Introduction to pattern recognition

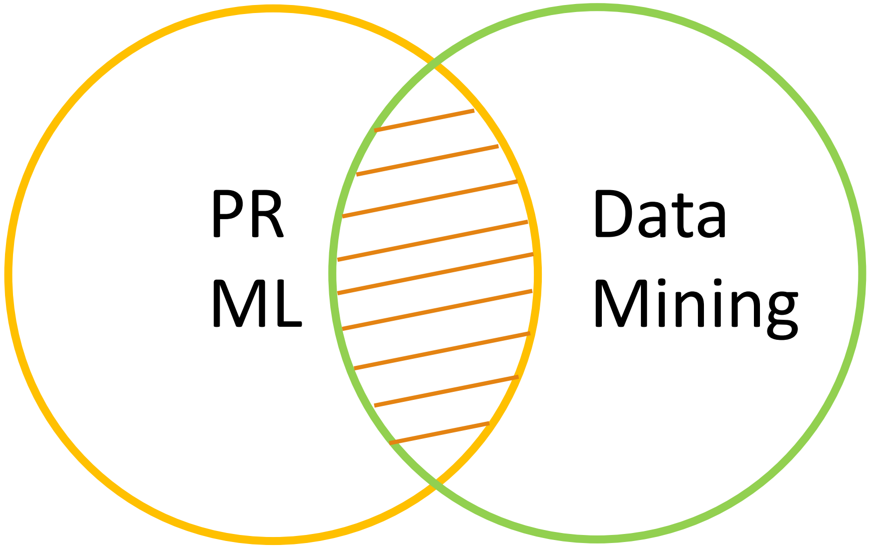
LIN ZHANG

SSE, TONGJI UNIVERSITY

SEP. 2016

Pattern recognition, machine learning, and data mining

Pattern recognition ≈ machine learning



How do you make a decision?

How to pick a “good” watermelon?

How do you know you have a cold?

Can you pick out the apple from bananas?

You are trained from the experience

You learn knowledge to make good decisions

Can a computer learn to make a decision like human?

How?

Experience: data

Knowledge: model

Task:

Learn a model from data

Pattern

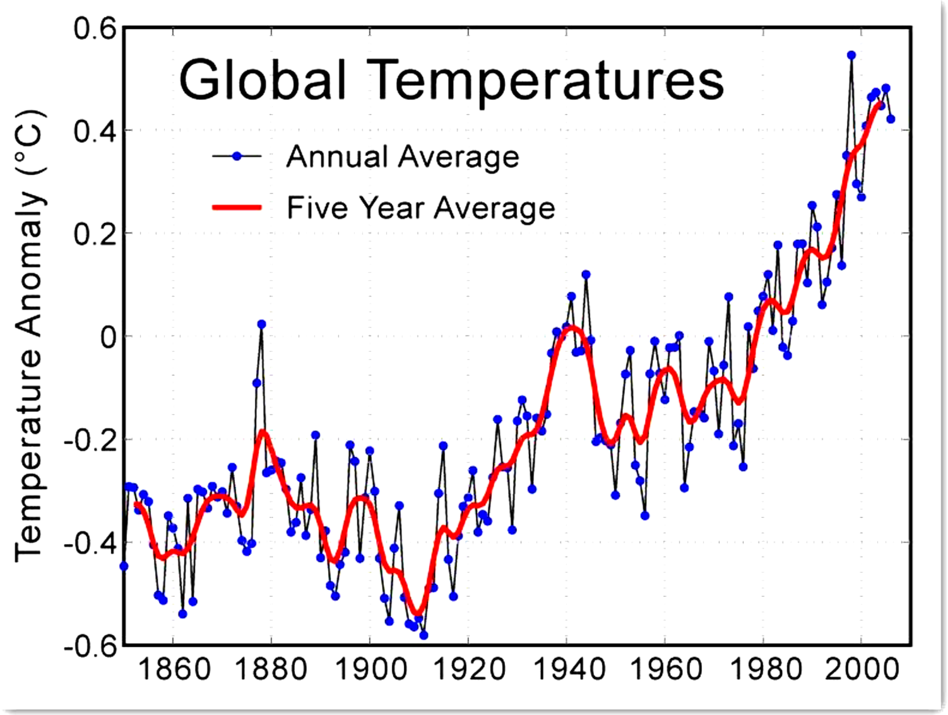
What is machine learning?

