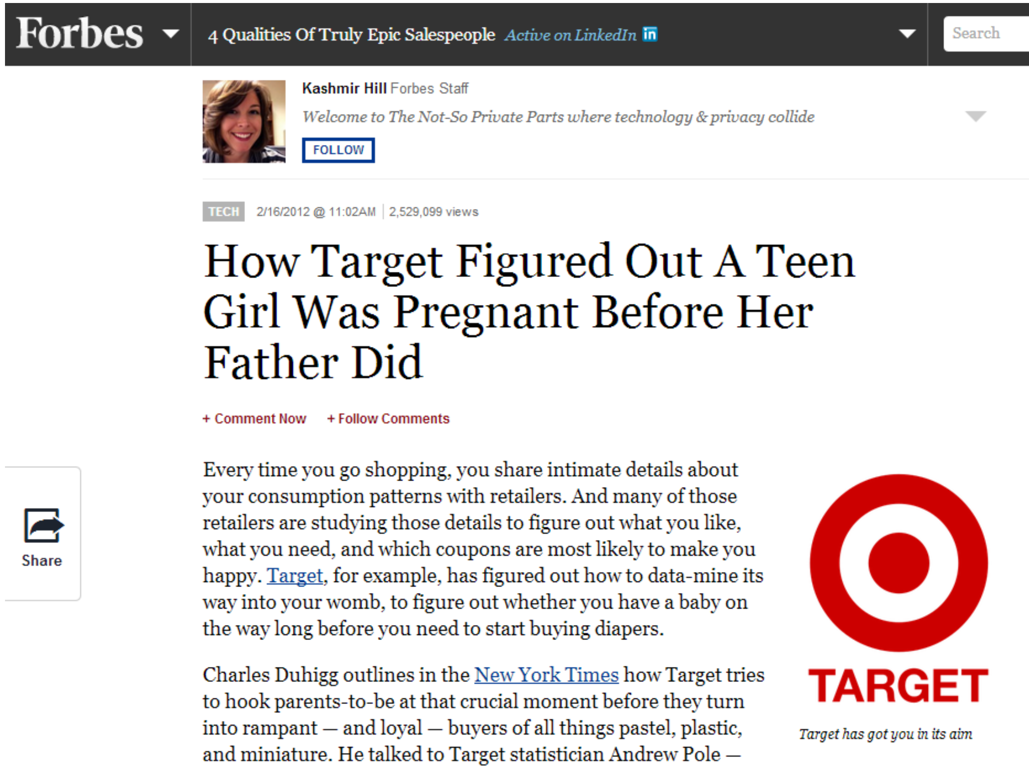
data:features

A measurable heuristic property

- Also called “attributes”
- Features turn a “thing” into a vector
- Based on experience
- Features are designed specifically for a dataset and desired prediction
- Much of the magic of ML is choosing the right feature vectors

prediction Think generalization, not telling the future

- Use a learning set to generalize
- Use generalizations to perform accurately on new, unseen examples


Forbes
4 Qualities Of Truly Epic Salespeople *Active on LinkedIn* in
Search
FOLLOW
Share
+ Comment Now
+ Follow Comments
TARGET

Kashmir Hill Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

TECH 2/16/2012 @ 11:02AM | 2,529,099 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

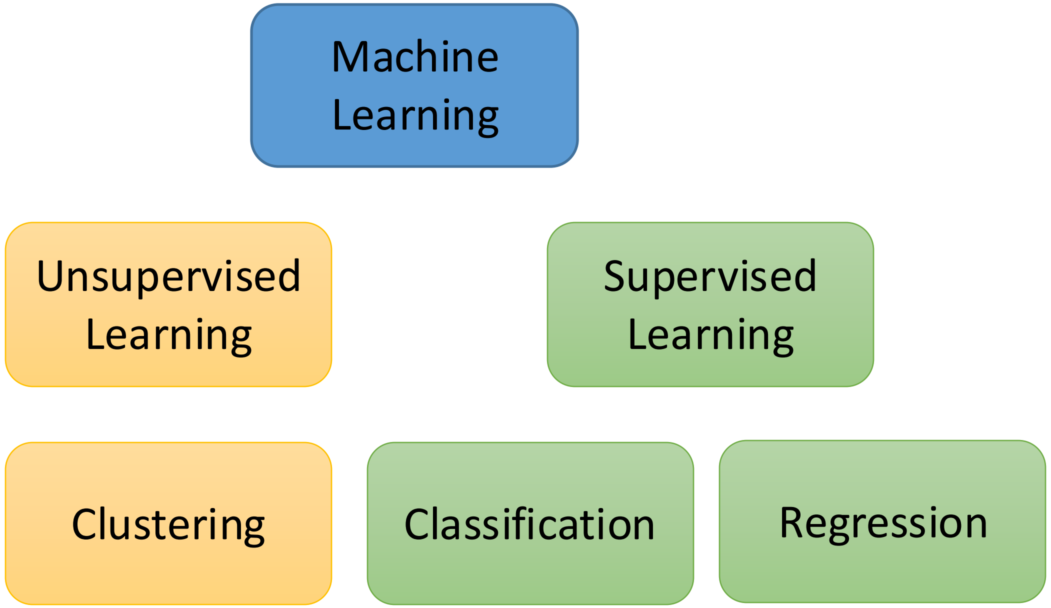
Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole —

Target has got you in its aim

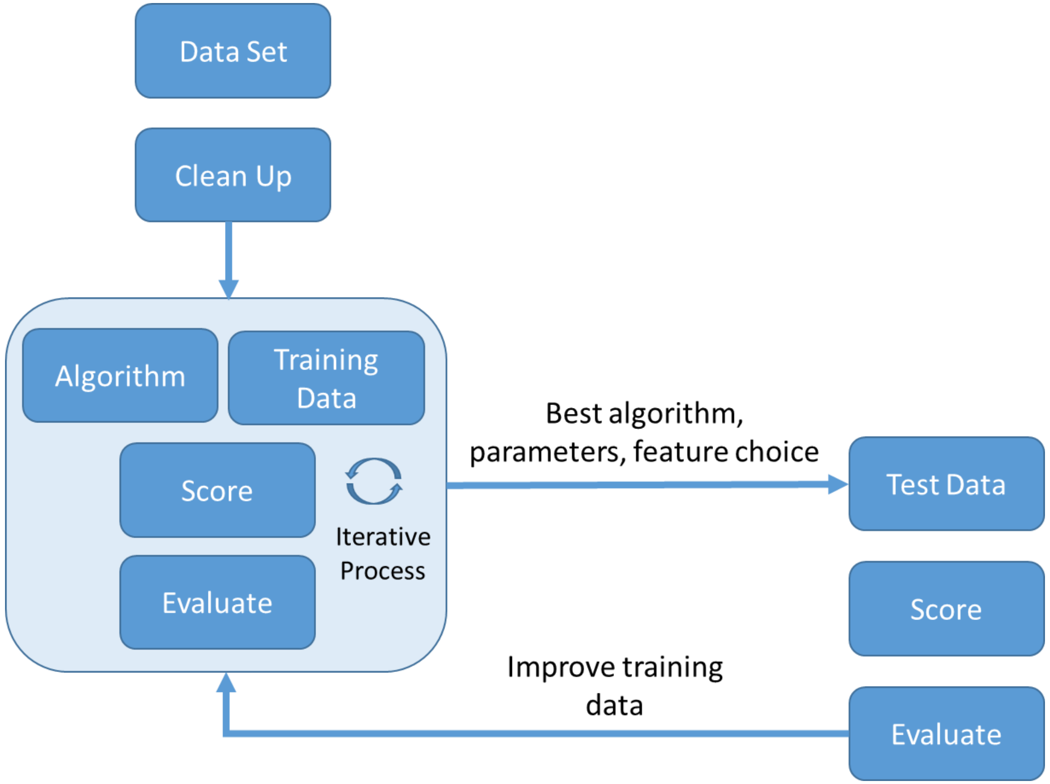
algorithm What do you want to do?

