Introduction to Advanced Machine Learning

Instructor: Dr. Muhammad Fahim

Self Introduction

- BS(Computer Science), Institute of Computing and Information Technology, Gomal University, 2007 (Gold Medalist)
- MS(Machine Learning), Department of Computer Science, National University of Computer and Emerging Sciences, 2009
- PhD(Artificial Intelligence), Department of Computer Engineering, Kyung Hee University, Feb. 2014
- Postdoc – Ubiquitous Computing Lab, Kyung Hee University, Aug. 2014
- Assistant Professor: Department of Computer Engineering, Istanbul S. Zaim University, Sep. 2014~ Till Aug.2017.
- Currently working with Innopolis University as an Assistant Professor.

Contact Information

- Muhammad Fahim
  - Office: 509
  - E-mail: m.fahim@innopolis.ru

- Vitaly Romanov
  - E-mail: v.romanov@Innopolis.ru

- Pavel Khakimov
  - E-mail: p.khakimov@innopolis.ru

- Course materials will be sent to you through moodle

Grading Criteria

• Top 10 Quizzes	10%
• Group Project (2 or 3 students per group)	20%
• Lab Participation (3 worst will be dropped)	20%
• Mid-term Exam + Lab Exam	20% + 5%
• Final Exam + Lab Exam	25% + 5%

Note: Please read the syllabus outline for late submission policy

Goals of this Course

- This course is designed for graduate students to provide comprehensive introduction and advance topics in machine learning.

- Student will learn to implement the machine learning models in Python programming environment from data science prospective.

- The end of the day they will able to apply machine learning algorithms to solve real-world problems.

Learning Outcome

- After this course, you will be able to:
  - Understand how machine can learn the concepts
  - Significant exposure to real-world implementations
  - To develop research interest in the theory and application of machine learning

Materials

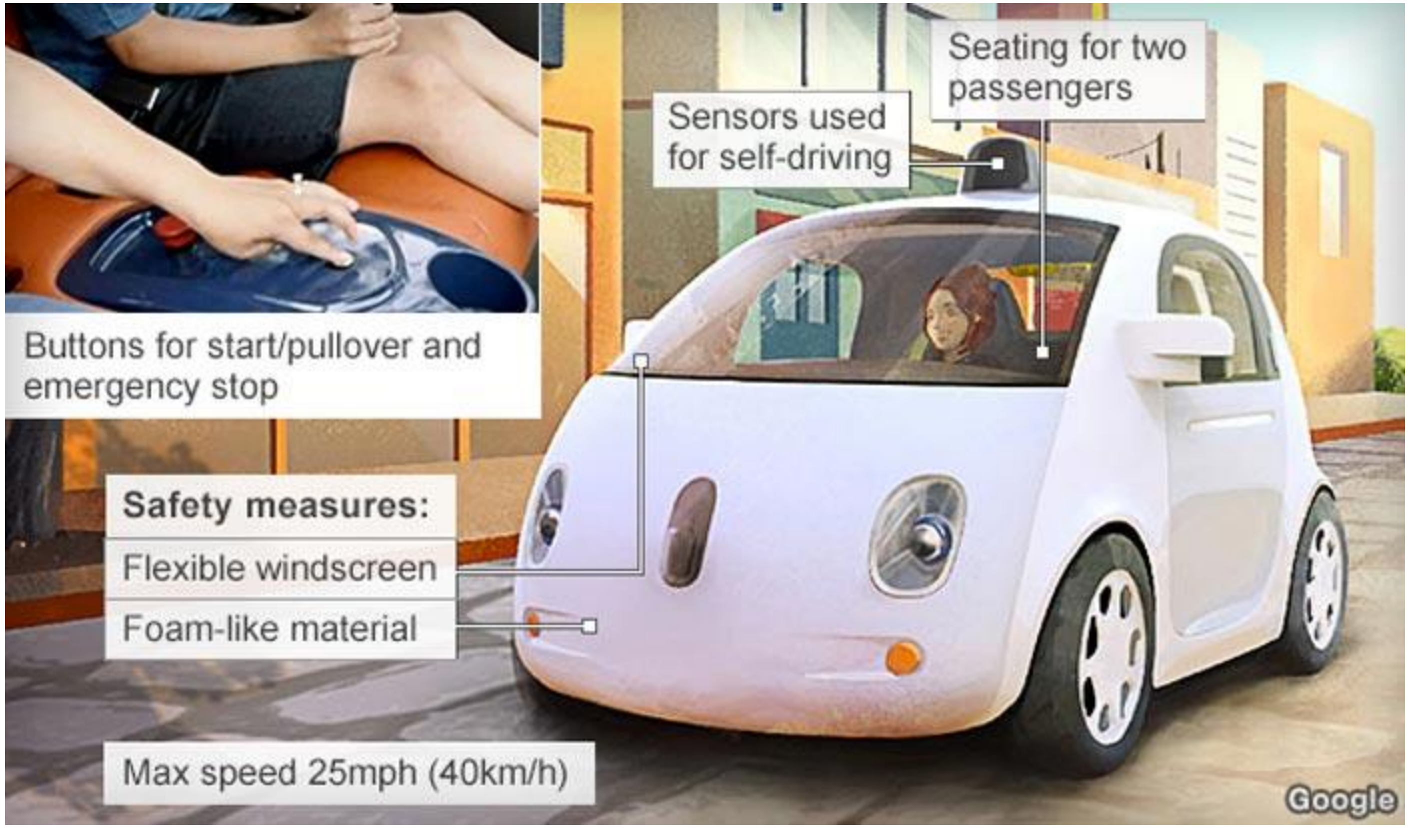
- **Text Book**
  - No specific text book

