CSC872 Pattern Analysis and Machine Intelligence

Fall 2017
Department of Computer Science
San Francisco State University

Ascending and Descending (1960), Waterfall (1961), by Maurits C. Escher
Introduction: the PAMI framework

CSC872
Pattern Analysis and Machine Intelligence

What is PAMI?

• Pattern Analysis and Machine Intelligence is a study for:
  – A modern artificial intelligence
  – Understanding the foundation of different approaches to make machines behave intelligently
  – Applying AI techniques to various engineering tasks
  – Type of researches that get published in IEEE trans on PAMI …
Collectively, we call them PAMI studies.

Well, which one should I use for my program...

QUESTIONs:
- Commonalities?
- Differences?
- Relationships?

Enormous!!!
The 3 questions: Common Framework

- What is PATTERN?
  - Codifying Properties of World
  - Data & Knowledge Representation

- What is MACHINE INTELLIGENCE?
  - Formalizing Intelligence for Machines
  - Problem Formulation

- What is ANALYSIS?
  - Analyzing Data & Knowledge to solve formulated problem
  - Problem Solving

Data & Knowledge Representation

- How to formally describe data/knowledge?
  - Algebraic Variables
    - Boolean, Scalar, Vector, Matrix, Tensor
  - Probabilistic Variables and Distributions
    - Random Variables, Probabilistic Mass/Density Function
  - Formal Rules
    - Rational Statement, Causality
  - Discrete & Continuous Relations
    - Tree, Graph, Function, Ontology
Problem Formulation: Problems?

- **Problems: what is the computational task?**
  - Inference
  - Modeling
  - Learning
  - Classification
  - Regression

Problem Formulation: Formulations?

- **Formalisms: How to describe the task?**
  - Agents
  - First Order Logic
  - Bayesian Inference/Classification
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriori Estimation (MAP)
  - Statistical Regression
  - Energy/Error Minimization
  - Maximum Information
  - Ensemble Learning
Problem Solving: Basics

- How to solve the problem w/ given data?
  - Search: Depth-First, Width-First, A*
  - Logical Inference: Resolution
  - Kernel Density Estimation (KDE)
  - Expectation-Maximization (EM) Algorithm
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Hill-Climbing/Gradient Descent
  - Simulated Annealing
  - Back Propagation
  - Support Vector Machine (SVM)
  - Markov Chain Monte Carlo (MCMC)
  - AdaBoost, Random Forest, ConvolutionNet, XGBoost...

PAMI Framework

- KR = Data & Knowledge Representation
- PF = Problem Formulation
- PS = Problem Solving
- Make your habit to think everything in the form of (KR-PF-PS)
- Example: you as a PAMI problem…
  - KR: your brain with all the details therein
  - PF: maximize amount and quality of learning
  - PS: taking and working in this course
# Course Overview

<table>
<thead>
<tr>
<th>Topic</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro &amp; Agent</td>
<td>AI: Ch1-2</td>
</tr>
<tr>
<td>Search Methods</td>
<td>AI: Ch3-4</td>
</tr>
<tr>
<td>Logic and Inference</td>
<td>AI: Ch7-9</td>
</tr>
<tr>
<td>Bayesian Framework</td>
<td>PR: Ch13-14 etc</td>
</tr>
<tr>
<td>Statistical Modeling</td>
<td>PR: Ch20 etc</td>
</tr>
<tr>
<td>Statistical Classification</td>
<td>PR: Ch3 etc</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>ML: Ch18 etc</td>
</tr>
<tr>
<td>Supervised Classification</td>
<td>ML: Ch3 etc</td>
</tr>
<tr>
<td>Supervised Regression</td>
<td>ML: Ch3 etc</td>
</tr>
<tr>
<td>Function Learning</td>
<td>NN: Ch20 etc</td>
</tr>
</tbody>
</table>

# Course Information

- **URL:** [https://bidal.sfsu.edu/~kazokada/csc872](https://bidal.sfsu.edu/~kazokada/csc872)
- **Instructor**
  - Kaz Okada, kazokada@sfsu.edu
  - OH: TH911, Wed: 4:00 – 5:00pm
- **TA:**
  - Andrew Scott, ats@mail.sfsu.edu
  - OH: SCI241, Tue: 2:00-3:00pm
- **Grading:** homework/project/
  - 50% Homework
  - 25% Final Report
  - 10% Final Presentation
  - 15% Fast Prototyping
- **Policies:** Please read the web.
  Be aware of the deadline and the late policy!!!
Evaluations (Exams?)

