

CSE 151 Intro. to AI: A Statistical Approach

Instructor: Kamalika Chaudhuri

Course Staff

Instructor: Kamalika Chaudhuri (kamalika@cs)

Office Hours: F1-2pm, CSE 4110

TAs:

Shuang Song

Songbai Yan

Shang Liu

Tutor: Shuang Liu

Website:

<http://cseweb.ucsd.edu/classes/sp18/cse151-a/>

What is Machine Learning?

How to use data to learn to make better predictions

Example I: Recommendation Systems

Suggestions to Watch Instantly



Wallace & Gromit: Loaf and Death

Because you enjoyed:

- The Iron Giant
- Howl's Moving Castle
- Coraline

Play

★★★★★

Not Interested



Iron Man: Armored Adventures

Because you enjoyed:

- The Iron Giant
- How to Train Your Dragon
- Kung Fu Panda

Play

★★★★★

Not Interested



Traffik

Because you enjoyed:

- This Is Spinal Tap
- The Third Man
- Rashomon

Play

★★★★★

Not Interested

What is Machine Learning?

How to use data to learn to make better predictions

Example 2: Spam Detection

UCSD School of Jacobs Engineering

Mail ▾

1–51 of 50

COMPOSE

Dictionary.com Word of the Day - [catechize](#): to question closely. - 5 days ago Web Clip

[Delete all spam messages now](#) (messages that have been in Spam more than 30 days will be automatically deleted)

☐	<input type="star"/>	<input type="trash"/>	a bc	Call for Papers—ICECC2012 - Call for Papers Welcome We are ple	Mar 29
☐	<input type="star"/>	<input type="trash"/>	Webmaster	Helpdesk Important Message.! - Your mailbox have exceeded 3.5 M	Mar 29
☐	<input type="star"/>	<input type="trash"/>	AcademicPub	Discover Content for Summer and Fall Classes - Professors Are Te	Mar 27
☐	<input type="star"/>	<input type="trash"/>	Antibody Design and Disc.	Antibody Design and Discovery Conference - If you can't view this	Mar 27
☐	<input type="star"/>	<input type="trash"/>	Pearson Higher Education	New from Pearson: Creating a Sustainable Organization by Soyka	Mar 26
☐	<input type="star"/>	<input type="trash"/>	Control Engineers Bullet.	Top Tuning Secrets Revealed - Article - In this weeks issue of Cont	Mar 24
☐	<input type="star"/>	<input type="trash"/>	CitiFinancial® Loans	Refinance Loans for investment property - Dear Sir/ Madam, Are y	Mar 24
☐	<input type="star"/>	<input type="trash"/>	Advances in Robotics & A.	Editorial board Member Invitation: Advances in Robotics & Auton	Mar 23
☐	<input type="star"/>	<input type="trash"/>	mcanese	Warning! Your mailbox is almost full. - Your mailbox has exceeded	Mar 23
☐	<input type="star"/>	<input type="trash"/>	Control Engineers Bullet.	Includes: Take a Project Management Class Online - Advance Yo	Mar 22
☐	<input type="star"/>	<input type="trash"/>	Jessica Hall	ATTENTION WEB-MAIL USER - ATTENTION WEB-MAIL USER Yo	Mar 21
☐	<input type="star"/>	<input type="trash"/>	Delores Rosa	It is so easy to stop bronchospasms that it is stupid not to do it!	Mar 21

What is Machine Learning?

How to use data to learn to make better predictions

Example 3: Link Prediction

People You May Know



James G. (Jimi) Shanahan,
Experienced Data Mining and
[+ Connect](#)

×



Joshua Widen, System Software
Engineer-Camera at nVIDIA
[+ Connect](#)

×



Chris Pfeifer, Client Service
Manager at CellFire
[+ Connect](#)

×

[See more »](#)

What is Machine Learning?