One possible definition

a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty

Example: detect patterns

How the temperature has been changing in the last 140 years?



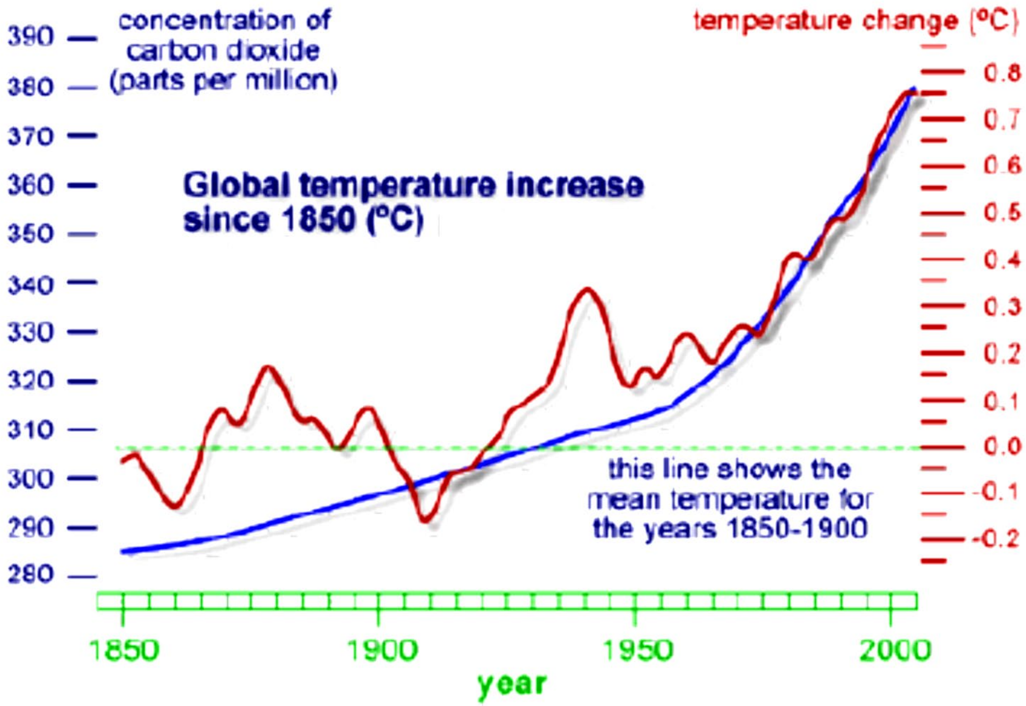
Patterns

- We see repeated periods of fluctuation
- General trend is that temperatures are rising

Repetition frequently and near together —
Recent - learn to short. Law of frequency, recency
and resultant satisfaction - main factor.
4 - a - Instinct to put things into mouth - toys, lu
of sugar - etc. discovered sugar was sweet
so agreeable - formed habit of putting candy
in mouth whenever it could reach it -
b - Instinct of puppy to chew - shoes, trouse

How do we describe the pattern?

