CS 726
Advanced Machine Learning
Course Overview

Sunita Sarawagi
Spring 2025

Welcome!

Scope of the course

Learning to represent, generate, and reason on objects:

- High dimensional $x = \{x_1, \ldots, x_n\}$, space of x is large
- Inter-dependent components

Examples:

- Image
- Video
- Time-series
- Text

Examples of high dimensional spaces

1024

1024

This image is very high-dimensional: comprising of 1024*1024*3 ≈ 3 million dimensional real space

Words in a sentence

If you ask a question, you are a fool only once. If you do not ask, you are a fool forever.

Assume a vocabulary size of 50 K.

The sentence of 25 words has 25*50 K = 1.25 million dimensional discrete space

Different task settings

Given training data D, train a model M that can be used for

- Generation
  - Unconditional: Generate a sample X that is representative in D
  - Conditional: Given an input prompt X, generate a likely sample Y.
- Density estimation:
  - What is the probability that a given sample X is part of the training distribution D
- Other forms of reasoning:
  - Causality, Counter-factual reasoning, recourse on predictions.

Text to text generation

- Write a poem

$\underset{X}{\text{Instruction}} \xrightarrow{\text{LLM}} \underset{Y}{\text{Poem}}$

- Translation

- Text-to-tree generation

Translation

Input: x

Predicted sequence: y

Where can I find healthy and traditional Indian food? → स्वस्थ और पारंपरिक भारतीय भोजन कहां मिल सकता है?

y_1 y_2 y_3 y_4 y_5 y_6 y_7 y_8

- Each token in the output is a random variable and there is inter-dependence in the output tokens.

- We want to output a probability with the output translation, and not just produce one translation.

- We cannot predict the whole sentence in one shot but need to decompose it into parts

Text to image generation

- Imagen
- Stable diffusion

Topics for Generation

Goal: Output a distribution $P_\theta(\boldsymbol{y}|x)$ over a structured output $\boldsymbol{y} = y_1, \dots, y_n$, optionally conditioned on an input x.

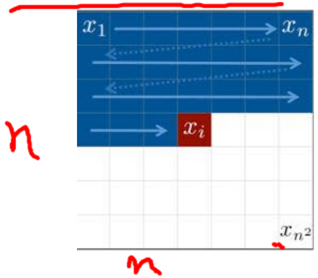
- **Representation/Modeling:** Form of P_θ, how to represent $P(\boldsymbol{y})$ of high-dimensional y for easy learnability and efficient inference.
- **Training or learning:** How to parameterize the distribution and learn the parameters
- **Inference**: How to efficiently generate?

Key insight from the course

Decompose high-dimensional objects into smaller manageable sub-parts

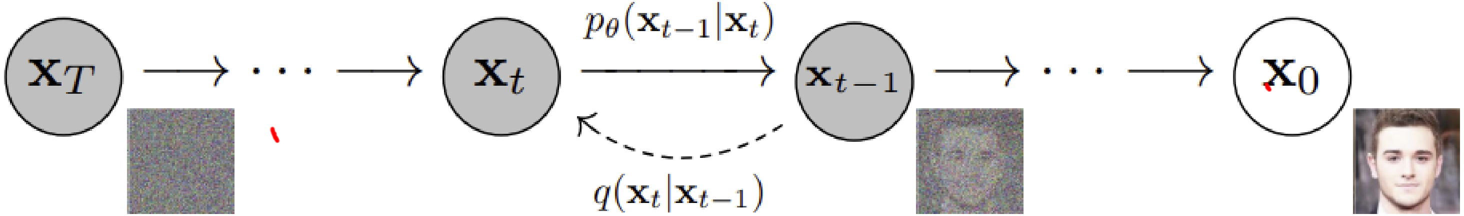
Representation

- With observed variables



$$P(x_1, x_2 \ldots x_n) = P(x_1)\,P(x_2|x_1)\,P(x_3|x_1, x_2) \cdots P(x_i|\underbrace{x_1 \cdots x_{i-1}}) \cdots P(x_{n^2}|\ldots)$$

- With latent variables



Can we make the dependency graph simpler via factorization?

Representation

- Represent the rate of change of a random variable (stochastic differential equations)

$$P(\underline{X} \mid t)$$

time

$$\frac{\partial}{\partial t} P(x \mid t)$$

stochastic differential equation

Learning

- How to parameterize the joint distribution for sample-efficient learning

- How to efficiently learn the parameters θ of the distribution
  - Training data (conditional): $D = \{(x^1, \boldsymbol{y}^1), \ldots, (x^N, \boldsymbol{y}^N)\}$
  - Training data: (unconditional) $D = \{x^1, x^2, \ldots, x^N\}$

Adapting trained distributions

- In-context learning for regression, time-series, and language tasks
- Parameter efficient fine-tuning