- **Reference Books**
  - *Machine Learning Probabilistic Approach by Kevin Murphy, MIT Press*
  - *Deep learning by Ian Goodfellow, MIT press*
  - *Pattern Recognition and Machine Learning by Christopher M. Bishop, Springer*
  - *Machine Learning by Tom M Mitchel, McGraw Hill*

- **My slides and shared references of research papers**

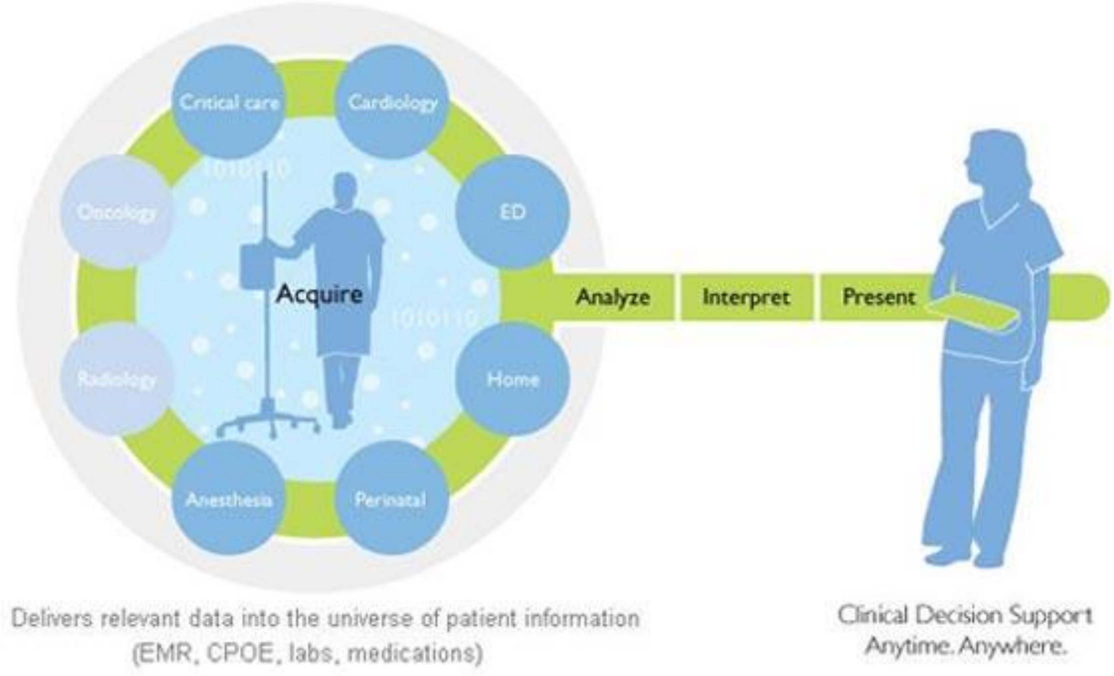
Ready to go .....☺

Machine Learning in Action



Machine Learning in Action

- Computers learning from medical records which treatments are most effective for new case



http://www.healthcare.philips.com/main/products/hi_pm/products/clinical_support.wpd

Machine Learning in Action

- Houses learning from experience to optimize energy costs based on the particular usage patterns of their occupants

Machine Learning in Action

- Helicopters can learn aerial tricks by watching other helicopters perform the stunts first

Machine Learning in Action

- Document Classification

Sports
Science
News

Machine Learning in Action

- Stock Market Prediction

Machine Learning in Action

- Weather Prediction

Machine Learning in Action

- Many, many more…
  - Speech recognition
  - Natural language processing
  - Computer vision
  - Sensor networks
  - Social networks
  - …

What is Machine Learning?

- **Definition:** "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

Tom M. Mitchel

Designing a Learning System

- **Choosing the Training Experience**
  - The first design choice we face is to choose the type of training experience from which our system will learn
  - The type of training experience available can have a significant impact on success or failure of the learner
  - Types of training experience
    - Direct or indirect
    - Teacher or not?