Build a model: fit the data with a polynomial function

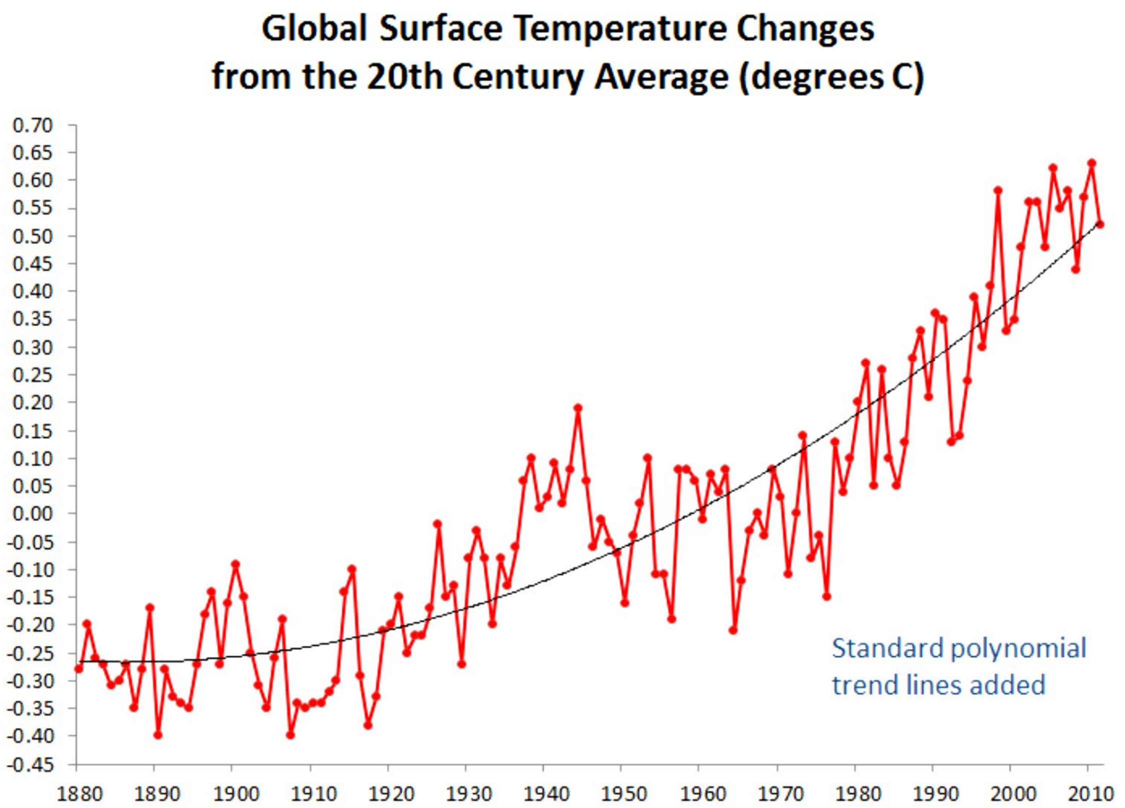


The model is not accurate for individual years

But overall, the model captures the major trend

Predicting future

What is the temperature of 2010?



This particular polynomial model is not exactly accurate for that specific year, but it is pretty close

What we have learned from this example?

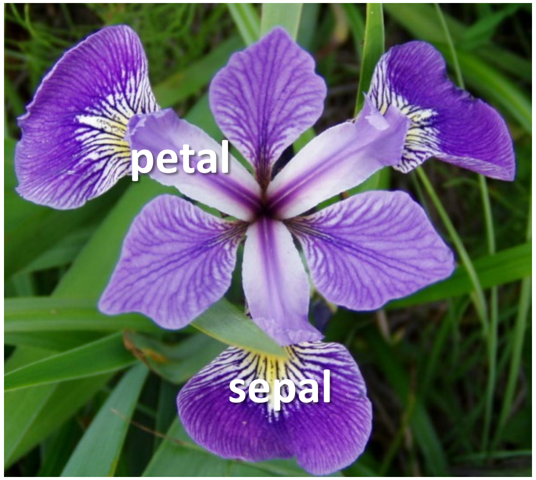
Key ingredients in the machine learning task

- Data: collected from past observations (training data)
- Modeling: devised to capture the patterns in the data
  - The model does not have to be true -- as long as it is close, it is useful
  - We should tolerate randomness and mistakes -- many interesting things are stochastic by nature.
- Prediction: apply the model to forecast what is going to happen in future

A rich history of applying statistical learning methods

Recognizing flowers (by R. Fisher, 1936)