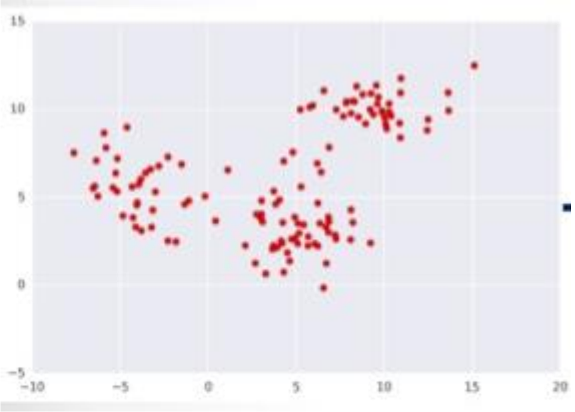
Inference

- Given a x, how to efficiently find the most likely $y_1, \ldots, y_n$: MAP Inference.
- How to generate multiple representative examples from estimated model: Sampling
  - Generate examples that are representative of the distribution

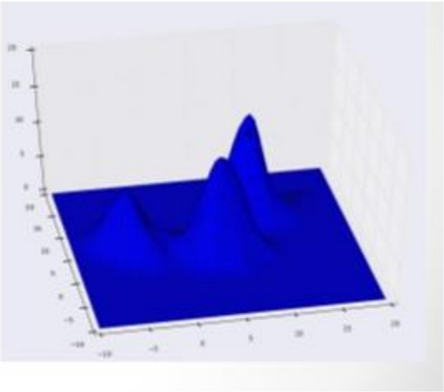
Density estimation

Given D = $\{x^1, x^2, ..., x^N\}$ learn a P(x), so that given a new x we can efficiently calculate the probability of “x”.

Applications: Out of distribution detection, outlier detection, classification



Density estimator



What is causal inference?

Inferring the effects of any treatment/policy/intervention/etc.

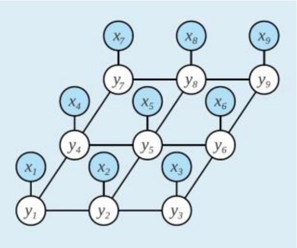
Examples:

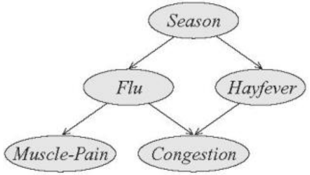
- Effect of treatment on a disease
- Effect of climate change policy on emissions
- Effect of social media on mental health
- Many more (effect of X on Y)
- Effect of Interest rate on inflation
- Effect of air pollution on cancer

Counterfactual reasoning

1. Would I have been happier if I went to IITK instead of IITB.
2. Would demand be higher if discount was offered.

Course contents





Representation of P(X) or P(Y|X)

- Probabilistic graphical models: Bayesian Networks and Markov Random Fields
    - Exact, efficient, but limited capacity
    - But, important to understand them to build a framework for probabilistic reasoning
    - Intuitive and easy to incorporate prior knowledge and biases
  - Special Graphical models
    - Gaussian processes: special structure that allow trivial computation of marginals

Representation (continued)

- Deep latent variable models:
  - VAEs, GANs, Discrete diffusion models – technology behind latest image generation models such as ImageGen
- Representation via variable transformation: Normalizing flows
- Stochastic differential equations P(Y|X) where X is time and distribution represented as rate of change → continuous time diffusion model

Course contents

Learning

- Parameterization (model architectures for efficient learning)
  - Feature-based like in CRFs
  - Deep neural methods e.g. transformers
- Training algorithms
  - Maximum likelihood learning
  - Generalized Expectation Maximization: Variational Auto Encoders, diffusion models for images

Learning (continued)

- Advanced topics from deep learning:
  - In-context learning in foundation models
  - Parameter efficient fine-tuning
  - Model editing

Course contents

Inference

- Boolean queries on conditional inference
- Marginalization queries: P(Xi), max_x P(x)
    - Sum-product and max-product Inference in Graphical Models
- Sampling
    - Classical methods of sampling in tractable model: forward sampling, importance weighted sampling, Markov Chain Monte Carlo sampling (MCMC),
    - Recent methods usable in deep learning: Monte-Carlo with Langevin dynamics

Inference (Continued)

- Inference challenges in modern LLMs (a special Bayesian network)
    - Limitations of greedy decoding
    - Sampling multiple generations
    - Grammar constrained decoding
    - Speculative decoding

- Other forms of Inference
    - Causal effects
    - Algorithmic recourse

Who should take the course

- Students who are interested in doing research in machine learning
- Students who want to learn to think about learning from a probabilistic perspective in the context of modern deep learning
- Students who want to model learning tasks in a manner that cuts across applications.
    - The course will cite applications in NLP, vision, time-series, event sequences, and speech when relevant but it is not primarily about any of these applications.

Mode of running the course

- Two 85 minute slots per week:
- SAFE/Moodle quiz on the material covered in the **prior** week
  - 20 minute duration at a pre-announced time.
  - Grading will be done on top n-2 out of n quizzes. No compensation for missed quizzes.
  - First quiz on Jan 15^{th} on probability and ML basics
- All materials will be uploaded on Moodle, announcements via Moodle, questions on Moodle or cs726@googlegroups.com
  - Forum for each topic for discussions and questions.

Evaluation

Approximate credit structure

- 15% In-class Quizzes
- 20% 4—6 graded programming and paper homeworks (in teams of 3)
- 25% Mid-semester exam
- 35% End semester exam
- 3% Scribing
- 2% Attendance and class participation

Course calendar https://www.cse.iitb.ac.in/~sunita/cs726/