How to use data to learn to make better predictions

Algorithm behavior changes based on data

This class: some basic machine learning methods

Two Types of Machine Learning

Supervised Learning

Given examples of data and their labels, predict labels of new (unseen) data

Unsupervised Learning

Given data, build a model or cluster

There are other types, but we won't get to it in this class

Supervised Learning

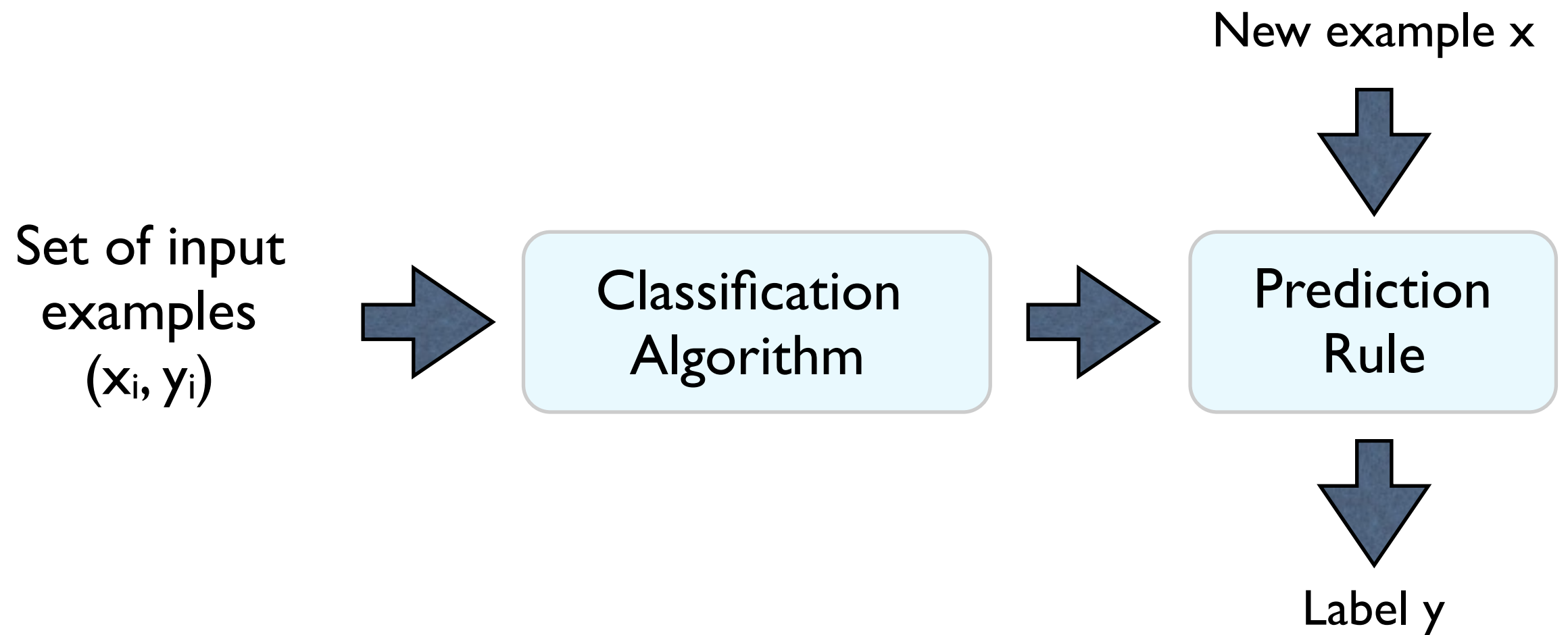
Classification:

Given labeled data:

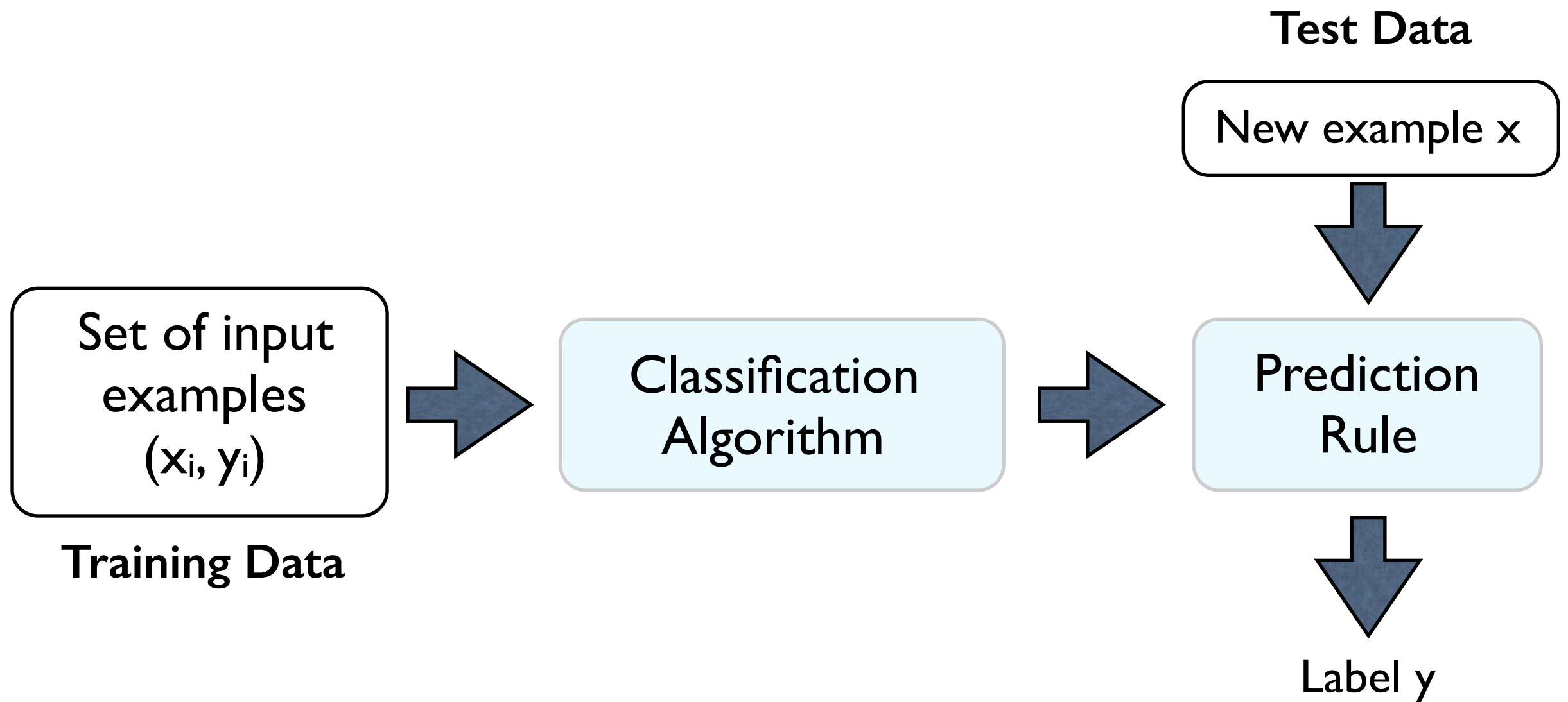
$$\begin{array}{ccc} (x_i, & y_i) & i=1, \dots, n \\ \text{feature} & \text{label} & \\ \text{vector} & & \end{array}$$