Iris Setosa

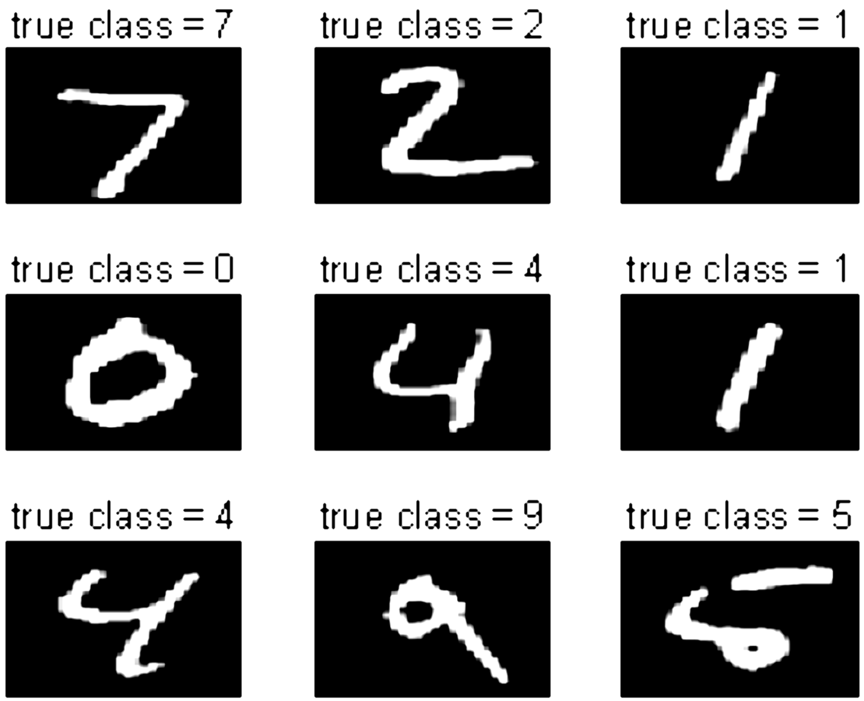


Iris Versicolor

Iris Virginica

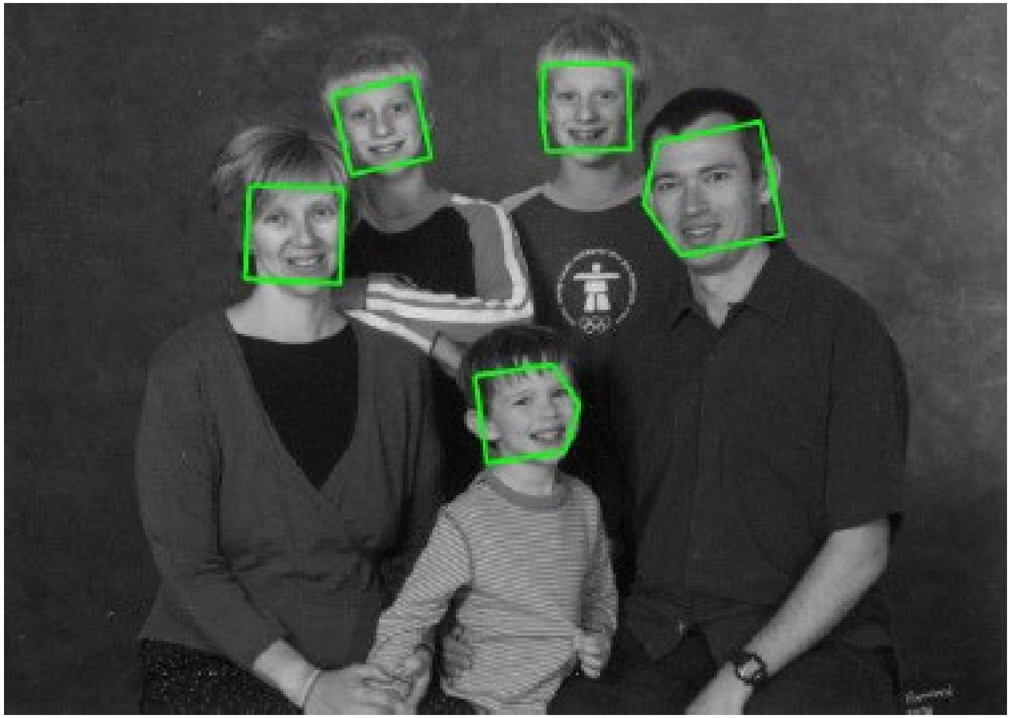
Huge success 20 years ago

Recognizing handwritten zipcodes and checks (AT&T Labs, circalate 1990s)



More modern ones, in your social life

Recognizing your friends on Facebook



Learn your preferences

Recommending what you might like

Why is machine learning so hot?

Flood of data leads to several high-impact applications

Consumer applications:

- speech recognition, information retrieval and search, email and document classification, stock price prediction, object recognition, product recommendation, robot...
- Highly desirable expertise from industry: Google, Facebook, Microsoft, Twitter, LinkedIn, Amazon, BAT, SAIC, Tesla...