Designing a Learning System

- **Choosing the Target Function**
  - The next design choice is to determine exactly what type of knowledge will be learned and how this will be used by the performance program.

- **Choosing a Representation for the Target Function**
  - Now that we have specified the ideal target function, we must choose a representation that the learning program will use to describe the function that it will learn.

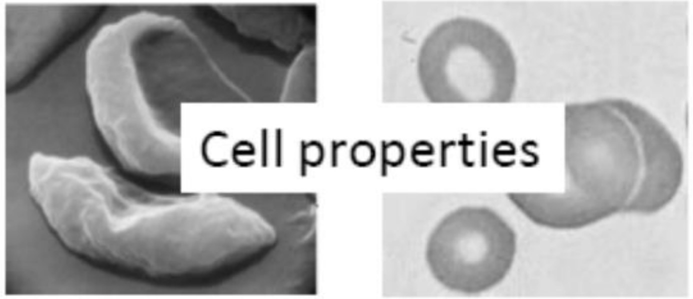
- **Choosing a Learning Algorithm**
  - Mechanism to learn from the experiences.

Types of Machine Learning

- **Supervised learning**
  - Where we get a set of training inputs and outputs. The correct output for the training samples is available
- **Unsupervised learning**
  - No specific output values are supplied with the learning patterns
- **Semi-supervised learning**
  - Where we get a small amount of labeled data with a large amount of unlabeled data
- **Active learning**
  - A special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points
- **Reinforcement learning**
  - Where there are no exact outputs supplied, but there is a reward (reinforcement) for desirable behavior

Discrete Labels

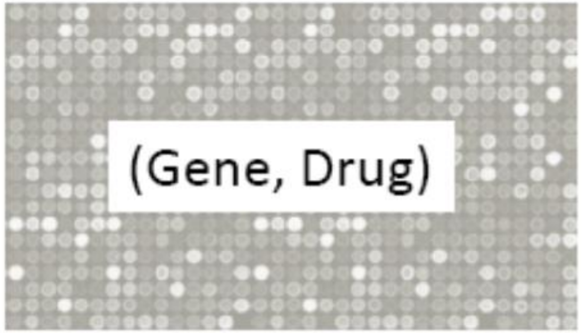
“Sports”
“News”
“Science”
…



“Anemic cell”
“Healthy cell”

Continuous Labels

Share Price
"$ 24.50"



Expression level
"0.01"

Machine Learning vs Statistical Modeling

- **Machine Learning is …**
  - *a subfield of computer science and artificial intelligence which deals with building systems that can learn from data, instead of explicitly programmed instructions.*

- **Statistical Modelling is …**
  - *a subfield of mathematics which deals with finding relationship between variables to predict an outcome.*

- Both the branches have learned from each other a lot and will further come closer in future

Source: https://www.infogix.com/blog/machine-learning-vs-statistical-modeling-the-real-difference/

Terminologies in Machine Learning

- Instance
- Label
- Feature
- Feature Column
- Example
- Model
- Metric
- Objective
- Pipeline

Terminologies in Machine Learning

- **Instance**: The thing about which you want to make a prediction. For example, the instance might be a web page that you want to classify as either "about cats" or "not about cats".

- **Label**: An answer for a prediction task either the answer produced by a machine learning system, or the right answer supplied in training data. For example, the label for a web page might be "about cats".

- **Feature**: A property of an instance used in a prediction task. For example, a web page might have a feature "contains the word 'cat'".

- **Feature Column**: A set of related features, such as the set of all possible countries in which users might live. An example may have one or more features present in a feature column.

Terminologies in Machine Learning

- **Example**: An instance (with its features) and a label.

- **Model**: A appropriate representation of a task. You *train* a model on examples then use the model to make predictions/classification etc.

- **Metric**: A number that you care about. May or may not be directly optimized.

- **Objective**: A metric that your algorithm is trying to optimize.

- **Pipeline**: The infrastructure surrounding a machine learning algorithm. Includes gathering the data from the front end, putting it into training data files, training one or more models, and exporting the models to production.

Important Points for ML Engineers

- Understand your data.
- Keep the first model simple and get the infrastructure right.
- Test the infrastructure independently from the machine learning.
- Turn heuristics into features (if possible).
- Know the freshness requirements of your system.
- Give feature column owners and documentation.
- Starting with an interpretable model makes debugging easier.
- Plan to launch and iterate
  - Combine and modify existing features to create new features in human understandable ways.
  - Tweak the model with different parameters (if applicable)
- You are not a typical end user – user experience strategies

Probability for Machine Learning

Random Variables

- A random variable is a **"probabilistic"** out-come
  - **For Example:** a coin flip or the height of a person chosen from a population

- Random values take on values in a sample space.

- This space may be discrete or continuous, and the space may be defined differently for different scenarios.