where y is **discrete**, find a rule to predict y values for **unseen** x

Typical Classification Algorithm

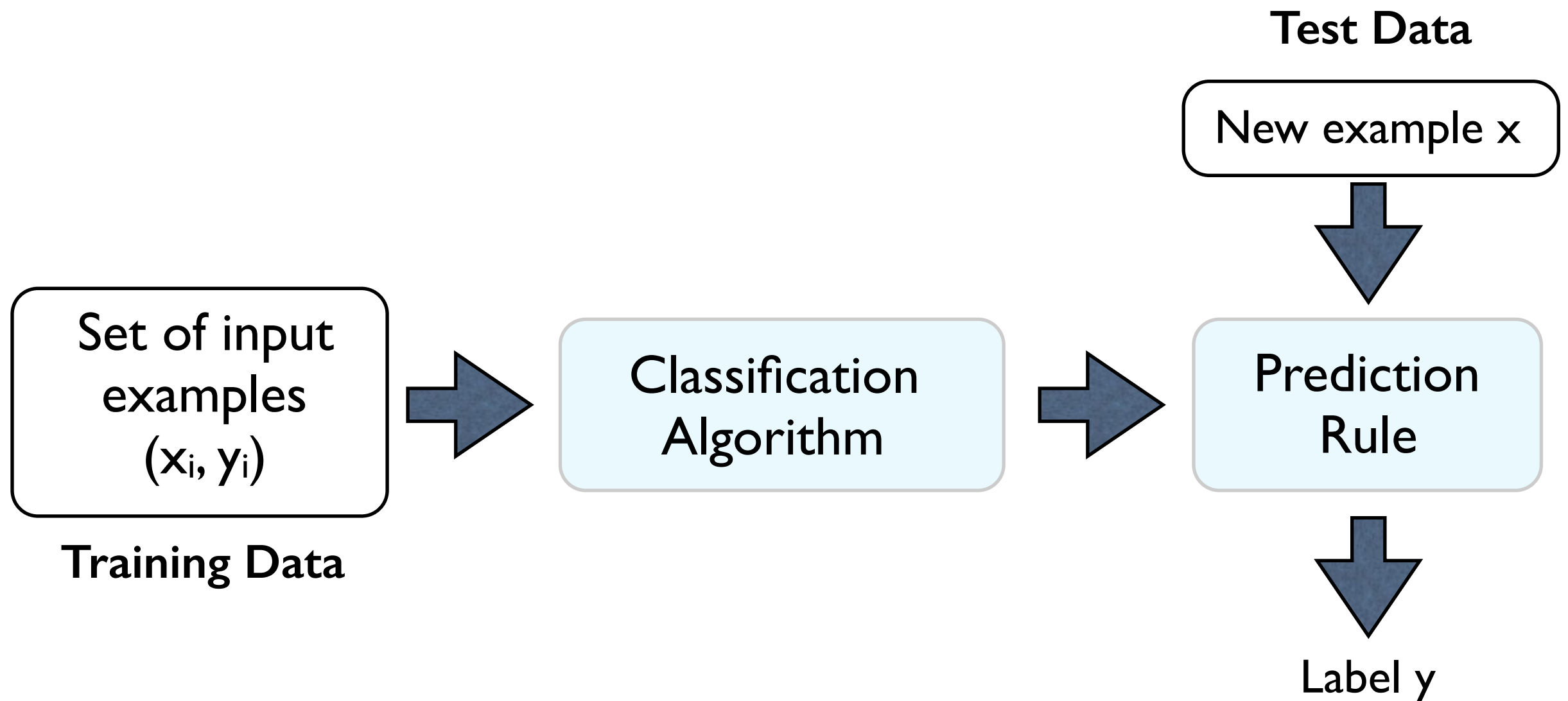


Typical Classification Algorithm



Training and test data must be separate!

Typical Classification Algorithm



Performance Measure:

Accuracy (or fraction of correct answers) on **test data**

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

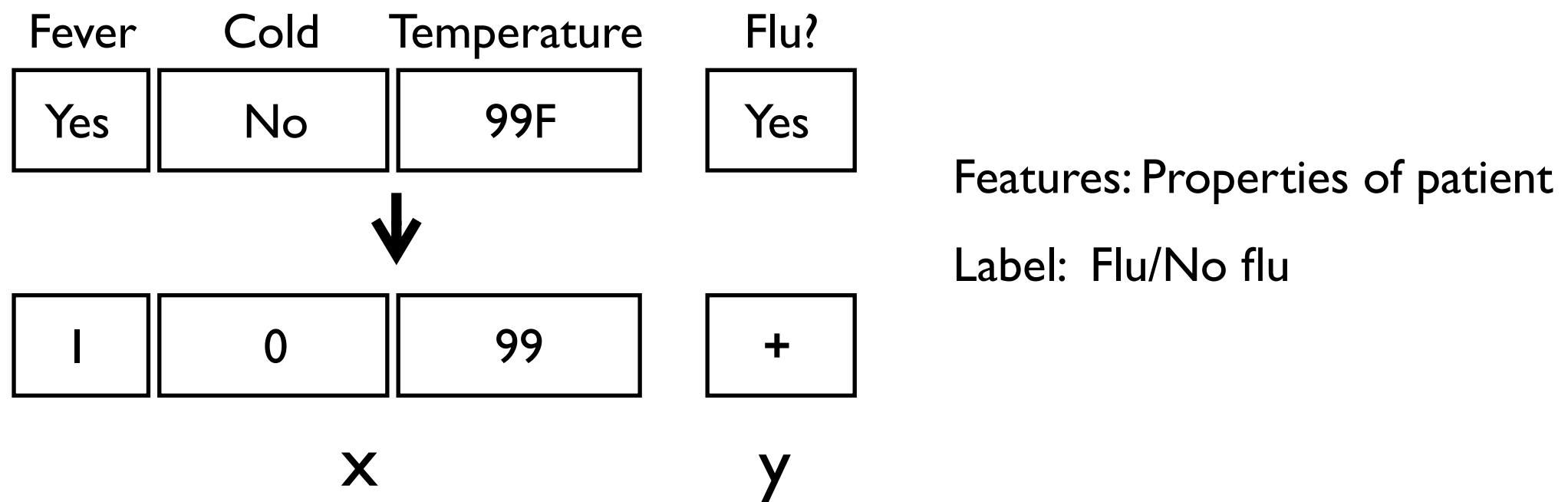
Example 1: Predict if a **new** patient has flu or not,
based on **existing** patient data