- No Midterm/Final Exams
- Homework (50% of total grades)
  - Five HWs (See course webpage for schedule)
  - Due in one/two week(s)
  - Involves some difficult analytical problem solving
- Final Project (25% report, 10% presen)
  - Final Presentation (Presentation) on Dec 12.
  - Assignments given in the course web (follow the link)
- Fast Prototyping (15%: 5% each)
- Extra Credit of 5% for completing them all

Text Books

- AIMA by Russell-Norvig: our text, general AI
  - Duda-Hurt: for PR foundation
  - Hastie: advanced ML
  - Gonzales-Woods: comprehensive IP&CV

- Read the AIMA chapters before the classes
- Additional reading assignments given as appropriate
- Course Slides published online AFTER lectures
Course Schedule

• Consult the course homepage for details of the lecture plan

• First part: lecture 4:00 – 5:30 ca (90min)
• 10 min break
• Next part: in-class exercise 5:40 – 6:45 (65min)

• Drop deadline: Sep 13 (three weeks)
• Oct 31\textsuperscript{th}: No Lecture
• Nov 22\textsuperscript{nd}: No Lecture

MATLAB

• You will learn how to use a powerful prototyping software environment !!!
  – Exercise Tutorials
  – Fast Prototyping

• You need to bring a laptop with MATLAB by the next lecture
  – Student copy at bookstore (reasonably priced)
  – Free copy at COSE: network access. Standard version but only within the SFSU network.
  – Contact Andrew and I as soon as you can!
  – Free MATLAB clones are NOT RECOMMENDED.
In-Class Exercises

• MATLAB Exercises
  – Basics of MATLAB
  – Three exercise sessions
  – Hands-on tutorials
  – TA and my help during office hours
  – End up learning a useful tool

• Fast Prototyping Exercises
  – Hands-on MATLAB software prototyping guided exercise
  – Three algorithms: PCA, Mean Shift, LDA
  – 15% of the total grades!
  – 5% extra credit for completing all of them in-class
  – End up learning how to quickly implement your ideas
defying all the nice thing you learned in SE classes.

• Bring your own laptop with MATLAB!!!

Roll Call

• Adding to the course?
• Pre-reqs?
Review: Basic Concepts

• Some relevant mathematical ideas:
  – Calculus (high-school to lower-division)
  – Algebra (high-school to lower-division)
  – Probability (basic + some advanced)
  – Statistics (basic + some advanced)

• You want to make sure you are comfortable with these concepts and notations

OK… some refresher now;

KR: Variable: Scalar & Vector

• Variable is:
  – Symbolic representation of quantity
  – Unknown quantity that can change in algebraic sense
  – Measurable attribute of a system in statistics

• Scalar X : Variable indicating a single-valued entity

• Vector X : Variable indicating a multiple-valued entity

\[
x = (x, y)^T
x = a \text{ : area}
x = \mu \text{ : angle}
\]

Dimension := number of coeffs
**KR: Continuous vs. Discrete**

- **Continuous Variable** $X$
  - indicates real-value entities
  - $x \in \mathbb{R}$
  - $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$

- **Discrete Variable** $X_i$
  - only take a set of predetermined discrete values
    - $x_i \in \mathbb{N}$ Natural number: $i = 1, 2, \ldots$
    - $x_i \in \{MO, TU, WE, TH, FR, SU, SA\}$

**KR: Function**

- Deterministic dependence of two quantities/sets, associating input $X$ to output $Y$ by a binary relation

  ![Diagram of function](image)

- $f : X \rightarrow Y$
- Map, Mapping, Transformation = Function
- Inverse function: $g = f^{-1} : Y \rightarrow X$
KR: Function Properties

- Rules of $f$ described in a graphical plot or sometimes in analytic formula when known

- Continuous Function
- Differentiability
- Smooth Function
  - All-order differentiable over entire domain

KR: Matrix

- Product does not commute: $AB \neq BA$
- Transpose: $A^T: a_{ij} \leftarrow a_{ji}$ (swapping rows & columns): $(AB)^T = B^T A^T$
- Symmetric matrix $A$: $A^T = A : a_{ij} = a_{ji}$
- Inverse matrix of $A^{-1}: A A^{-1} = A^{-1} A = I_n$
- Orthogonal matrix $A$: $A^T = A^{-1}: AA^T = A^T A = I_n$
- Outer and inner product

$x = \begin{pmatrix} x \\ y \end{pmatrix}$

$x^T = (x \ y)^T$

$x x^T = \begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix}$

$x^T x = x^2 + y^2 = \text{tr}(xx^T)$
## PF: Matrix Equations

- **Linear equations**
  
  \[ ax_1 + bx_2 = e \]
  \[ cx_1 + dx_2 = f \]
  \[ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = A^{-1} \begin{bmatrix} b \\ e \end{bmatrix} \]

- **Linear transform**

  \[ \begin{bmatrix} y \\ x \end{bmatrix} = A \begin{bmatrix} q \\ a \\ b \\ c \\ d \end{bmatrix} \]

- **Eigen-values & vectors**

  \[ \lambda v = Av; \lambda \in \mathbb{R} \]

## KR: Graph

- **Undirected graph**: \( G = (V, E) \)
  - vertices and edges (no direction)