- For example sample space for:
  - a coin flip is {H,T}
  - a height might be defined as the positive real values in $(0,\infty)$
  - a temperature, it might be defined as real values in $(-\infty, \infty)$
  - the number of occurrences of a word in a document, it might be the positive integers $\{1, 2, . . .\}$.

Random Variables – Terminology

- The values of a random variable are called atoms.
- Random variables are written using capital letters and its realizations are written using lowercase letters.
  - **For Example:** X is a coin flip, and x is the value (H or T) of the coin flip.

Probability Functions

- A probability function maps the possible values of x against their respective probabilities of occurrence, such that
  - $f(x)$ is a number from 0 to 1.
  - The area under a probability function is always 1.

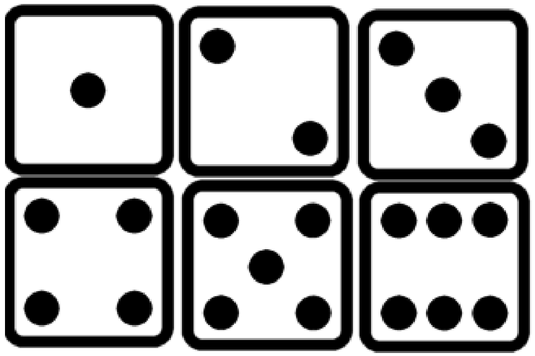
Discrete Random Variable

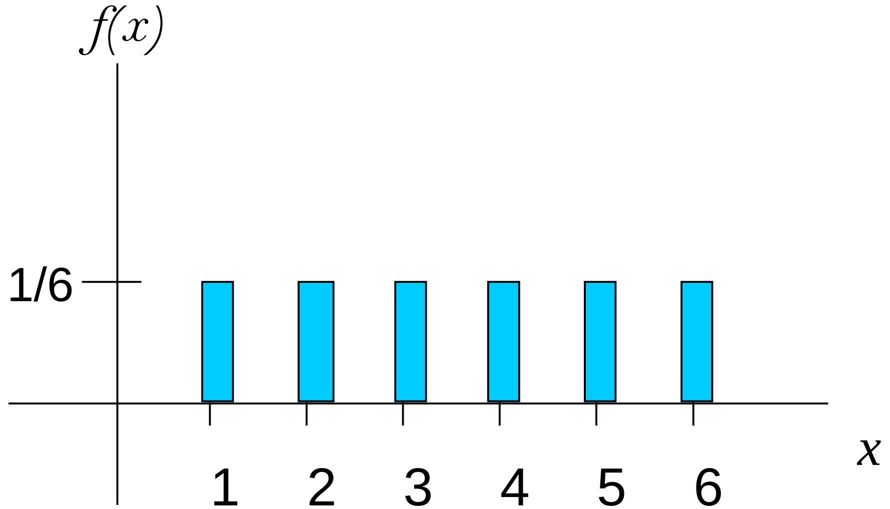
Probability Mass Function (pmf)

Cumulative Distribution Function (cdf)

Probability Mass Function *(pmf)*

x	p(X=x)
1	p(x=1) = 1/6
2	p(x=2) = 1/6
3	p(x=3) = 1/6
4	p(x=4) = 1/6
5	p(x=5) = 1/6
6	p(x=6) = 1/6

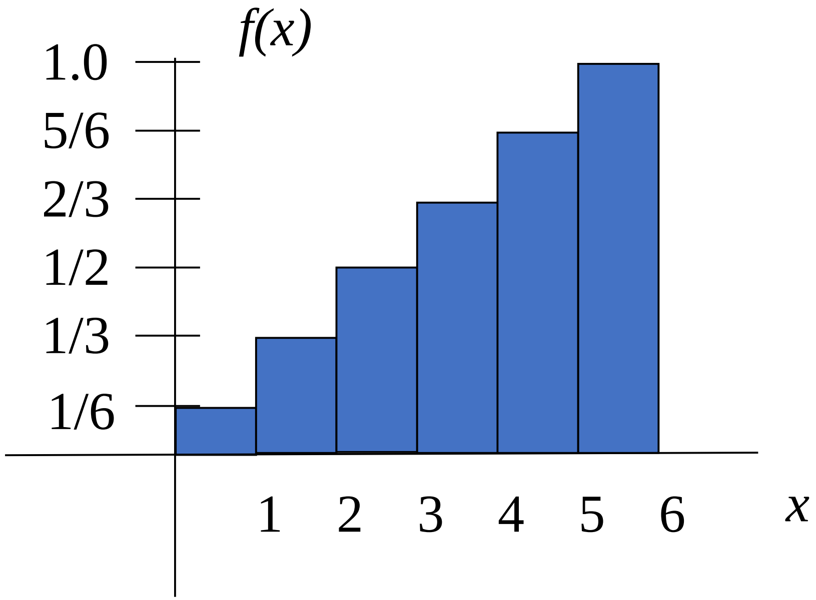




$$\sum_{\text{all } x} P(x) = 1$$

Cumulative Distribution Function (CDF)