What is x and y ?

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

Example 1: Predict if a patient has flu or not

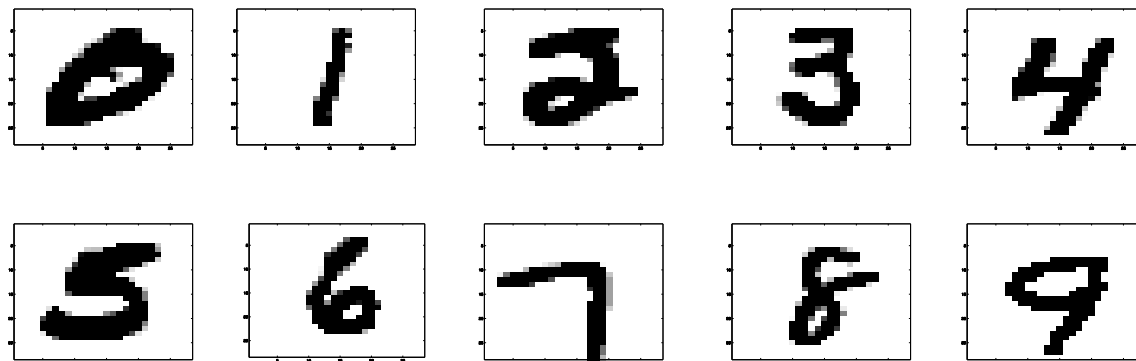


A **binary** (two-label) classification problem

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

Example 2: Which digit in the image ?



Label: 0,1,...,9

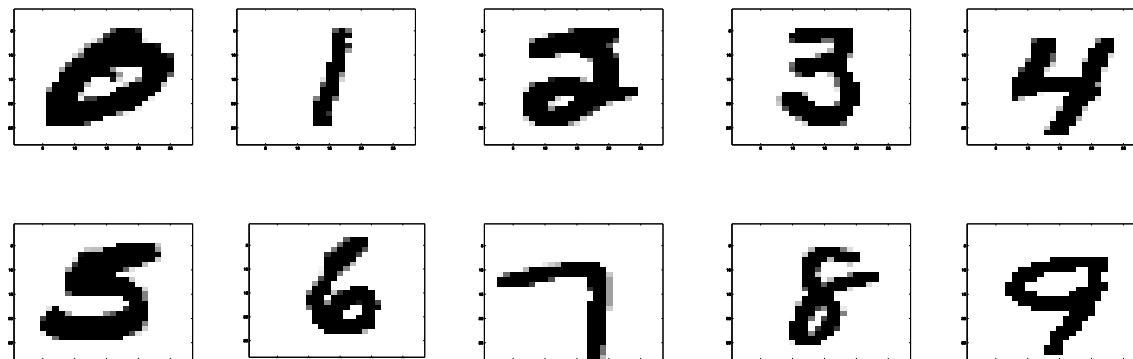
What are the features?

A multiclass classification problem

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

Example 2: Which digit in the image ?



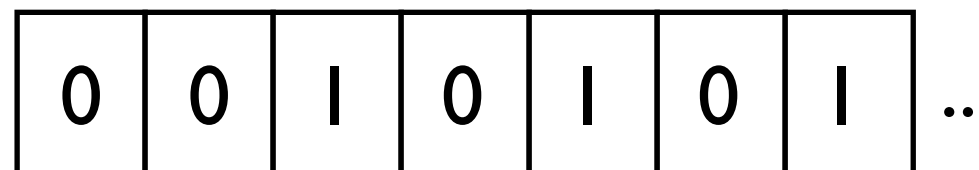
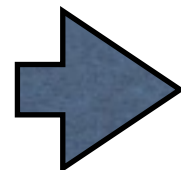
Label: 0,1,...,9

What are the features?

Option: vector of pixel colors



Image

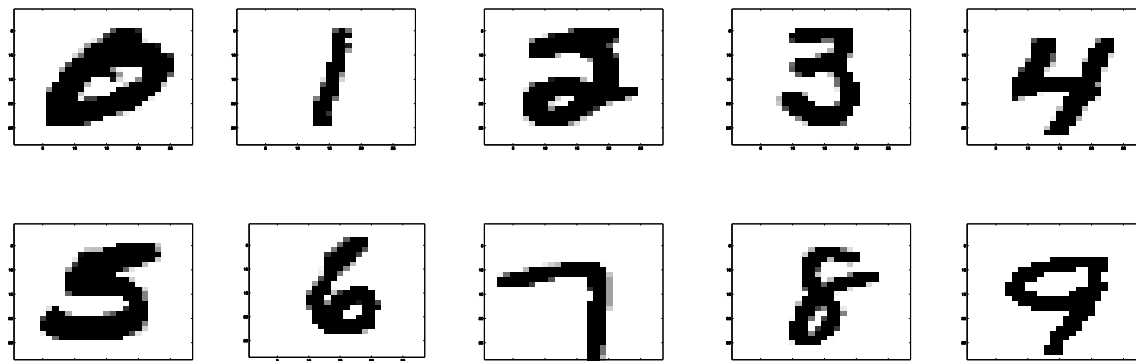


x (0 for white, 1 for black)

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

Example 2: Which digit in the image ?