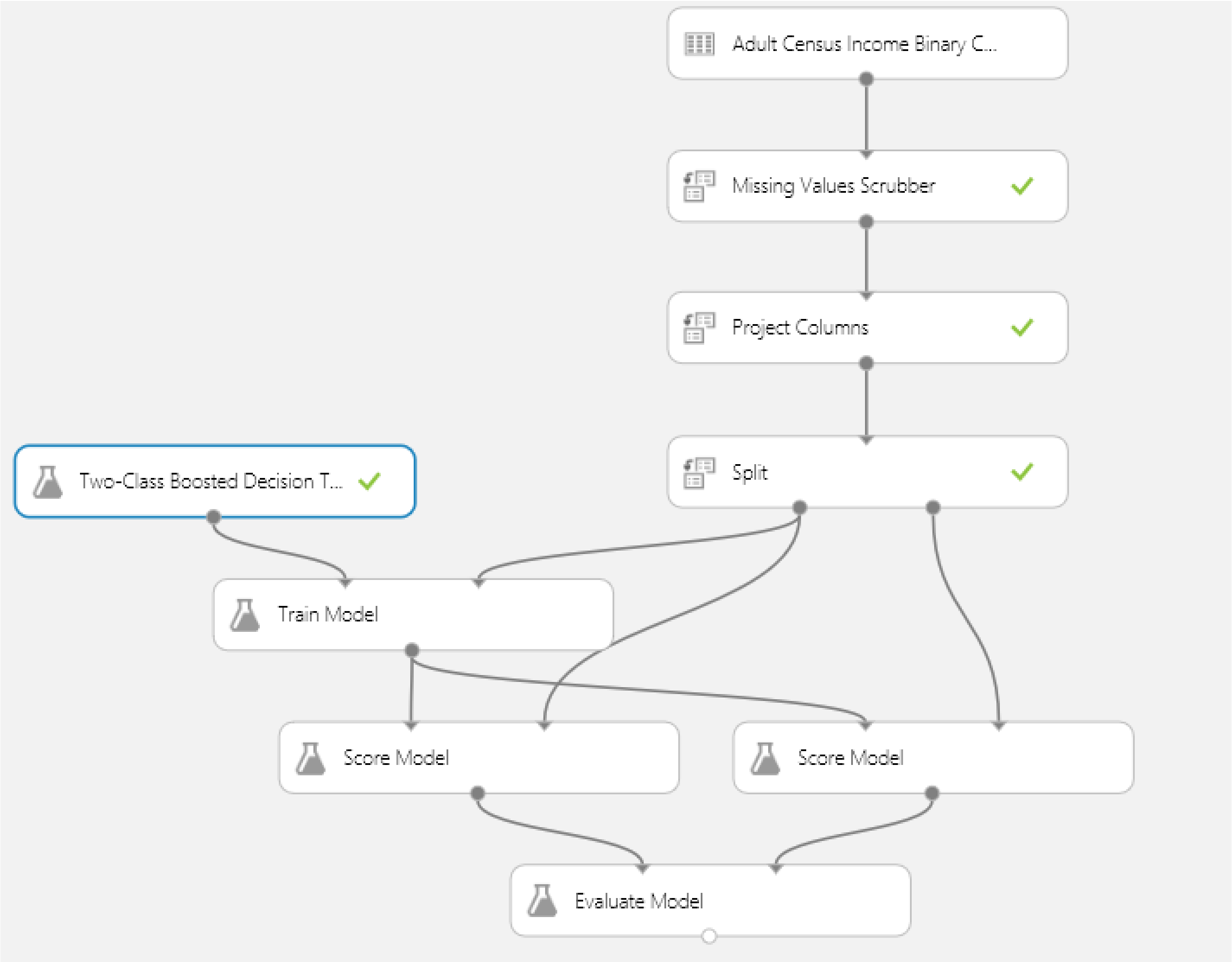
- Unsupervised learning tries to find structure in unlabeled data
- Supervised learning infers a function from labeled training data



the general process

- You have a dataset
- Clean up your dataset (project columns, hide columns)
- Separate dataset into training data and test data
- Choose an algorithm
- Train your algorithm with the training data
- Iterate until you have the best combination of algorithm, parameters, feature vectors
- Test your algorithm on your test data
- Analyze results
- Wash rinse repeat




Adult Census Income Binary C...
Missing Values Scrubber ✓
Project Columns ✓
Two-Class Boosted Decision T... ✓
Split ✓
Train Model
Score Model
Score Model
Evaluate Model

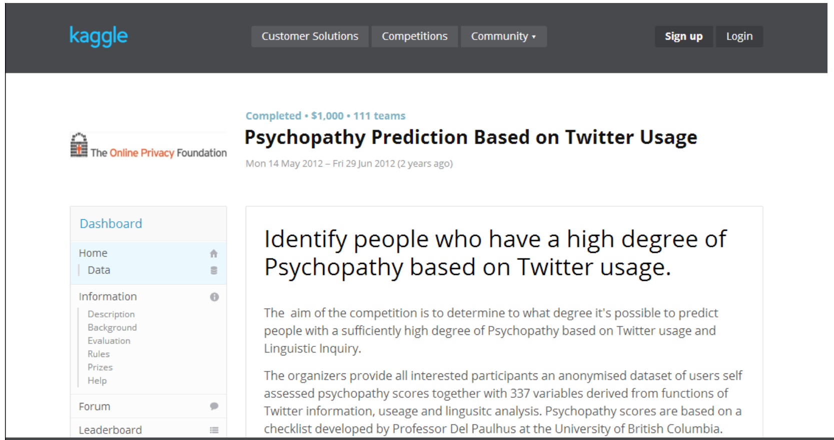
Data Set & Cleanup

Train algorithm, iterate on test data

Test the best and improve

what is ml good for?

- Classification (credit scoring, spam detection)
- Sentiment analysis
- OCR, Speech recognition
- Serving ads
- Finding psychos on Twitter (e.g. @jpalioto)
- Spying on you



Completed • $1,000 • 111 teams

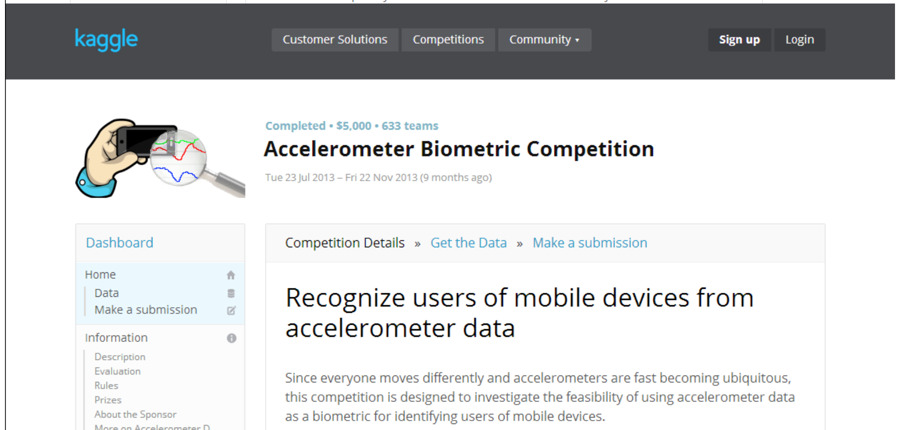
Psychopathy Prediction Based on Twitter Usage

Mon 14 May 2012 – Fri 29 Jun 2012 (2 years ago)

Identify people who have a high degree of Psychopathy based on Twitter usage.

The aim of the competition is to determine to what degree it's possible to predict people with a sufficiently high degree of Psychopathy based on Twitter usage and Linguistic Inquiry.

The organizers provide all interested participants an anonymised dataset of users self assessed psychopathy scores together with 337 variables derived from functions of Twitter information, useage and lingusitc analysis. Psychopathy scores are based on a checklist developed by Professor Del Paulhus at the University of British Columbia.



Completed • $5,000 • 633 teams

Accelerometer Biometric Competition

Tue 23 Jul 2013 – Fri 22 Nov 2013 (9 months ago)

Competition Details » Get the Data » Make a submission

Recognize users of mobile devices from accelerometer data

Since everyone moves differently and accelerometers are fast becoming ubiquitous, this competition is designed to investigate the feasibility of using accelerometer data as a biometric for identifying users of mobile devices.

messaging

What can we tell our communities?

- Microsoft is an innovator in the ML space
- ML Studio is unique among cloud providers
- ML Studio allows iterative, predictive analytics that would have taken weeks or months to be done in days
- ML Studio brings ML to everyone

how can I learn more?