X	p(X≤x)
1	p(X≤1) = 1/6
2	p(X≤2) = 2/6
3	p(X≤3) = 3/6
4	p(X≤4) = 4/6
5	p(X≤5) = 5/6
6	p(X≤6) = 6/6



Practice Problem

- The number of patients seen in the Emergency Room in any given hour is a random variable represented by *x*.
- The probability distribution for *x* is:

x	10	11	12	13	14
p(x)	0.4	0.2	0.2	0.1	0.1

Find the probability that in a given hour:

a. exactly 14 patients arrive $p(x=14) = .1$

b. At least 12 patients arrive $p(x \geq 12) = (.2 + .1 + .1) = .4$

c. At most 11 patients arrive $p(x \leq 11) = (.4 + .2) = .6$

Review Question

- If you toss a die, what's the probability that you roll a 3 or less?

a. 1/6
b. 1/3
c. 1/2
d. 5/6
e. 1.0

Answer: 1/2

Review Question

- Two dice are rolled and the sum of the face values is six. What is the probability that at least one of the dice came up a 3?

  a. 1/5
  b. 2/3
  c. 1/2
  d. 5/6
  e. 1.0

How can you get a 6 on two dice? 1-5, 5-1, 2-4, 4-2, 3-3
One of these five has a 3.
$\therefore 1/5$

Continuous Random Variable

Probability Density Function *(pdf)*
Cumulative Distribution Function *(cdf)*

Probability Density Functions (PDF)

- Let X be a **continuous** random variable
- Then a probability distribution or probability density function (pdf) of X is a function f(x) such that for any two numbers a and b with a ≤ b

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- The graph of $f(x)$ is often referred to as the *density curve*.

Probability Density Functions (PDF)

- **For Example:** x = continuous

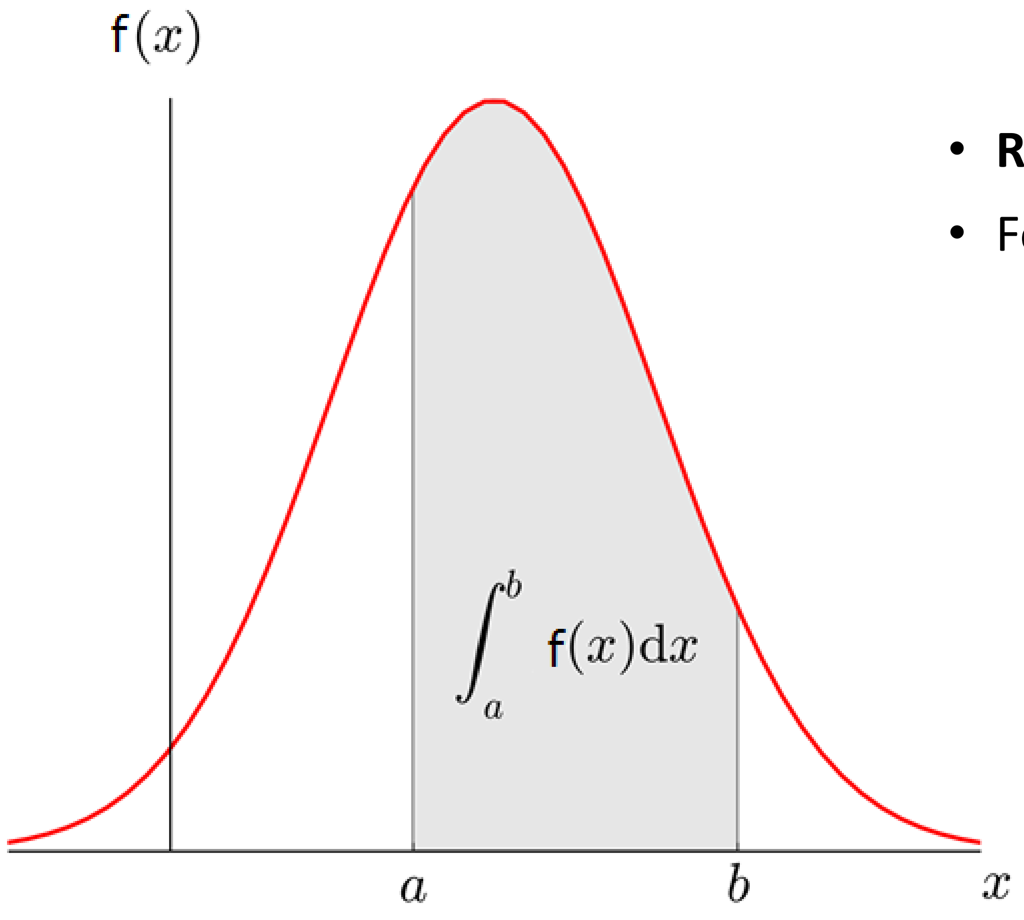
  = amount of rain tomorrow

- If we say tomorrow will be rain 2 mm and chances of rain is 90%.