Scientific applications:

- Biology and genetics: identify disease-causing genes and gene networks
- Climate science: predicting global warming trends
- Social science: social network analysis; social media analysis
- Business and finance: marketing, operation research
- Emerging ones: healthcare, energy,...

What is in machine learning?

Different flavors of learning problems

- Supervised learning: make prediction given labeled training observations, e.g., Spam detection, Iris
- Unsupervised learning: Discover hidden and latent patterns in data; data exploration, e.g., topic modelling in text data
- Many other paradigms

The focus and goal of this course

- Supervised learning
- Unsupervised learning
- Semi-supervise learning

Let’s start!

Let’s begin to explore the PR world!

Some terms

label

attribute/feature

Iris dataset (from UCI machine learning repository)

attribute value

	sepal length (in cm)	sepal width (in cm)	petal length (in cm)	petal width (in cm)	class
Sample 1	5.1	3.5	1.4	0.2	Iris-setosa
Sample 2	4.9	3.0	1.4	0.2	Iris-setosa
Sample 3	7.0	3.2	4.7	1.4	Iris-versicolor
Sample 4	6.4	3.2	4.5	1.5	Iris-versicolo
…					
Sample 149	6.3	3.3	6.0	2.5	Iris-virginica
Sample 150	5.8	2.7	5.1	1.9	Iris-virginica

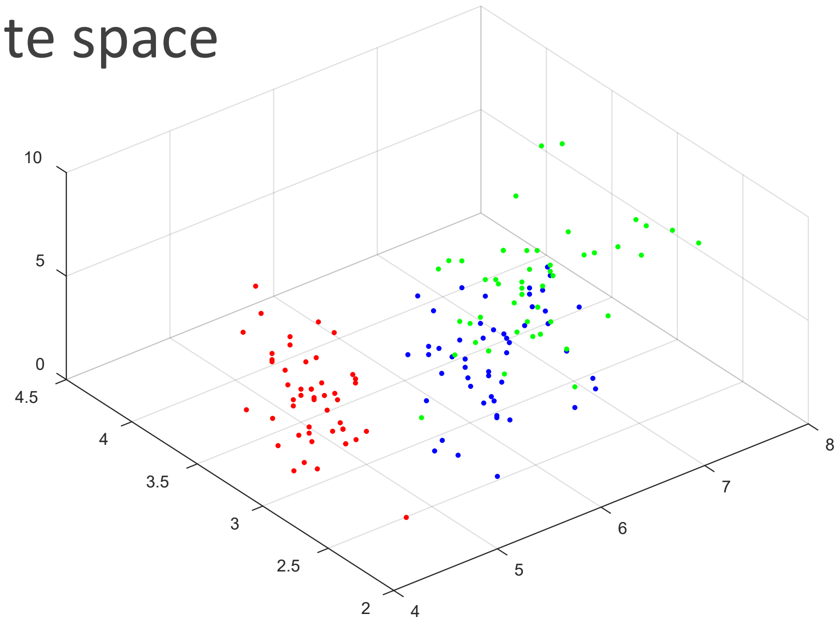
Iris setosa
山鸢尾

Iris versicolor
变色鸢尾

Iris virginica

Some terms

Sample/attribute space



Denote $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$ a dataset which contains m instances. Each instance has d features.

So $\boldsymbol{x}_i = (x_{i1}\ x_{i2}\ \ldots\ x_{id})$ is the i-th instance in the sample space $\mathcal{X}$, x_{ij} is the value of $\boldsymbol{x}_i$ on j-th feature, and d is the dimension of sample $\boldsymbol{x}_i$.