- 20 part lecture from Stanford on Youtube (bring your linear algebra knowledge)
- MIT Opencourseware and videos
- Lots of free ebooks

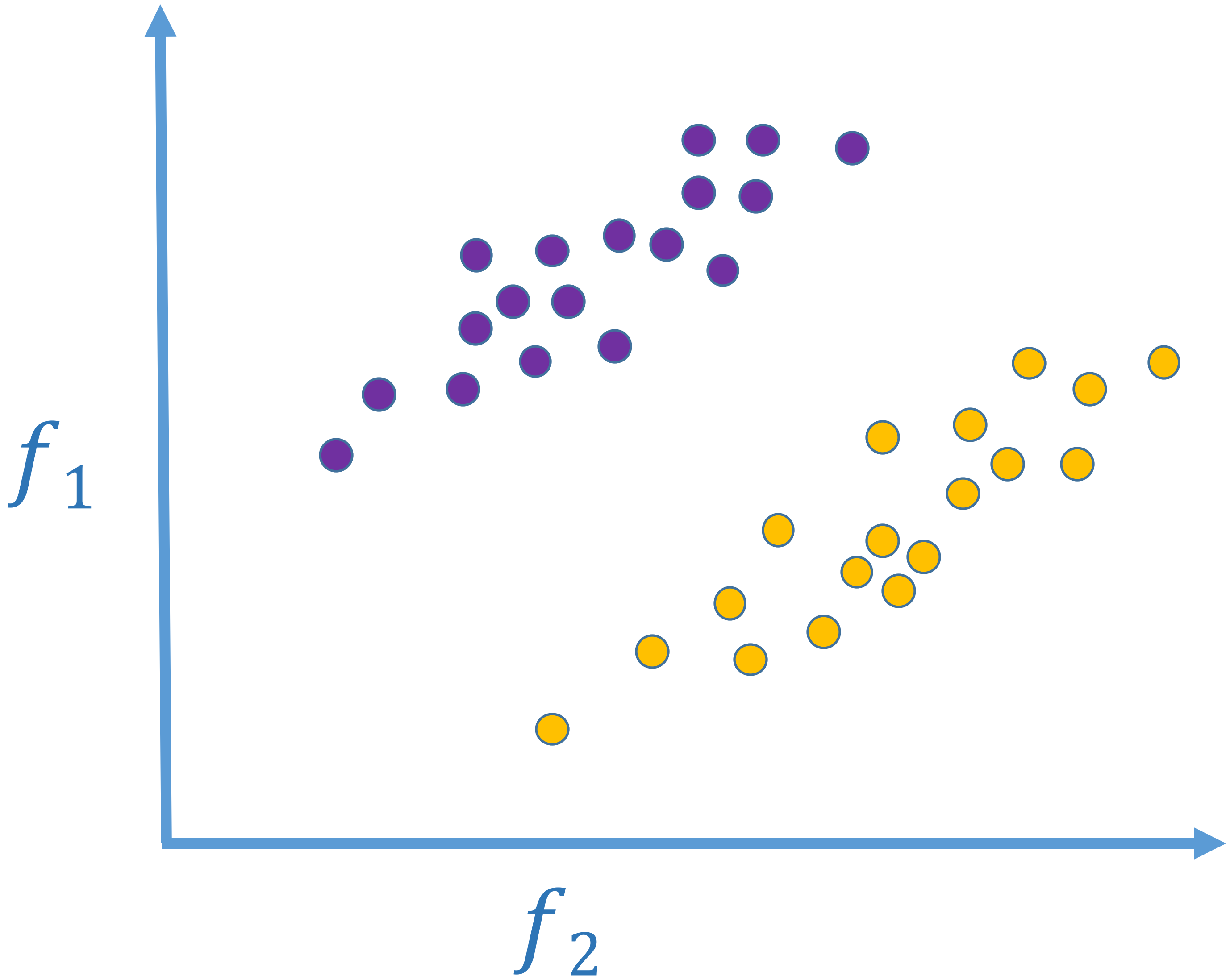
demo

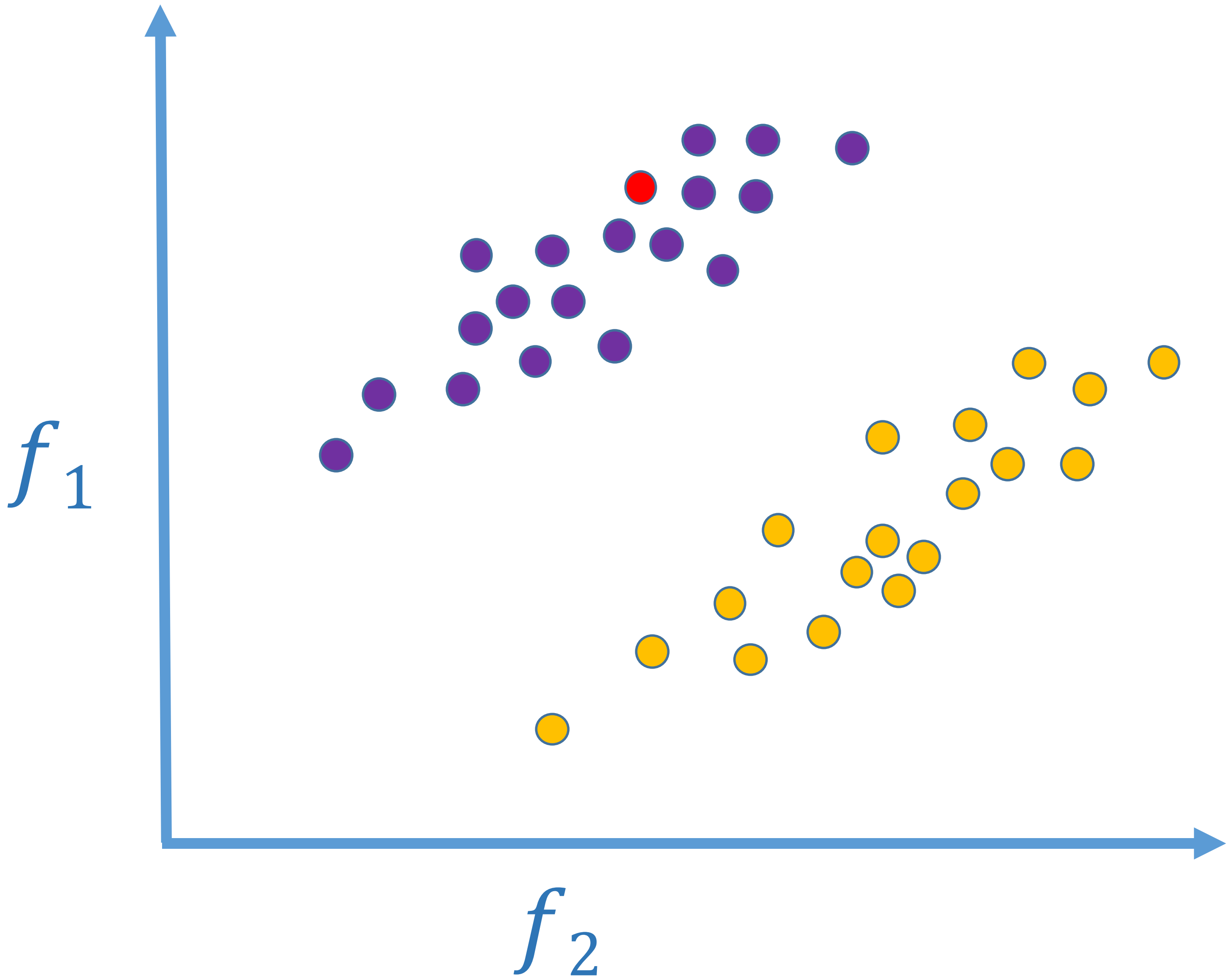
questions?