Label: 0,1,...,9

What are the features?

Option: vector of pixel colors

There are other options too

Lesson: Choosing features is non-trivial in real applications

Supervised Learning

Classification: Given labeled data (x_i, y_i)
where y is **discrete**, predict y values for unseen x

Example 3: Spam or not?

Email 1

From: Canadian Pharmacy
Subject: Offer ends now!

Email 2

From: Yuncong Chen
Subject: TA meeting

	Pharmacy	offer	meeting	TA	Spam?
Email 1	1	1	0	0	Yes
Email 2	0	0	1	1	No

Label: 0 (not spam), 1 (spam)

Features: Words in the email

Supervised Learning

Regression:

Given data:

$$\begin{array}{ccc} (x_i, & y_i) & i=1, \dots, n \\ \text{independent} & \text{dependent} & \\ \text{variable} & \text{variable} & \end{array}$$

where y is **continuous**, design a rule to predict y values for unseen x

Supervised Learning

Regression: Given data (x_i, y_i)
where y is **continuous**, predict y values for unseen x

Example 1: Predict house price from properties of house

Bedrooms	Bathrooms	Area	Price
3	2	2000	600K
2	1	1200	400K
x			y

Independent Variable: Property of house

Dependent variable: price

Two Types of Machine Learning

Supervised Learning

Given examples of data and their labels, predict labels of new (unseen) data

Examples: Classification, Regression

Unsupervised Learning

Given data, build a model

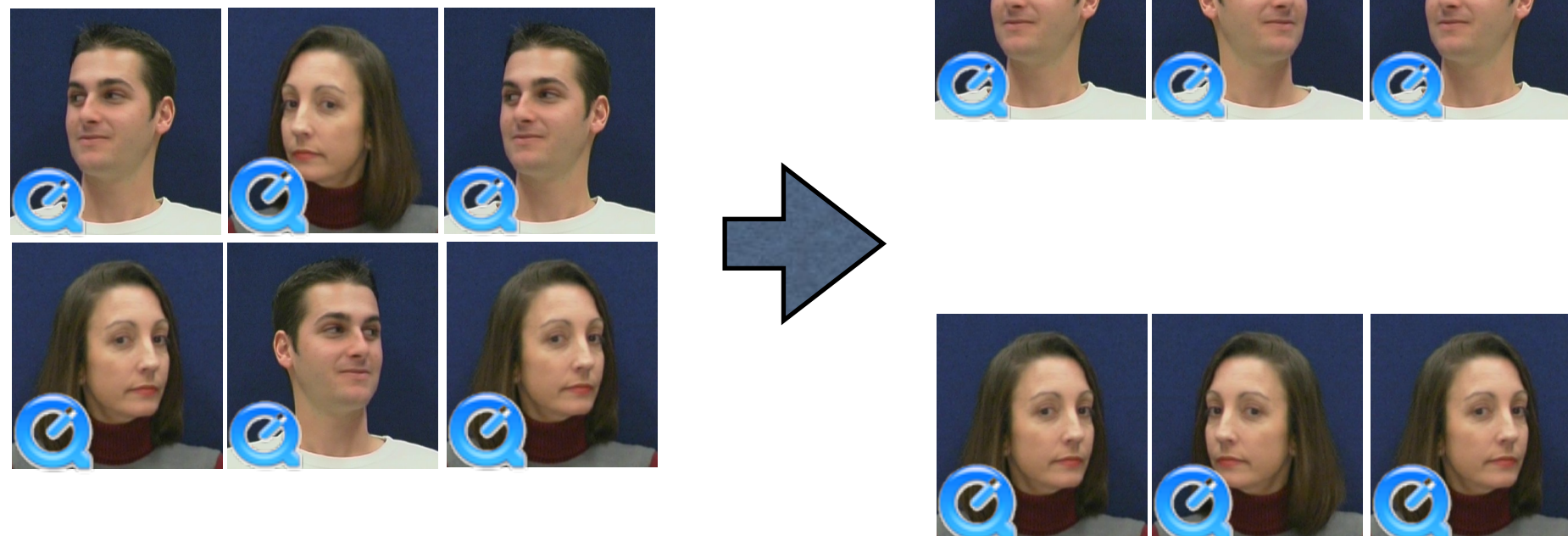
There are other types, but we won't get to it in this class