- **Directed graph**: \( G = (V, A) \)
  - vertices and arrows

- **Directed acyclic graph (DAG)**
  - Directed graph without a loop

- **Connected graph**
  - Can reach from any vertex from any other vertices

- **Connected DAG**
  - Tree
**KR: Boolean Variable**

- A is a Boolean variable if it indicates two-valued system, a statement or event
  - e.g., indicator variable $A = \{\text{Yes, No}\}$
  - e.g., $A = \text{My name is George}$
  - e.g., $A = \text{I teach CSC872}$

- Some event has intrinsic degree of **uncertainty** as to whether $A$ occurs
  - e.g., $A = \text{There will be an earthquake tomorrow}$
  - e.g., $A = \text{My stock price will go up tomorrow}$

- **Random Variable** is a function that chooses a value from the event space $\{\text{True, False}\}$ according to probability $P(A)$

---

**KR: Basic Probability**

- $P(A)$ means "the fraction of possible worlds in which $A$ is true"

  - Event space of all possible worlds
  - Its area is 1

  - A is true
  - A is false

  - $P(A)$ is the area of the pink circle

- The axioms of probability !!!
  - $0 \leq P(A) \leq 1$
  - $P(\text{True}) = 1$
  - $P(\text{False}) = 0$
  - $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

• Joint Probability
  – Probability of two events in conjunction
  – \( P(A \text{ and } B) := P(A \cap B) := P(A, B) \)

• Marginal Probability
  – Probability of one event (A) regardless of the other events (B)
  – Obtained by summing (integrating) a joint probability over the event space \( \Omega \) for unwanted events (B)

  \[
  P(A) = \sum_{v \in \Omega} P(A \cap B = v) = P(A \cap B) + P(A \cap \neg B)
  \]

  – \( P(\text{not } A) := P(\neg A) \)
  – \( P(A) + P(\neg A) = 1 \) \, total probability theorem


• Conditional Probability
  – Probability of an event (A) given other event (B)
  – \( P(A|B) = \frac{\text{area of } A \text{ and } B}{\text{area of } B} \)
  – \( P(A|B) = \frac{P(A \cap B)}{P(B)} \)

X = College Major
Y = Likes “XBOX”

• Product Rule
  – Joint probability can be written as a product of a conditional and a marginal
  – \( P(A \cap B) = P(A|B)P(B) \)
    \( = P(B|A)P(A) \)

• Statistical Independence satisfies
  – \( P(A \cap B) = P(A)P(B) \)
  – \( P(A|B) = P(A) \)
  – \( P(B|A) = P(B) \)
KR: Beyond Boolean Events

- When more than one state (over a discrete variable):
  - e.g., \( X = \text{day}, \Omega := \{\text{Mon,..,Sun}\} \)
  - Discrete Random Variable
    - \( p(X = v_i \text{ and } X = v_j) = 0 \text{ if } i \neq j \) (mutually exclusive)
    - \( p(X = v_i) = \sum_{j=1}^{t} p(X = v_j) = 1 \) (total prob. Th.)
    - \( p(Y) = \sum_{i=1}^{t} p(Y \text{ and } X = v_i) \) (marginal)

- When over continuous variable:
  - Continuous Random Variable
  - e.g., \( X = \text{temperature of SF} \)

KR: Probability Distribution

- For a discrete random variable \( X \)
  - Probability Mass Function
    \[ p(X = x_i) \]
    \[ \sum_{i=1}^{t} p(X = x_i) = 1 \]

- For a real-valued random variable \( X \)
  - Probability Density Function
    \[ p(x) \]
    \[ \int_{0}^{1} p(x) \, dx = 1 \]
    \[ P(a < X \leq b) = \int_{a}^{b} p(x) \, dx \]
**KR: Expectation**

- For a discrete random variable $X$
  - $E[X] = \sum_{\Omega} x_i P(X=x_i) = \mu$ (population mean)
  - $E[f(X)] = \sum_{\Omega} f(X=x_i)P(X=x_i)$

- For a real-valued random variable $X$
  - $E[X] = \int_{\Omega} xP(y)dy$

- Linearity
  - $E[aX+Y] = aE[X]+E[Y] = a\mu_X + \mu_Y$

**PF: Statistics**

- Independent and Identically-Distributed (i.i.d.) Random Variable
  - Rolling a fair dice for instance.
  - If $x_1, x_2, x_3, \ldots, x_i, \ldots, x_k$ are i.i.d. of $X$ then
  - $P(x_1, x_2, x_3, \ldots, x_i, \ldots, x_k) = P(X=x_1)P(X=x_2)\ldots P(X=x_k)$

- Central limit theorem
  - The sum of i.i.d. random variables with finite variance will be approximately normally (Gaussian) distributed as we go towards an infinite number of samples.
  - A reason why you see a lot of Gaussians …