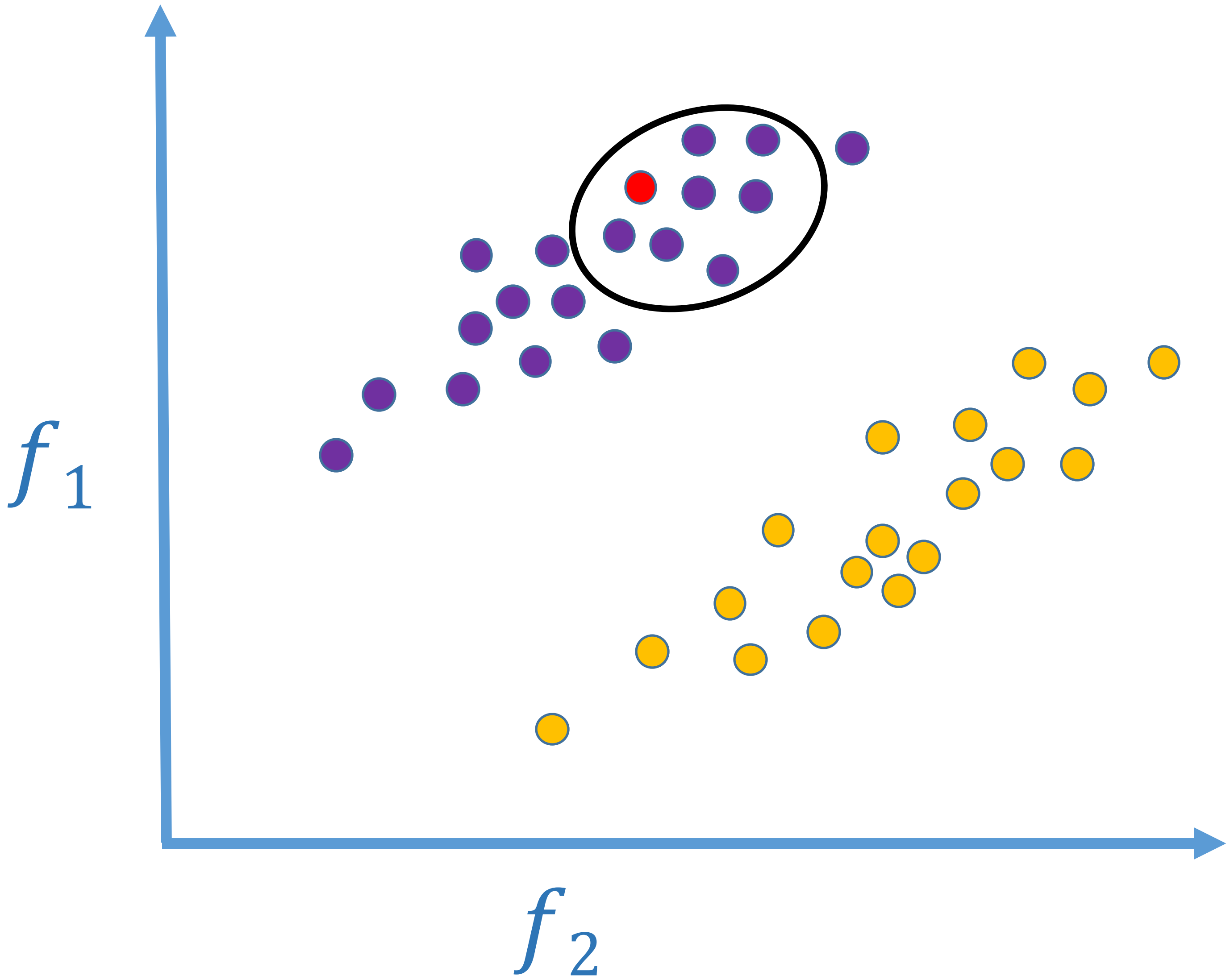
appendix

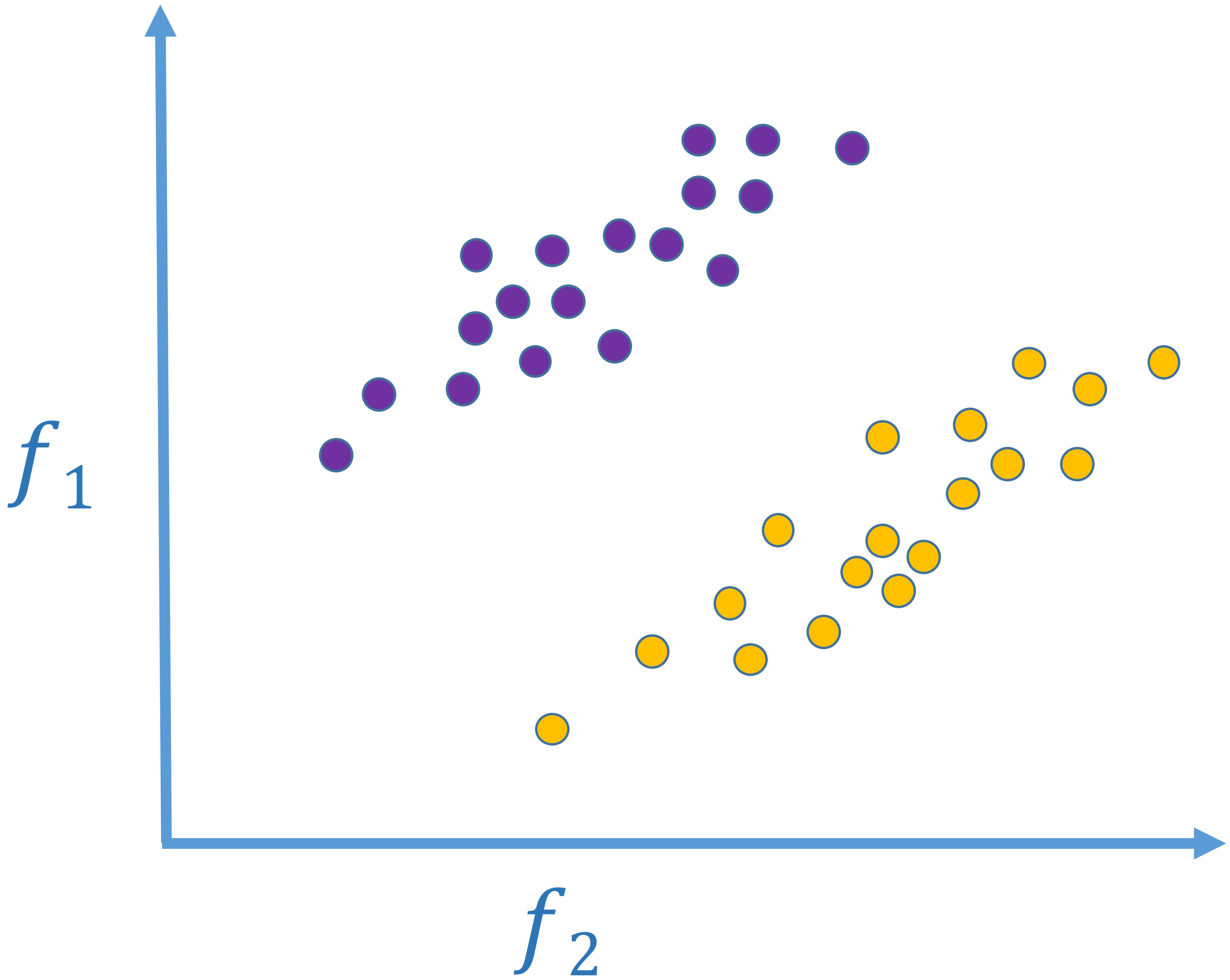
algorithm Classifiers

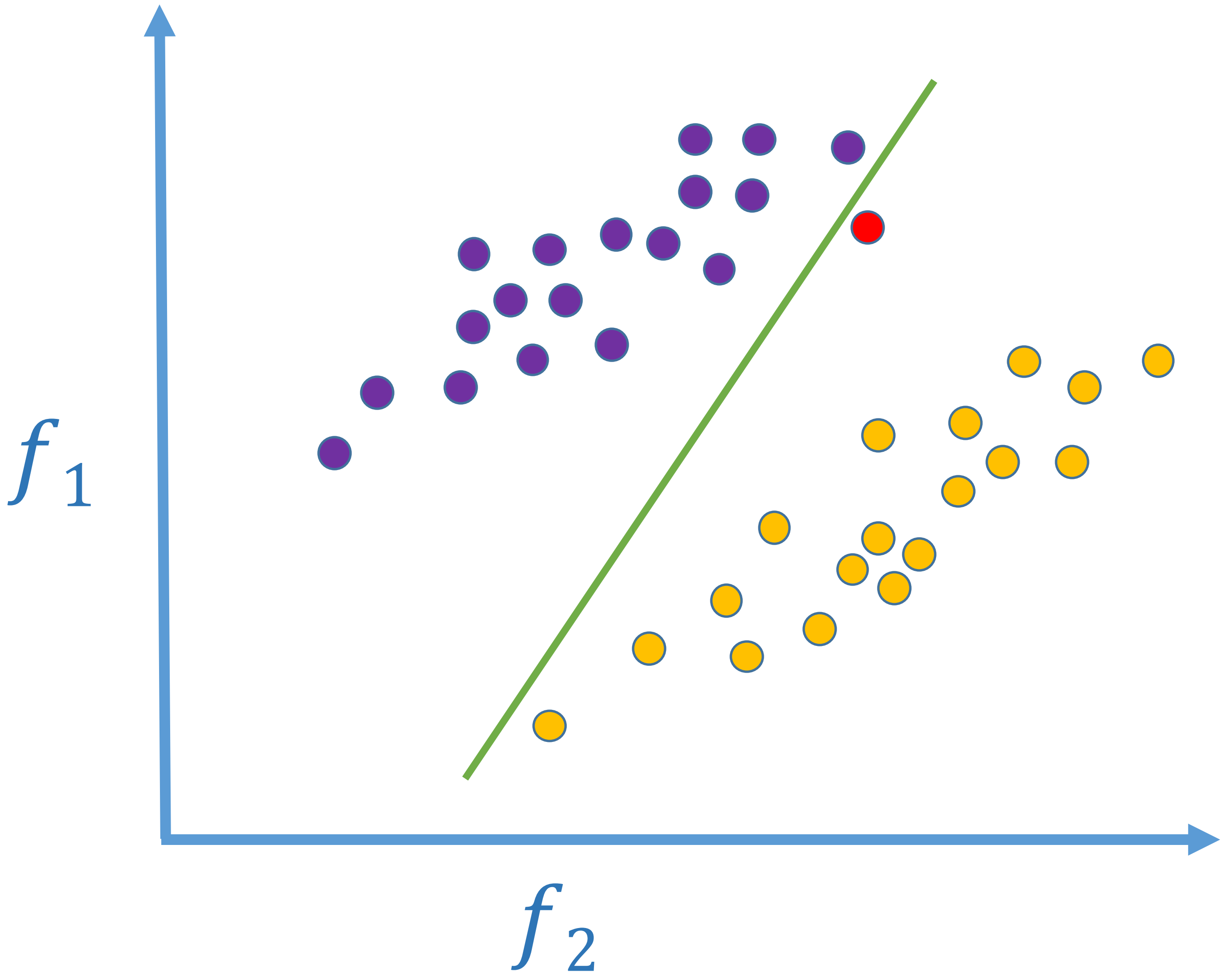
- What set does a new piece of data belong to?
- *k*-nearest neighbor
- Support Vector Machine (SVM)