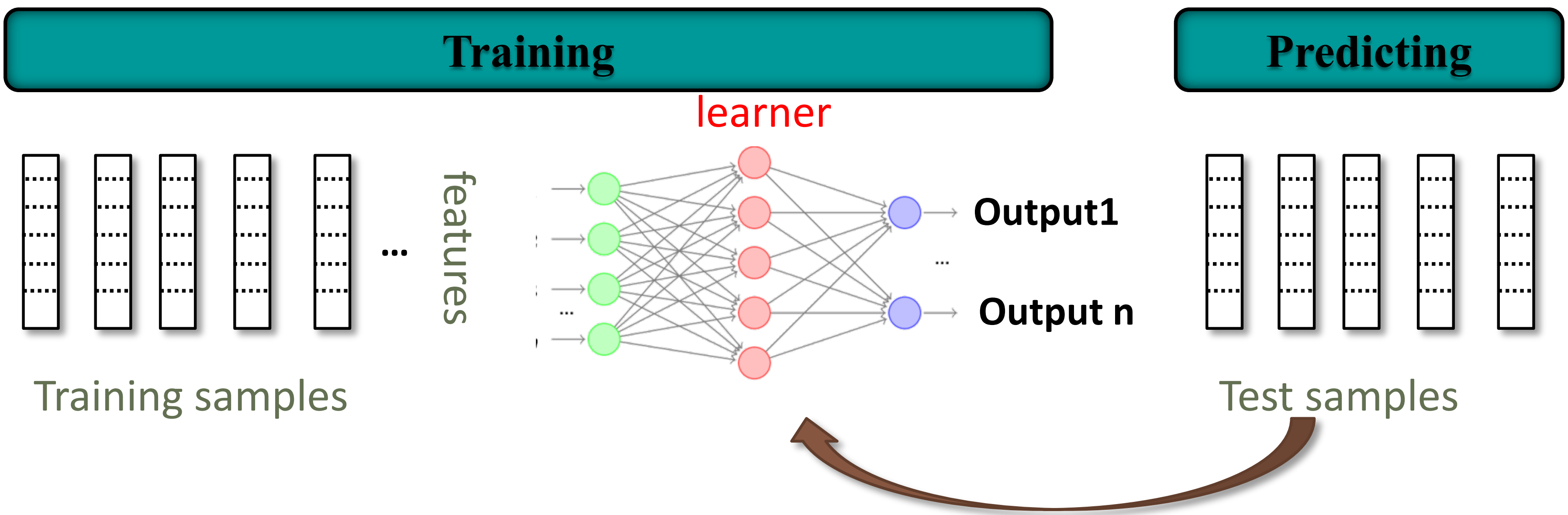
Some terms

The process of learning a model from a dataset is called learning/training process

Data used in the training process is called training data

Each sample in the training data is called training sample

All the training samples consist of a training set



Some terms

There are two types of prediction tasks

- Classification
- Regression

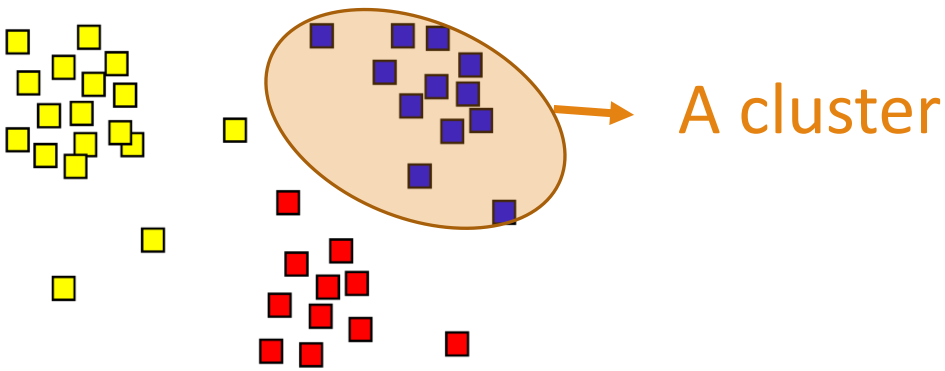
Classification

- Binary classification: a positive class and a negative class
- Multi-class classification
- A labels is used to represent the class that a sample belongs to

Denote y_i the label corresponding to the training sample $\boldsymbol{x}_i$, $y_i \in \mathcal{Y}$. The prediction task is to learn a mapping function f: $\mathcal{X} \longmapsto \mathcal{Y}$

Some terms

We can also do clustering on data if labels are unknown



Learning tasks can be divided into

- Supervised learning (classification + regression)
- Unsupervised learning (clustering)

Some terms

Generalization ability of a model

We assume that

- all the samples in a sample space obey a certain distribution (e.g. Gaussian distribution)
- and training samples are obtained by sampling from the space independently, i.e. training samples are independent and identically distributed (i.i.d)

