Introduction to Machine Learning - CS4~~1~~9 725/403

Instructor: Prof. Ganesh Ramakrishnan

Lecture 2 - Supervised vs. Unsupervised Learning and Method of Least Squares

Supervised vs Unsupervised

Task: Suppose you had a basket and it is fulled with some fresh fruits your task is to arrange the same type fruits at one place. Suppose the fruits are apple, banana, cherry, grape

Case: 1

- You already know: Shape (parametrize shape?), Color

- **Train data**: Pre-classified data
- Goal: Learn from the pre-classified data and predict on new unclassified fruits.
- This type of learning is called as **supervised learning**.

Case 2:

- In this case, you know nothing about the fruits, you are seeing them for the first time!
- How will you arrange fruits of the same type together?
- One approach is to consider various characteristics of a fruit and divide them on the basis of that.
- Suppose you divide the fruits on the basis of *color* first.
  - ... papaya, orange, mango
  - ... guava, pear
- Now you take another physical characteristic, size. The grouping will then be:
  - ... papaya, melon
  - ... guava, pear
  - ..
  - ...
- ..

Case 2:

- In this case, you know nothing about the fruits, you are seeing them for the first time!
- How will you arrange fruits of the same type together?
- One approach is to consider various characteristics of a fruit and divide them on the basis of that.
- Suppose you divide the fruits on the basis of *color* first.
  - **Red Color Group**: Apples and cheery
  - **Green Color Group**: Bananas and grapes
- Now you take another physical characteristic, size. The grouping will then be:
  - **Red color and big size**: Apple
  - **Red color and small size**: Cheery
  - **Green color and big Size**: Banana
  - **Green color and small Size**: Grapes
- This type of learning is **unsupervised learning**

No prior cases that are labeled

Supervised Learning

Explicitly Specifed goals

implict goals: ease of retrieval, brevity (small size)

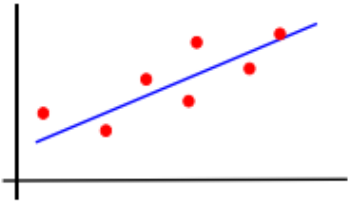
Unsupervised Learning

Goal of analysis is NOT explicit

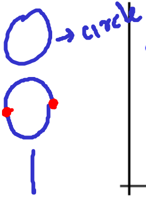
- In supervised learning, the desired outputs are provided which are used to train the machine whereas in unsupervised learning no desired outputs are provided, instead the data is analysed and studied through clustering, mining associations, reduce dimensionality, *etc.* into different classes

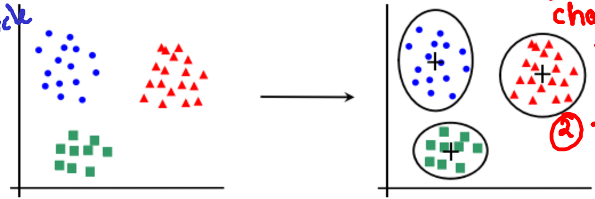
Three Canonical Learning Problems

1. Regression - Supervised
   - Estimate parameters, e.g. least square fit



2. Classification - Supervised
   - estimate class, eq handwritten digit classification

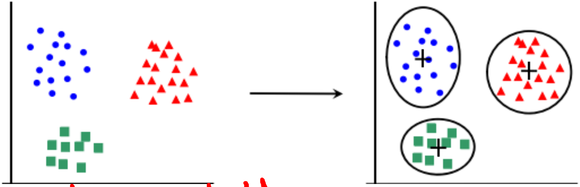




Features
① Params for polynomial characterization (piecewise)
② # of pixels

3 Unsupervised Learning - model the data

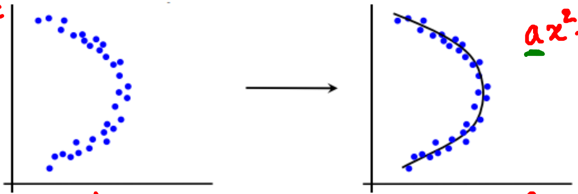
- clustering



Colors are hidden

- dimentionality reduction

Symmetric



ax^2+bx+c

set description

parabola (degree 2 poly)
=3 params

Supervised Learning

Functions F Training Data

$f : X \to Y \quad \{ (x^i, y^i) \in X * Y \}$

space of "interesting" functions
Eg: Linear fns

LEARNING

$$\boxed{\begin{array}{l}\text{find } \hat{f} \in \mathcal{F} \\ \text{s.t. } y_i \approx \hat{f}(x_i)\end{array}}$$

Learning machine

New data

PREDICTION $\boxed{y = \hat{f}(x)} \Longleftarrow \boxed{x}$

Need to do well
on new data

Next ....

We will start with linear regression and least square method to calculate parameters for linear regression problems.

Recap

- **Machine Learning in general**
  - Supervised Learning
  - Unsupervised Learning
  - Applications and examples
- **Canonical Learning Problems**
  - Regression Supervised
  - Classification Supervised
  - Unsupervised modeling of data

Agenda

- What is data? → Eg: Handwritten character recognition
  - Noise in data
- How to predict?
  - Fitting a curve
  - Error measurement
  - Minimizing Error
- Method of Least Squares

sum of squared error

set of sets of pixels

{ } , { }

1 2

{r, shift} 1

{length, shift} 2

{500 pixels} 1

{270 pixels} 2

We desire uniform representation for examples

What is data?