  $$P(X=2) = 0.9$$

- In reality we can't say exactly 2 mm rain tomorrow it can be little less or more.
- Let us assume tomorrow will be the rain but it may be 2.2 mm or even less then 2 mm (i.e., 1.9) (Right?)
  - Hence x is continuous random variable.
  - So that $P(1.9<X<2.2)$
  - $P(a < X < b)$

Probability Density Functions (PDF)

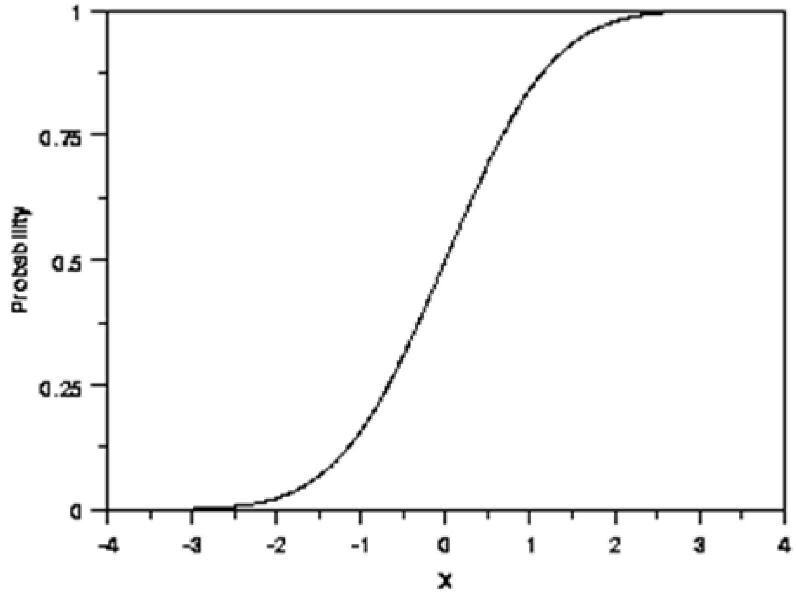


- **Remarks**
- For $f(x)$, it must satisfy the following two conditions:
  - $f(x) \geq 0$ for all x
  - $\int_{-\infty}^{\infty} f(x)\,dx = area\ under\ the\ entire\ graph\ of\ f(x) = 1$

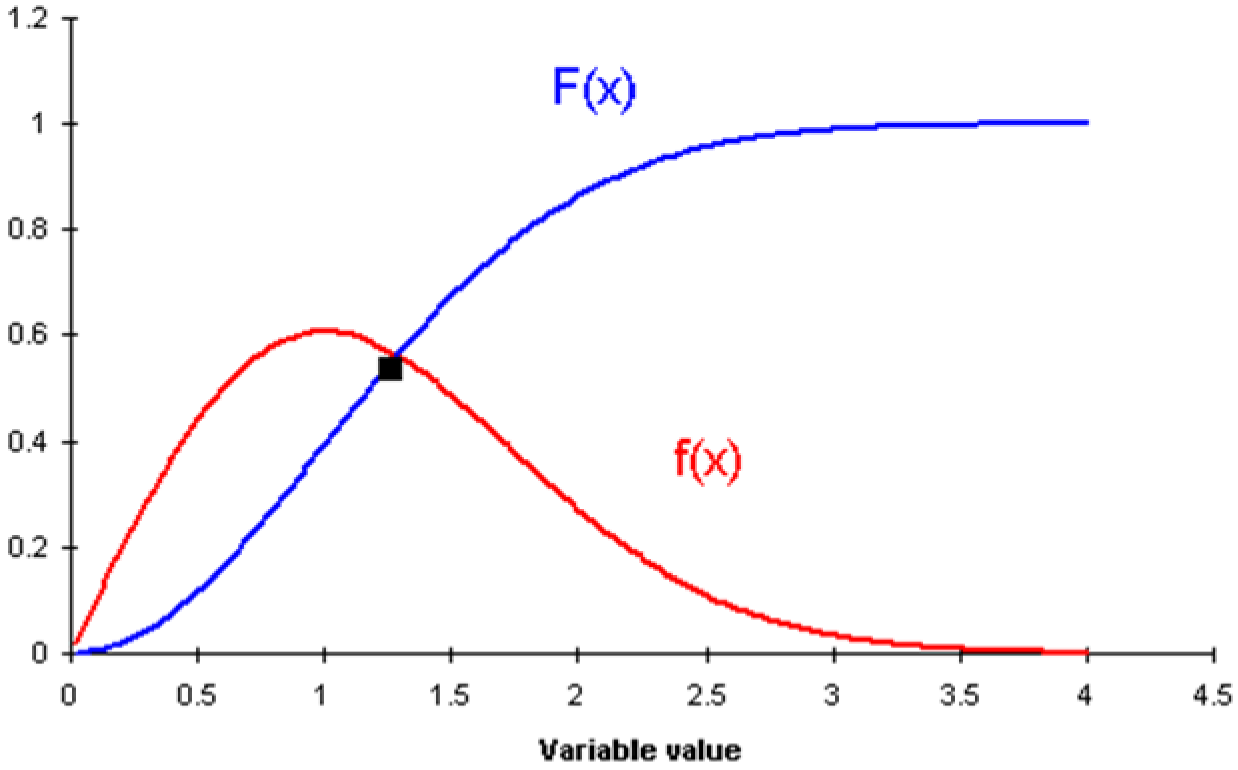
Cumulative Distribution Function (CDF)