f_1
f_2

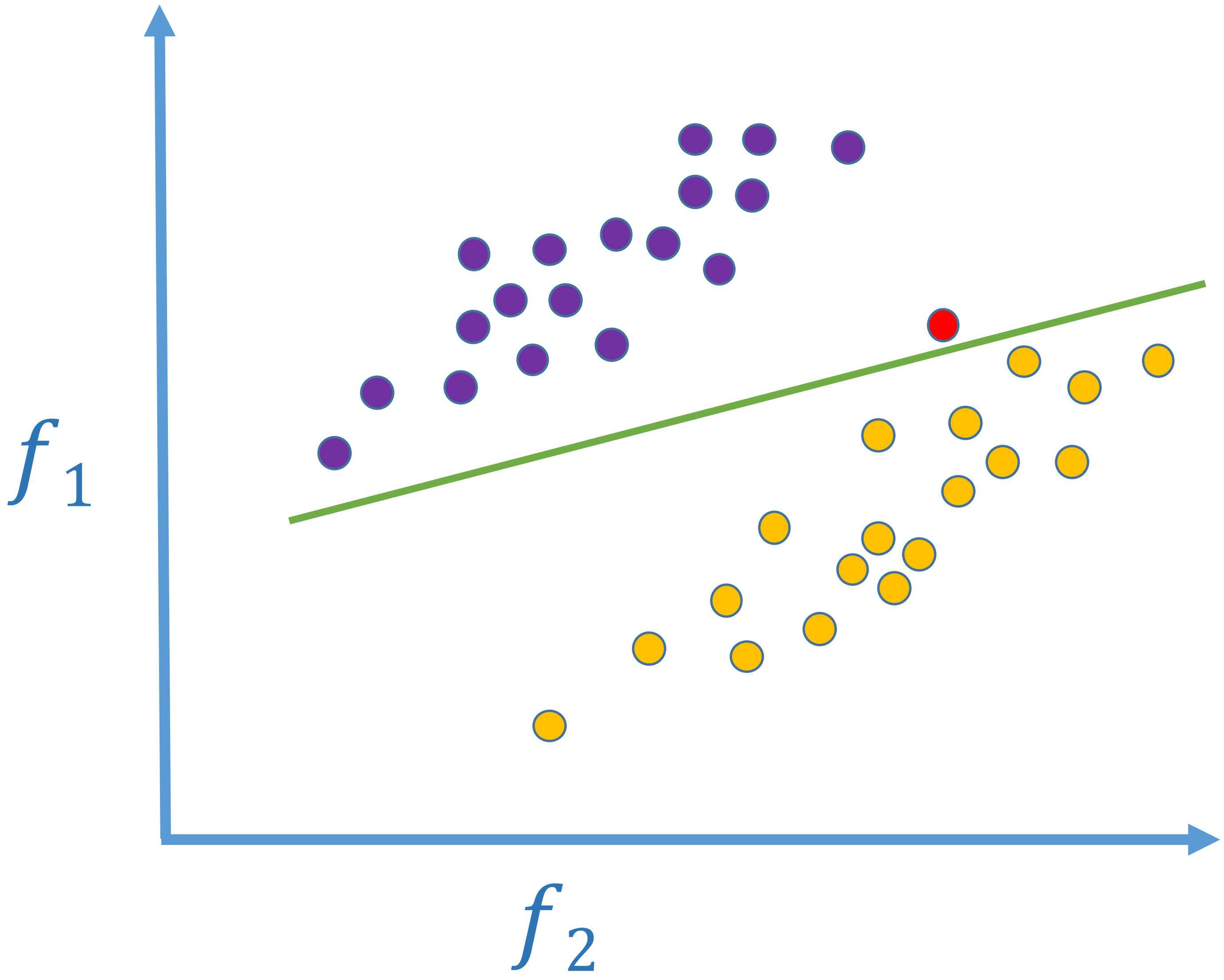

f_1
f_2

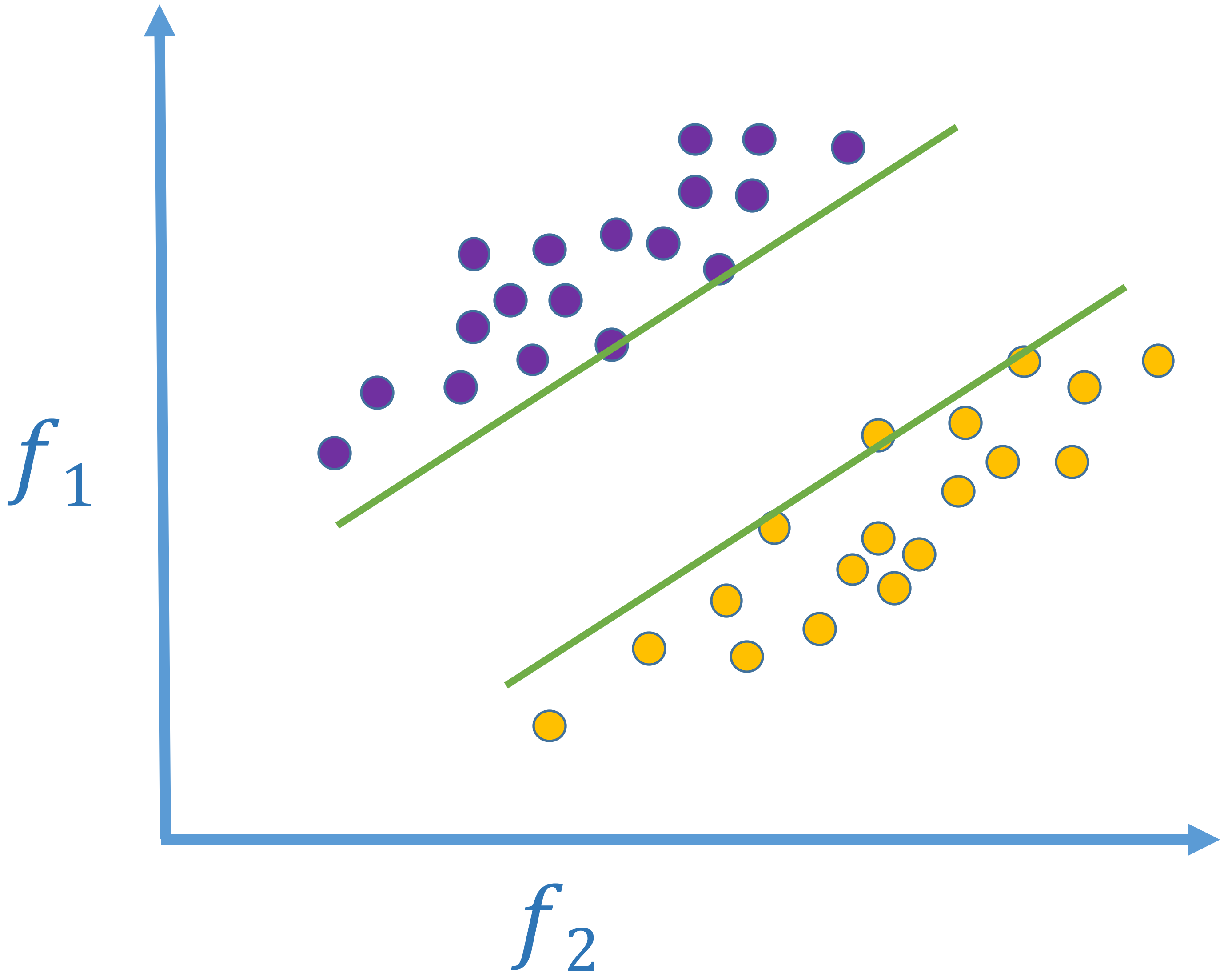

f_1
f_2


f_1
f_2

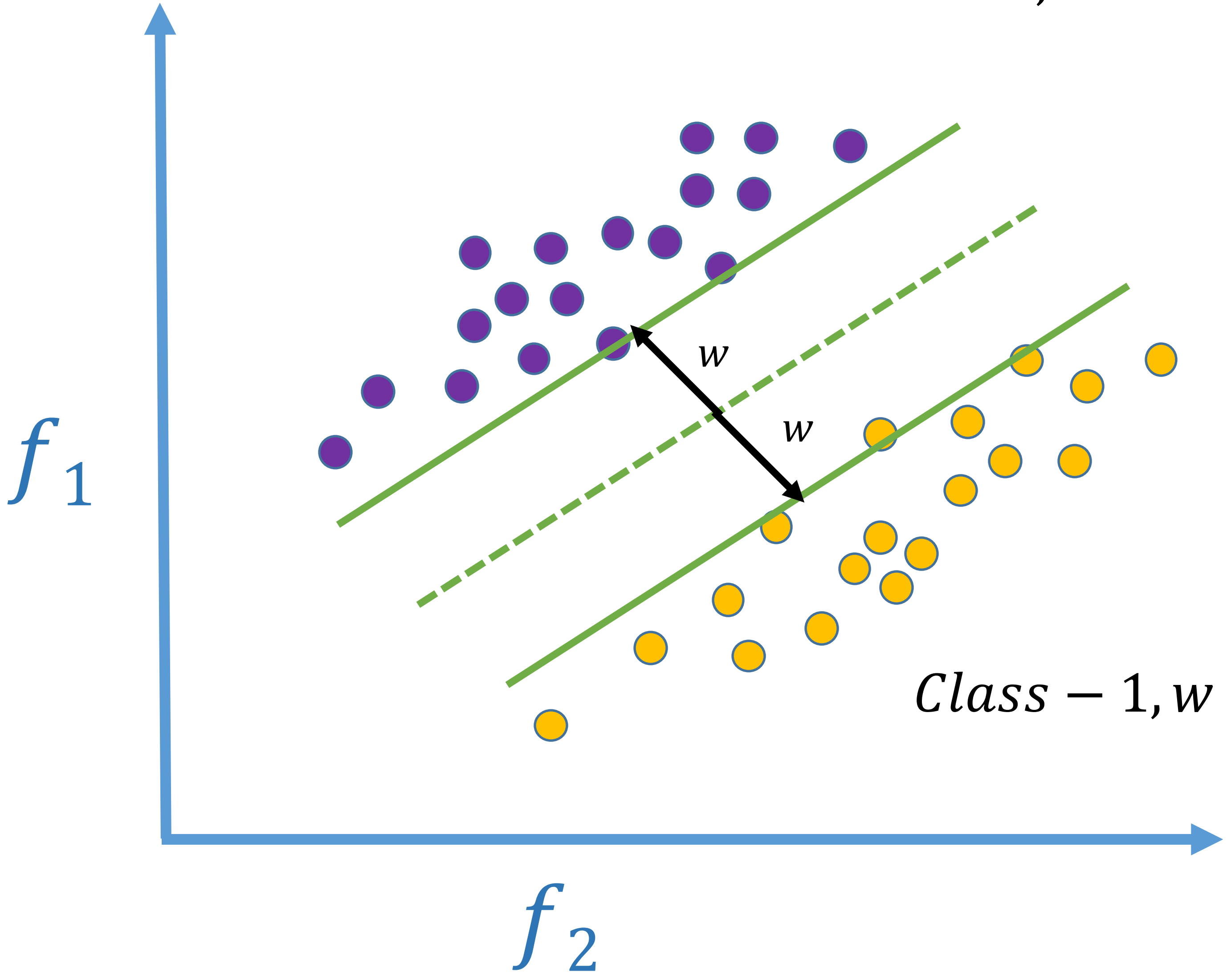


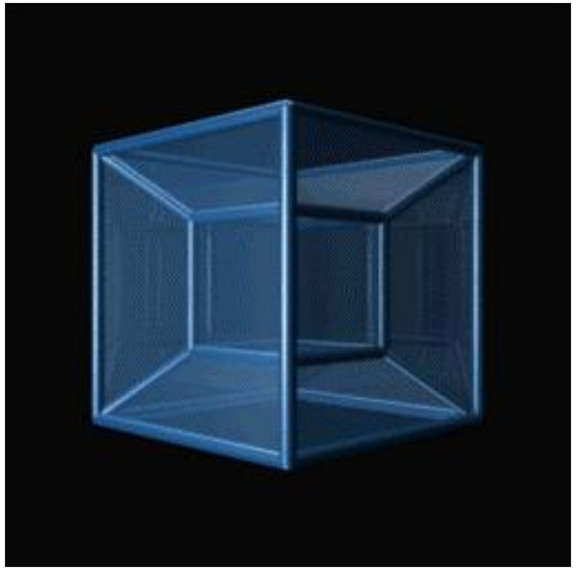
f_1
f_2


f_1
f_2

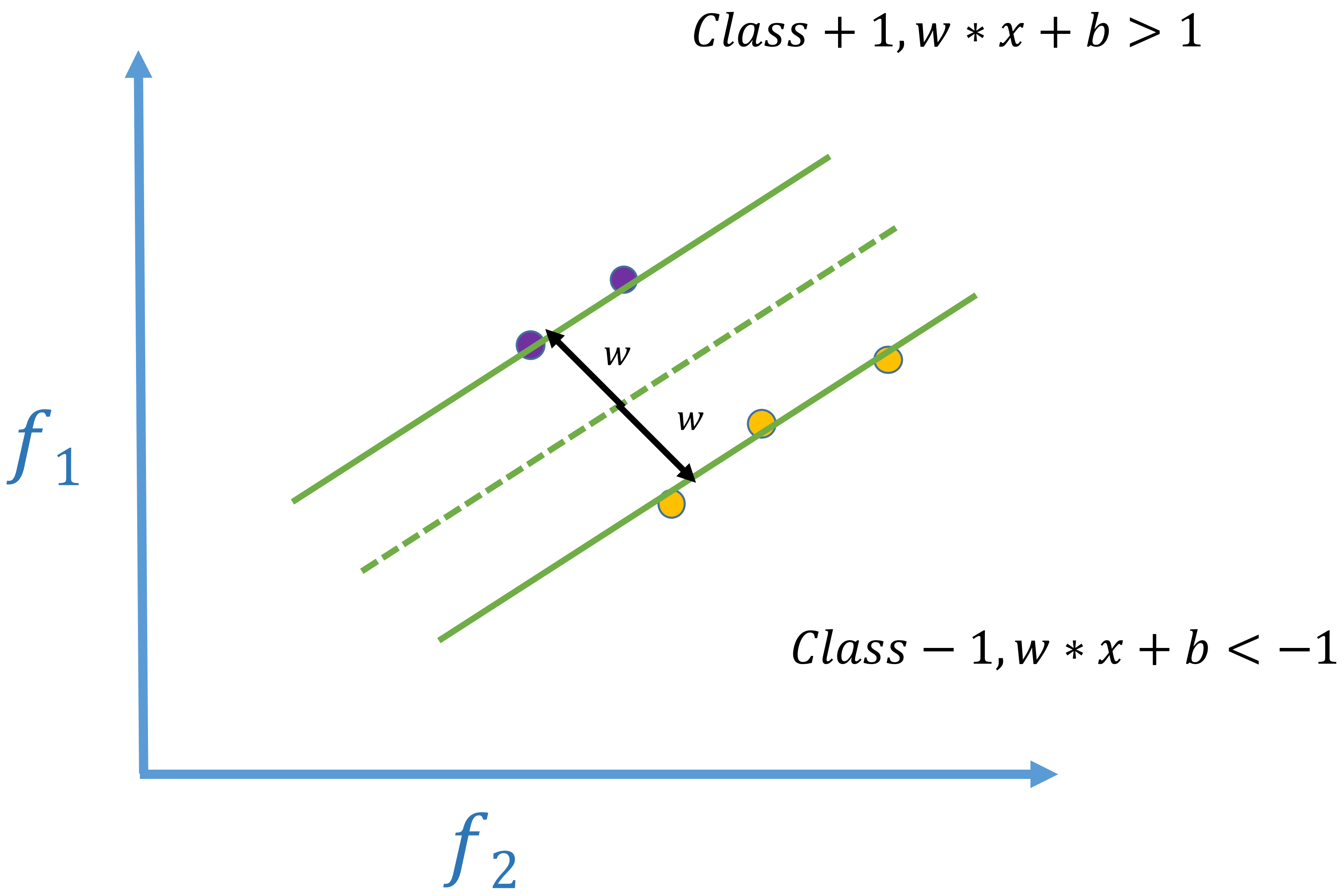

f_1
f_2