The more samples are obtained, the more information about the distribution we can have, and the higher generalization ability of a learned model.

hypothesis space

Induction vs deduction

- Induction: special -> general
- Deduction: general -> special

Inductive learning

Hypothesis is a model or pattern learned from training data

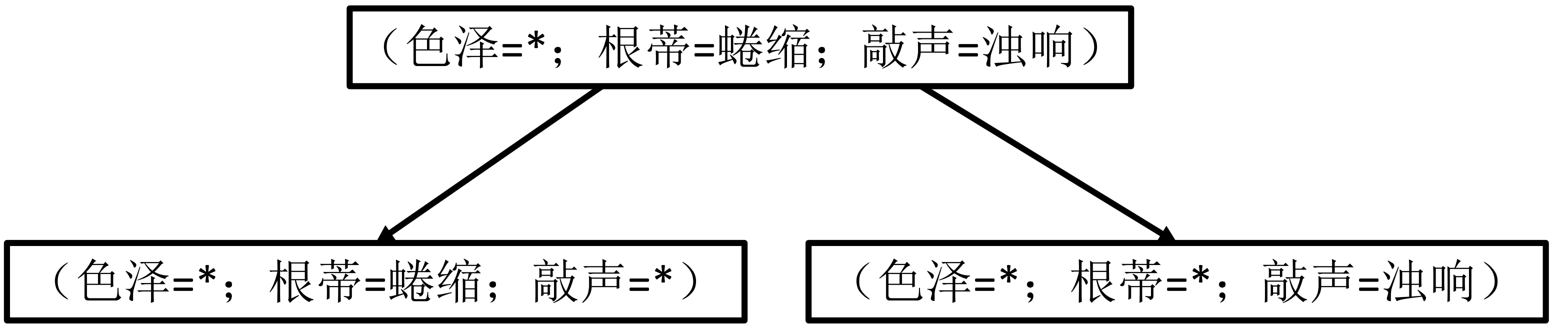
编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

hypothesis space

The hypothesis space is much larger than the (training) sample space

There may exist more than one hypothesis corresponding to the same training set

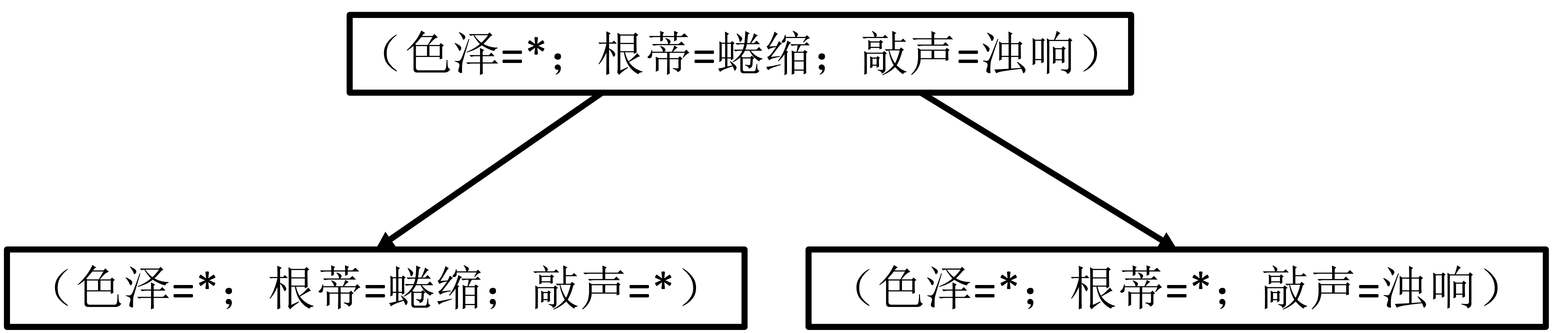
These hypothesis forms a hypothesis set called version space