- For us, data is the information about the problem, you are solving using ML, in quantized form
- This data can be from any source, some examples are
  - Prices of stock and stock indexes such as BSE or Nifty
  - Prices of house, area and size of the house
  - Temperature of a place, latitude, longitude and time of year
- The objective of ML is to predict or classify something using the given data
- Hence, one or more than one parameters of the data must also represent the output of our program

Noise in Data

- Data in real life problems are generally collected through surveys, measuring instruments (scan a document)
- And surveys may have random human errors / machine errors
- Hence most methods we will be using deals with expectations as they minimize the effect of error in our predictions
- It is better to find outliers and clean data in the first step. This is known as data cleansing

often it requires unsupervised techniques

$$E[X] = \sum_x x p(x)$$

Example dataset for this lecture

- For this lecture we will consider variation of cost of the house with the area of the house
- In this example we want to find a pattern or curve which this dataset follows, hence predict the price for any value of area

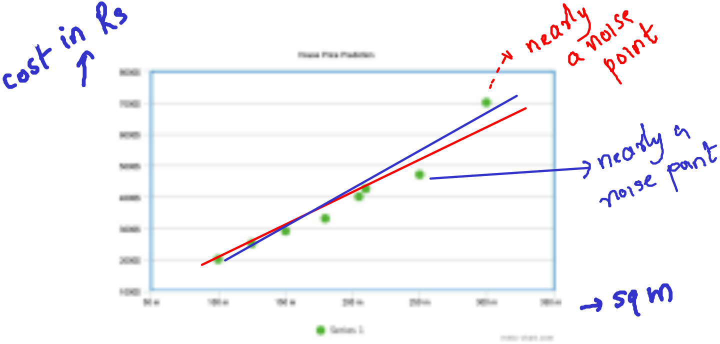
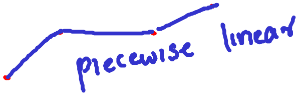


Figure: House purchase data - for illustration purpose only

How to predict?

Constraints: (1) degree of curve
(2) type of curve (smoothness reqd?)

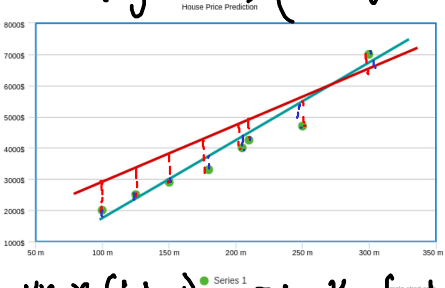
- Curve fitting is the process of constructing a curve, or mathematical function, that has the best fit to a series of data points, possibly subject to constraints. - Wikipedia
- Thus we need a critera to compare two curves on a dataset
- We describe an error function F(f, D) which takes a curve f and dataset D as input and returns a real number (goodness of f)
- Error function must be such that it can capture how worse is our



Example

- Consider the example below where we have two curves on our dataset defined by blue(f_b) and red(f_r) line respectively. We want to find which is the better fit.

For asymmetric view (regression) (Σ diff of y coords is interest-ing)



For symmetric views (later) Σ length of $\perp$ is interest-ing

Figure: House purchase data curve fit

Question

$$\sum_{x_i \in D} (y_i - f(x_i))^2$$

OR $$\sum_{x_i \in D} |y_i - f(x_i)|$$

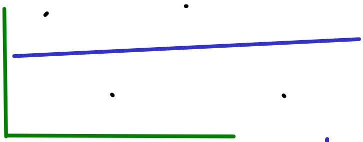
What are some options for F(f,D)?

Hint: Measurement of difference from original value.

Examples of F

- $\sum_D f(x_i) - y_i$
- $\sum_D |f(x_i) - y_i|$
- $\sum_D (f(x_i) - y_i)^2$
- $\sum_D (f(x_i) - y_i)^3$
- and many more

Problem: Cancellation effect



Close to "0" error!

Question

What F do you think can give us best fit curve and why?
Hint: Intuition of distances.

$\sum (f(x_i) - y_i)^2 =$ eucledian distance in predition space

Squared Error

$$\sum_{D}(f(x_{\text{i}}) - y_{\text{i}})^2$$

- To find the best fit curve we try to minimize the above function
- It is continuous and differentiable
- It can ve visualized as square of Euclidean distance between predicted points and actual points (distance in space of all pts)
- How we can perform mathematical treatment over this function will be covered in further lectures.
- This mathematical treatment is known as method of least squares. Can you find the reason why it is known as "Method of Least Squares"?
  Hint: Unit square is the basic unit in a graph.

Positive Definite (PD) matrix:

- M is p.d if $x^T M x > 0 \;\; \forall \; x \neq 0$
  (all eigenvalues are > 0)

Generally, we require M to be symmetric

- M is negative definite if $-M$ is positive definite
- M is positive semi-definite if
  $x^T M x \geq 0 \;\; \forall \; x$ (all eigen values are ≥ 0)
- M is negative semi-definite if $-M$ is positive semi-definite