$Class + 1, w * x + b > 1$



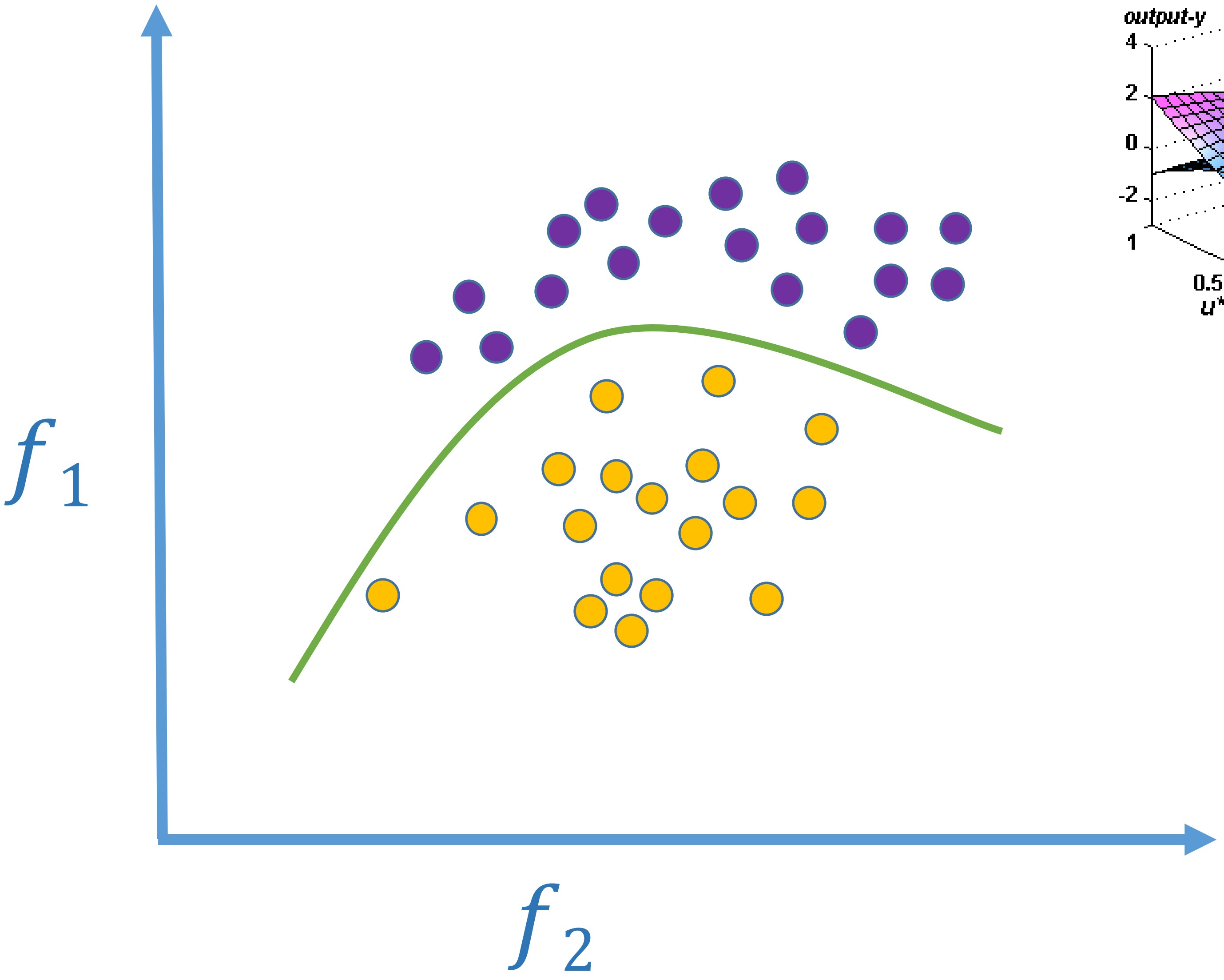


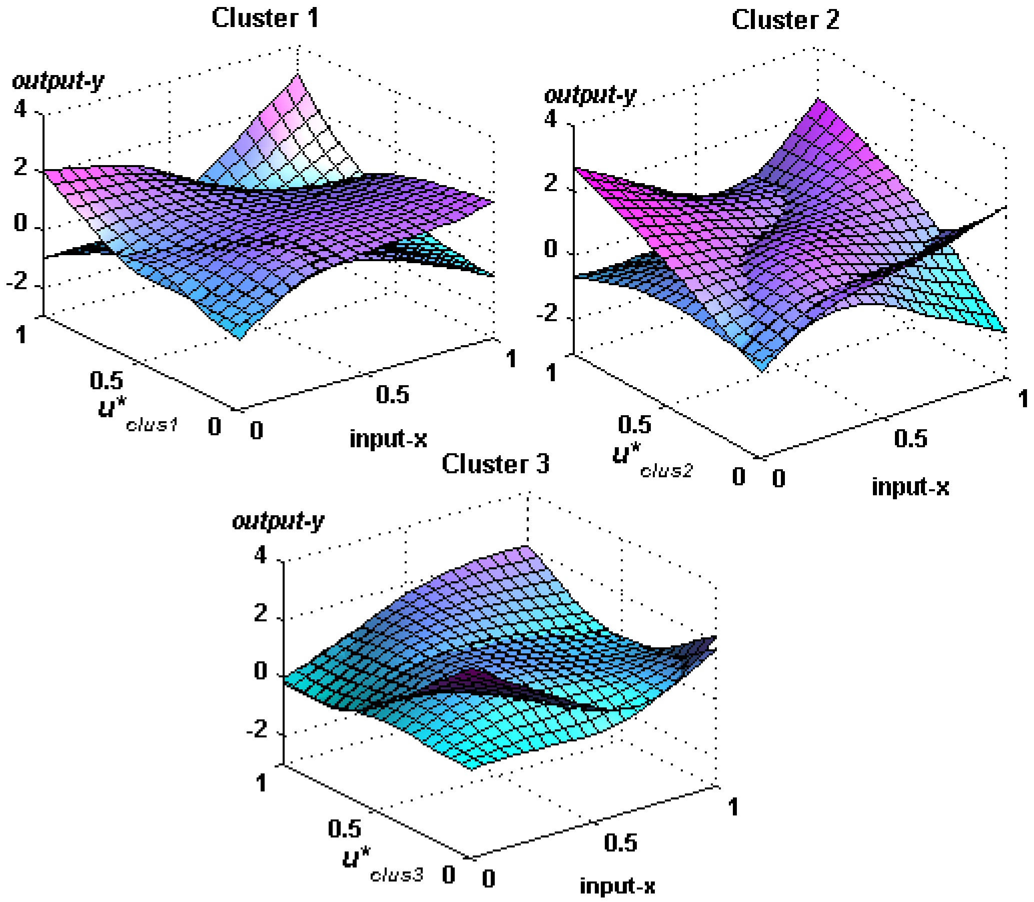
$Class - 1, w * x + b < -1$



$Class + 1, w * x + b > 1$

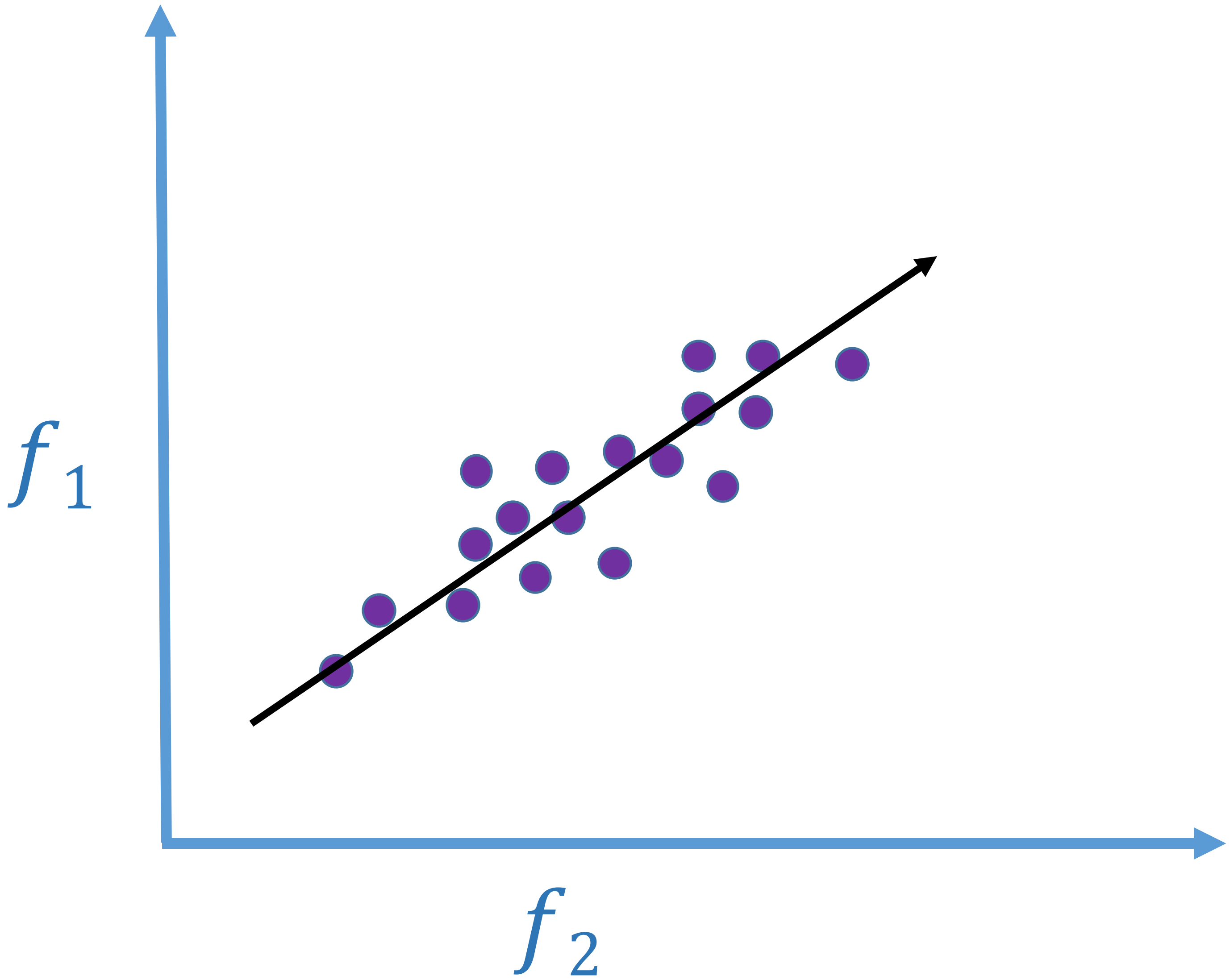
$Class - 1, w * x + b < -1$


f_1
f_2


Cluster 1
output-y