Unsupervised Learning

Clustering

Given a set of input objects, group them to clusters by similarity

Example 1: Cluster videos by people in them

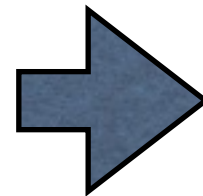
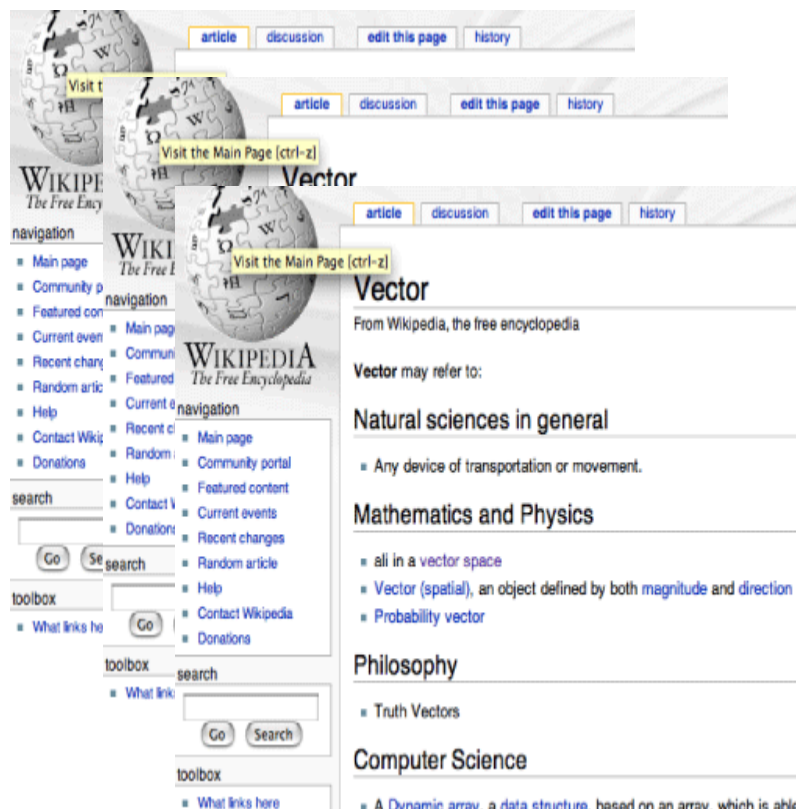


Unsupervised Learning

Clustering

Given a set of input objects, group them to clusters by similarity

Example 2: Cluster documents by topic



Physics

Gravity

Laws of Motion

Electricity

Math

Geometry

Algebra

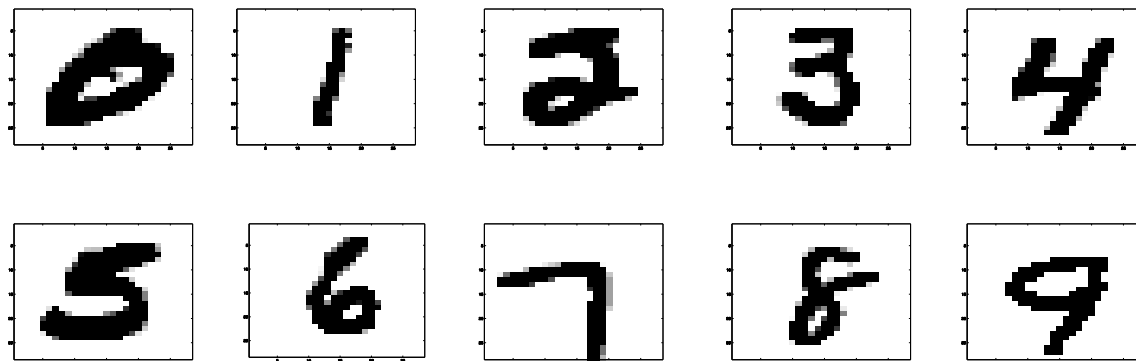
Features: Words in the document

Unsupervised Learning

Dimensionality Reduction

Given high dimensional data, find a good low dimensional representation

Example 1: Images



Number of pixels = 768, so 768-dimensional object
Can we find a lower dimensional representation?

Two Types of Machine Learning

Supervised Learning

Given examples of data and their labels, predict labels of new (unseen) data

Examples: Classification, Regression

Unsupervised Learning

Given data, build a model

Examples: Clustering, Dimension Reduction, learning HMMs

There are other types, but we won't get to it in this class

Logistics

Instructor: Kamalika Chaudhuri

Email: kamalika@cs.ucsd.edu

Lecture: MWF 11-11:50

Sections (optional): M 2-2:50pm, 3-3:50pm

Website: <http://cseweb.ucsd.edu/classes/sp18/cse151-a/>

Administrivia

Textbooks:

No textbook for this class

Syllabus:

Classification -- k-NN, Perceptron, Boosting, etc.

Linear Least Squares Regression

Unsupervised learning -- k-means, hierarchical clustering

Prerequisites

Probability: Events, random variables, expectations, joint, conditional and marginal distributions, independence

Linear Algebra: Vector spaces, subspace, matrix inversion, matrix multiplication, linear independence, rank, determinant, bases, orthonormality, solving systems of linear equations

Calculus: Minima, maxima of functions, derivatives, integrals

Programming: Write programs in a language of your choice.
No hand-holding provided

Prerequisites

Calibration homework HW0 on class website

Calibration quiz Quiz0 on Fri Apr 13

HW0 covers most (but not all) of the material
you need to know as a pre-requisite

Assessment

Homeworks (7): 0% (for self-study only)

Quiz (7): 50%

Programming Assignments (5): 25%

Final: 25%

We will be using GradeSource

Homeworks

There are seven homeworks, one on each topic

Homeworks are for self-study only

Programming Assignments

There are **five** programming assignments

Lowest one will be dropped in calculating your grade

You can use **any language** and **any libraries** for your programming assignments

If you use external libraries, it is **your responsibility** to make sure they give you **correct answers**

Email us a copy of your code with your assignment (instructions later)

Quizzes

There are **seven** quizzes, one for each HW

Each quiz is 20 mins

If you understand how to do a HW,
you should be able to do the corresponding quiz