- $F(x)$ describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x

$$P(X \le x) = F(x) = \int_{-\infty}^{x} f(t)dt$$



PDF vs. CDF



Overview of Probability Distribution

	Density	Cumulative
Discrete	PMF	CDF
Continuous	PDF	

Expectation

Motivation

- Often need to evaluate risk and decide how to proceed.

- **For Example:**
  - How much of an investment portfolio should go to stocks, and how much to bonds?

Expectation

- The **expectation** of a **discrete random variable** X taking the values a_1, a_2, . . . and with probability mass function p is the number:

$$E[X] = \sum_i a_i P(X = a_i) = \sum_i a_i p(a_i)$$

- We also call E[*X*] the ***expected value*** or ***mean*** of *X*.

- Since the expectation is determined by the probability distribution of *X* only, we also speak of the expectation or mean of the distribution.

Expectation

- **Example**
  - Let X represent the outcome of a roll of a fair six-sided die.
  - More specifically, X will be the number of pips showing on the top face of the die after the toss.
  - The possible values for X are 1, 2, 3, 4, 5, and 6, all equally likely (each having the probability of 1/6).
  - The expectation of X is

$$\mathrm{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

  - If you think about it, 3.5 is halfway between the possible values the die can take and so this is what you should have expected

Continuous Random Variable

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

Variance

- In probability theory and statistics, variance measures how far a set of numbers is spread out.
- A variance of zero indicates that all the values are identical.
- Variance of random variable X is defined as:

$$\operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)^2\right].$$

This can also be written as:
$\operatorname{Var}(X) = E(X^2) - \text{mean}^2$

$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx = \int x^2 f(x)\,dx - \mu^2$$

Covariance

- In probability theory and statistics, covariance is a measure of how much two random variables change together

$$\sigma(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

- Where E[X] is the expected value of X, also known as the mean of X.

- **Special case**
  - When the two variables are identical

$$\sigma(X,X) = \sigma^2(X).$$

Math for
Machine Learning

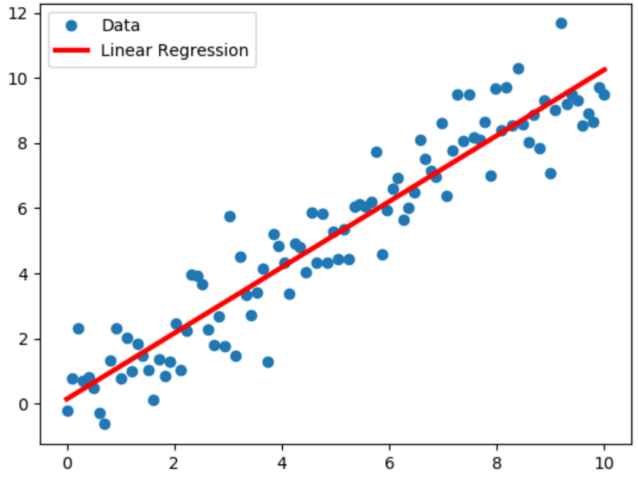
Math for Machine Learning

- **Calculus**
  - Calculus is classically the study of the relationship between variables and their rates of change. However, this is not what we use calculus for.
  - We use differential calculus as a method for finding extrema of functions
  - We use integral calculus as a method for probabilistic modeling.

Differential Calculus

- A classical statistics problem is linear regression.
- Suppose that we have a bunch of points:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$$



- We want to fit a line of the form

$$y = mx + b$$

- If we have lot of points, it's pretty unlikely that there is going to be a line that actually passes exactly through all of them.
- So we can ask instead for a line that lies as close to the points as possible.

Differential Calculus

- One easy option is to use squared error as a measure of closeness.