4
2
0
-2
1
0.5
u^*_{clus1}
0
0
0.5
input-x
1
Cluster 2
output-y
4
2
0
-2
1
0.5
u^*_{clus2}
0
0
0.5
input-x
1
Cluster 3
output-y
4
2
0
-2
1
0.5
u^*_{clus3}
0
0
0.5
input-x
1

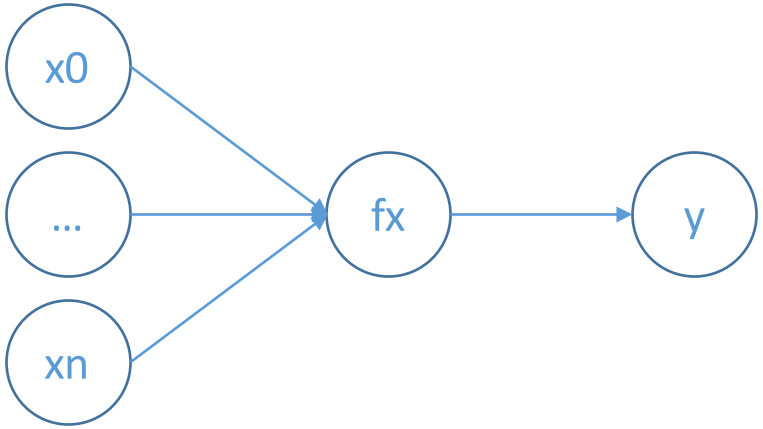
algorithm Regression

- Most commonly best fit or least squares “best fit line”
- Can be non-linear (logarithmic, quadratic, exponential)


f_1
f_2

software development

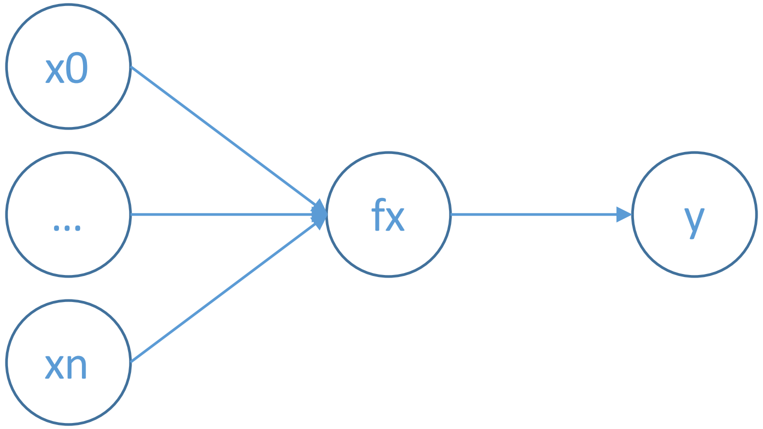
input data function output



known written computed

machine learning

input data function output



known learned known