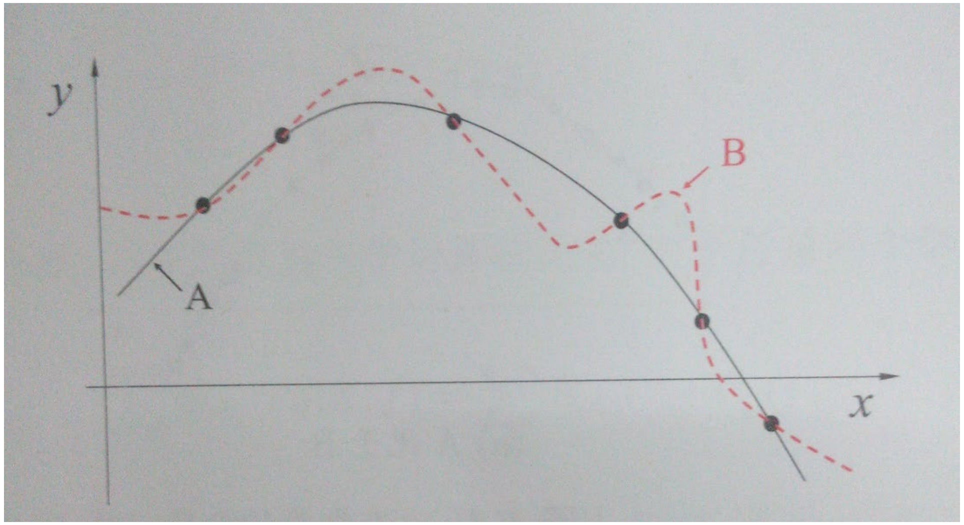
Inductive bias

Given a new sample: （色泽=青绿；根蒂=蜷缩；敲声=沉闷）

Is it good or bad?



Inductive bias



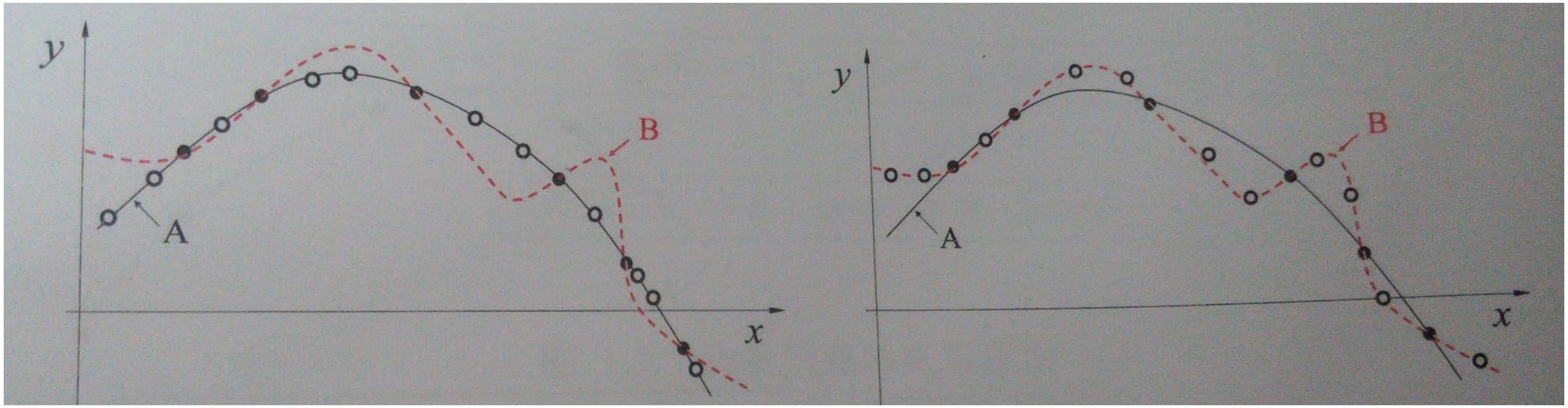
Occam's razor

（色泽=*；根蒂=蜷缩；敲声=浊响）

（色泽=*；根蒂=蜷缩；敲声=*）

Inductive bias is an assumption of "what is a good model"

Inductive bias



No Free Lunch Theorem

$$\sum_{f} E_{ote}(\mathfrak{L}_a | X, f) = \sum_{f} E_{ote}(\mathfrak{L}_b | X, f)$$

Inductive bias

$$E_{ote}(\mathfrak{L}_a|X,f) = \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\mathbb{I}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big)P(h|X,\mathfrak{L}_a)$$

$$\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\mathbb{I}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big)P(h|X,\mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \sum_f \mathbb{I}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1
\end{aligned}$$