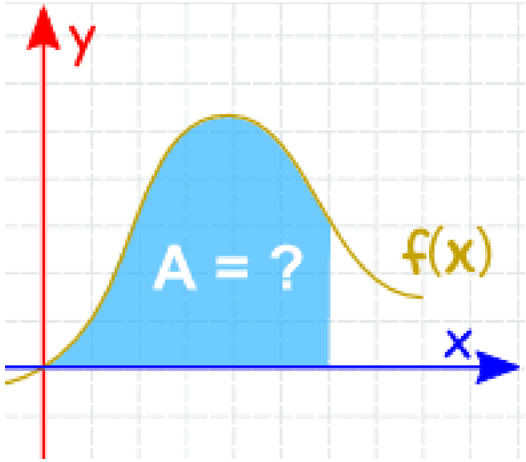
$$J(m,b) = \sum_{n=1}^{N} \left[(mx_n + b) - y_n\right]^2$$

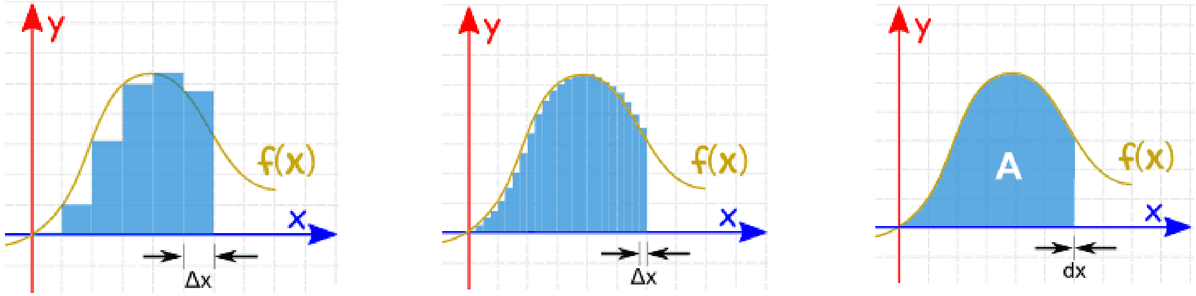
- **Note:** We have written the error J as a function of m and b, since, for any setting of m and b, we will get a different error.

- Our goal is to find values of m and b that minimize the error.
- How can we do this?
  - Differential calculus tells us that the minimum of the J function can be computed by finding its derivatives.

Integral Calculus

- An integral is the "opposite" of a derivative.
- Its most common use, at least by us, is in computing areas under a curve.
- We will never actually have to compute integrals by hand, though we should be familiar with their properties.





$$\int_a^b f(x)\mathrm{d}x$$

https://www.mathsisfun.com/calculus/integration-introduction.html

Convexity

- A convex function is, in many ways, "well behaved."
- Although not a precise definition – you can think of a convex function as one that has a single point at which the derivative goes to zero and this point is a minimum.
- One usually thinks of convex functions as functions that "hold water" – i.e., if you were to pour water into them, it wouldn't spill out.
- The opposite of a convex function is a concave function
- Convex functions look like valleys, concave functions like hills.
- **Why we care about convexity?**

Convexity

- The reason we care about convexity is because it means that finding minima is easy.

- **For instance:** the fact that $f(x)=2x^2-3x+1$ is convex means that once we've found a point that has a zero derivative, we have found the unique, global minimum.

- **For instance:** consider the function $f(x)=x^4+x^3-4x$. This function is non-convex.

More formally, a function f is convex on the range [a, b] if its second derivative is positive everywhere in that range.

Convexity

- **Example I:**
  - Consider the function $f(x)=2x^2-3x$. We've already computed the first derivative of this function: $\partial x f(x)=4x-3$.
  - To compute the second derivative of f, we just re-differentiate the derivative, yielding $\partial x \partial x f(x) = 4$.
  - Clearly, the function that maps everything to 4 is positive everywhere, so we know that *f* is convex.
- **Example II:**
  - Consider the non-convex function $f(x)=x^4+x^3-4x^2$.
  - The first derivative is $\partial x f(x)=4x^3+3x^2-8x$ and the second derivative is $12x^2+6x-8$.
  - It's fairly easy to find a value of x for which the second derivative is negative: 0 is such an example.
  - It is moderately interesting to note that while this f is not convex everywhere, it is convex in certain ranges, for instance the open intervals $(-\infty, -1)$ and $(0.5, \infty)$ are ranges over which f is convex.

Linear Algebra

- A large part of statistics and machine learning has to do with modeling data.
- For Example: we might characterize a car by it's length, width, height and maximum velocity.
- A given car can then be realized by a point in 4-dimensional space, where the value in each dimension corresponds to one of the properties we are measuring.
- Linear algebra gives us a set of tools for describing and manipulating such objects.
- **More details in lab sessions.**

Reference

- 1st Chapter of Tom Mitchell's book
- Read this article: http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf

Thank You ☺