Lowest one will be dropped in calculating your grade

Quiz Schedule

No makeup quizzes except for medical emergencies

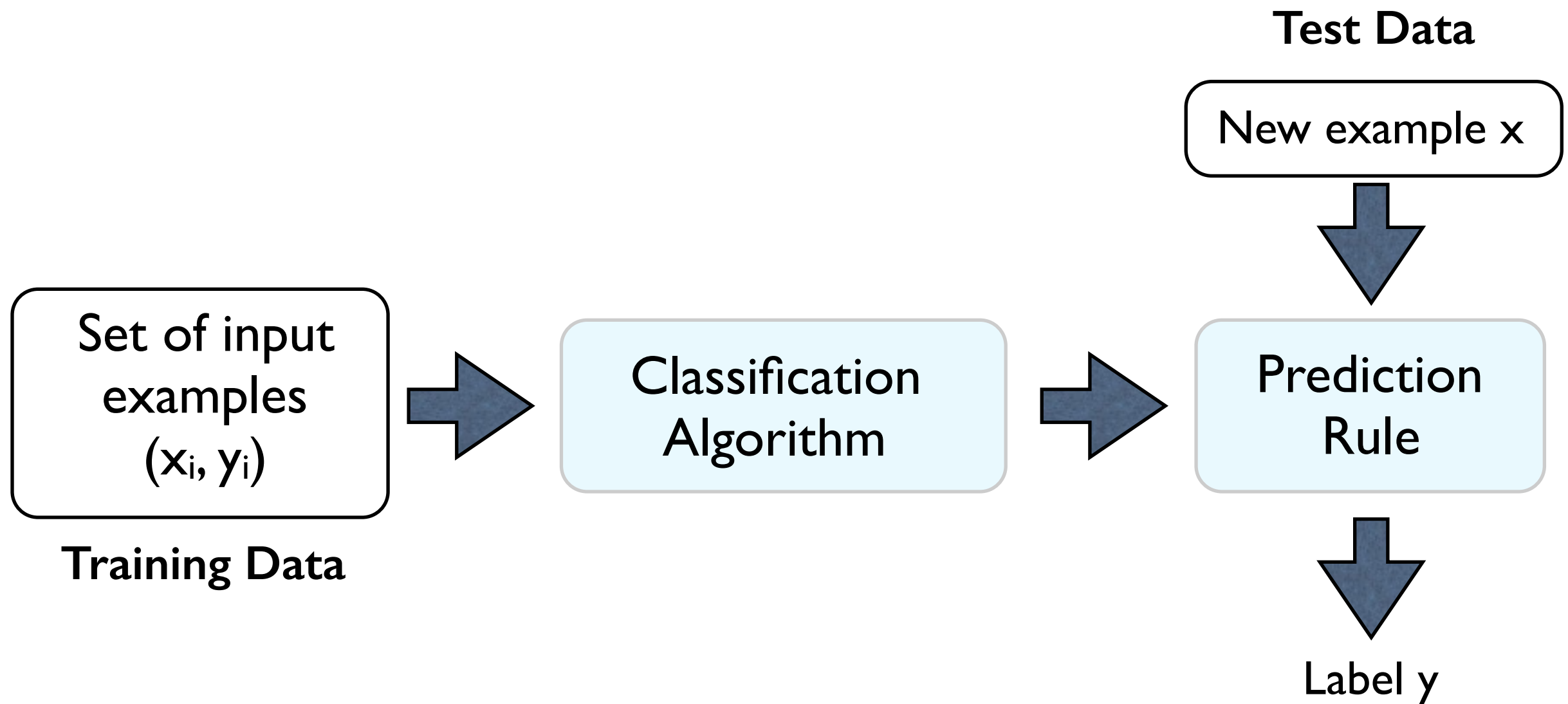
Quiz #	Date
0	4/13
1	4/20
2	4/27
3	5/11
4	5/18
5	6/1
6	6/8

Questions

Message board for this class on Piazza

Please post your questions on the message board!

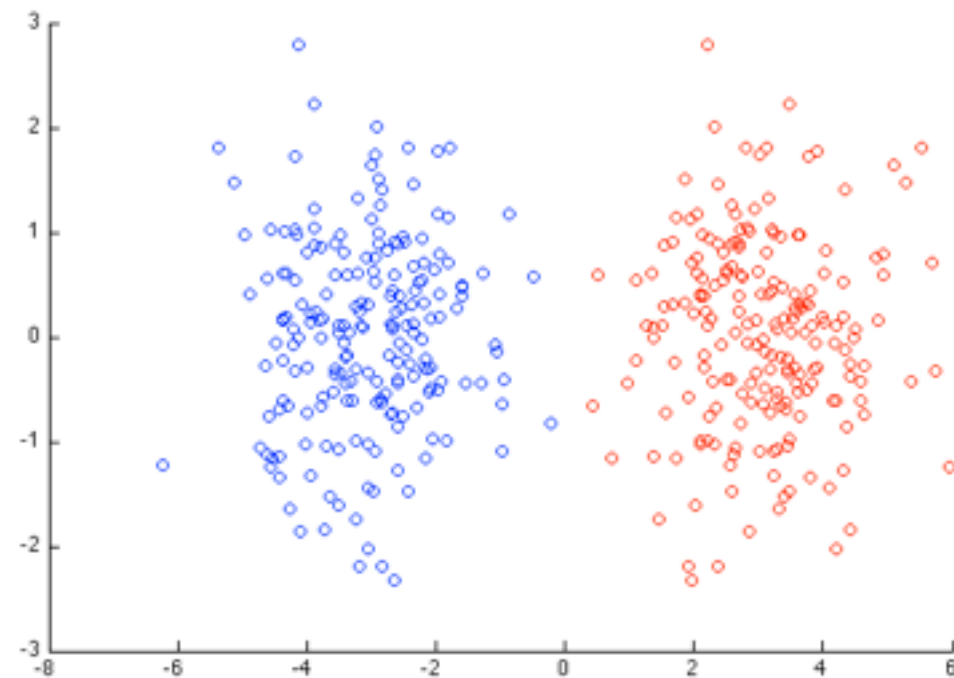
Typical Classification Algorithm



Training and test data must be separate!

Generative Classification

Goal: Classify **red** from **blue**



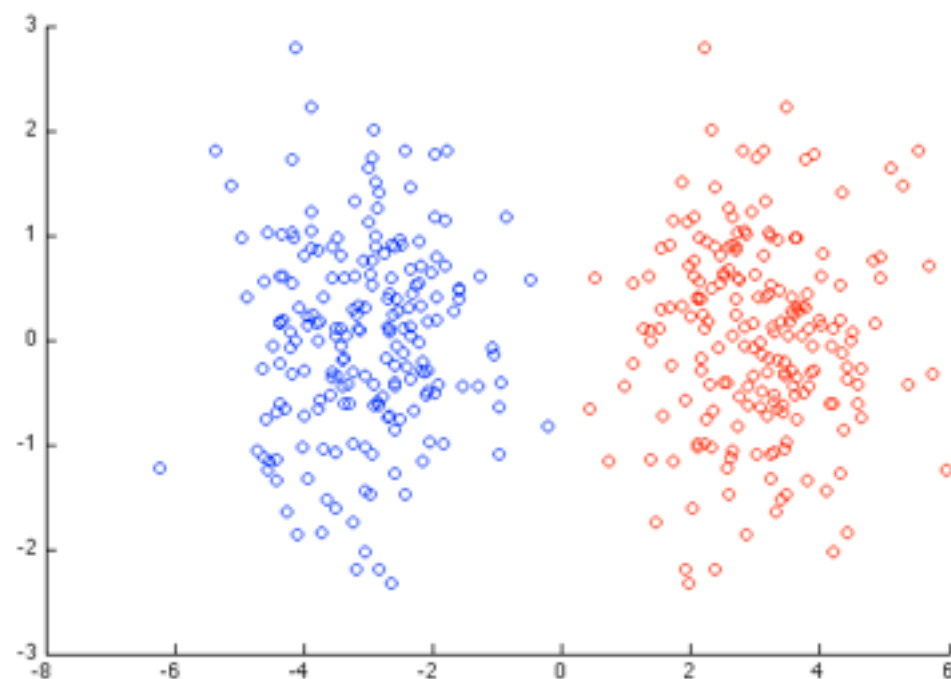
Generative:

- Model each class probabilistically

- Learn the parameters of each class from data

Generative Classification

Goal: Classify **red** from **blue**



Generative:

- Model each class probabilistically

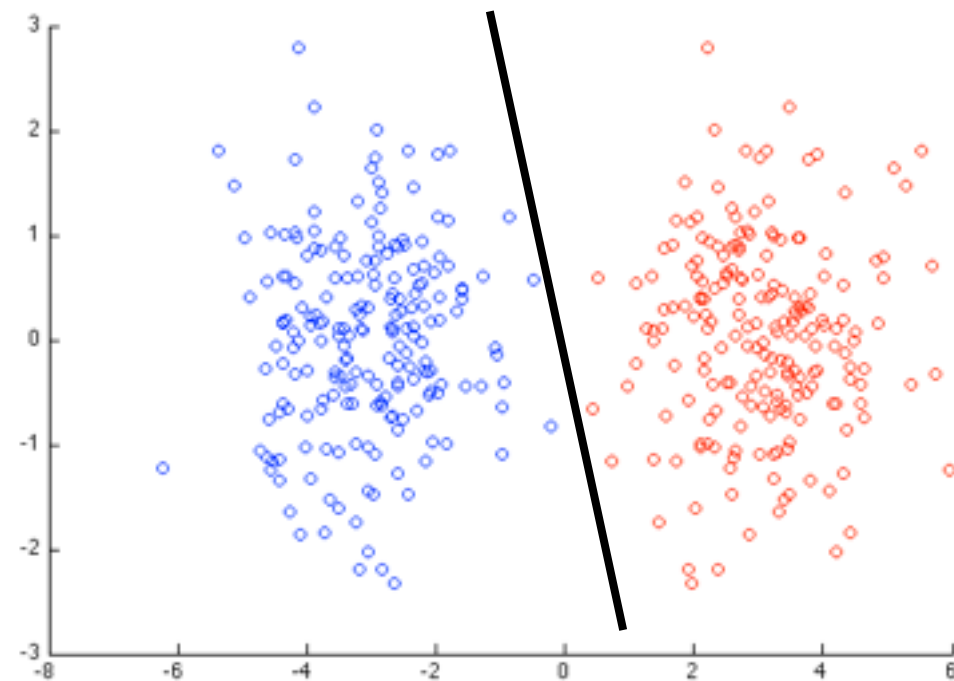
- Learn the parameters of each class from data

For a test example x , find $P(\text{class 1} | x)$ and $P(\text{class 2} | x)$

Report 1 if $P(\text{class 1} | x) > P(\text{class 2} | x)$, 2 otherwise

Discriminative Classification

Goal: Classify **red** from **blue**



Discriminative:

- No need to model each class probabilistically

- Find a suitable separator (say a linear separator) that mostly separates red from blue

Advantages and Disadvantages?

Generative vs. Discriminative

This class we will mostly cover discriminative models