CS 726: Course Overview

Sunita Sarawagi

Scope of the course

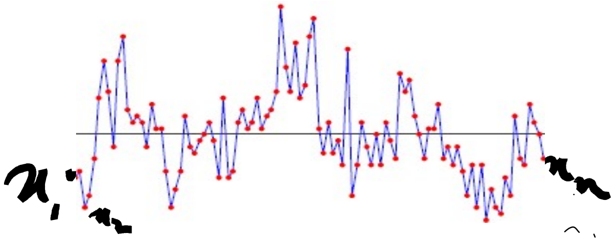
Learning to represent, predict, generate, and reason on objects comprising of several inter-dependent variables.

- Object: high dimensional x = {x_1,.....,x_n}, space of x is large
  - ○ Sentence, POS tags to each word in a sentence, translation

This is a sentence
x_1 x_2 x_3 x_4

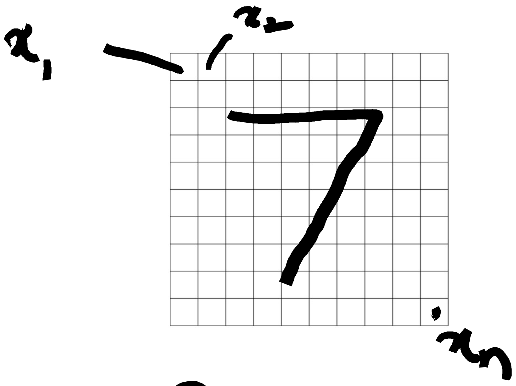
Det, verb, article, Noun
y_1 y_2 y_3 y_4

  - ○ Time series data



$x_i \in \mathbb{R}$.

Examples of Objects



$x_i \in \{0, 1\}$

Pixels of an image

y_i = depth of each pixel

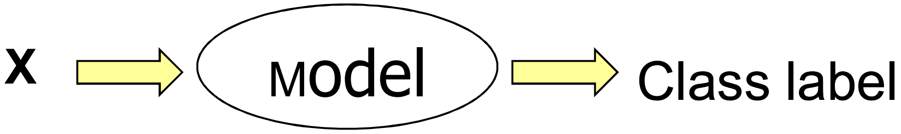
$y_i \in \mathbb{R}$.

Different task settings

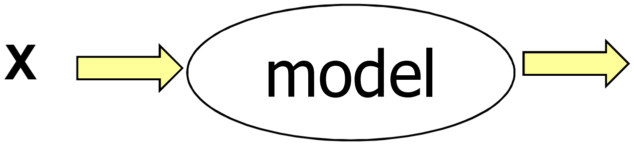
- Prediction:
    - Two types of variables: input and output. Output comprises of multiple variables.
- Generation: Two types
    - All variables are observed during training, during inference we need to generate new samples.
    - Variables are partitioned into an observed and hidden set. During training samples contain only observed variables. We need to reason about hidden variables to understand the data or generate more data
- Density estimation:
    - What is the probability that a given sample is part of the training distribution
- Other forms of reasoning:
    - Causality, Counter-factual reasoning, recourse on predictions.

Structured prediction

- Standard classification



- Structured prediction



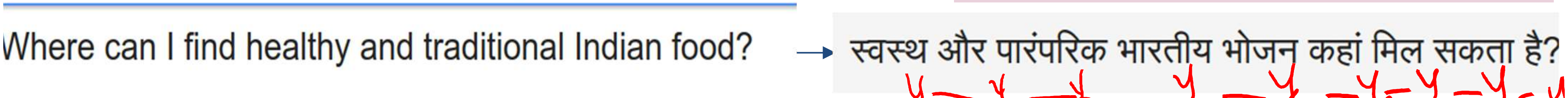
Structured y

1. Vector: y_1 ,y_2,..,y_n
2. Segmentation
3. Tree
4. Alignment
5. ..

Translation

Input: **x**

Predicted sequence: **y**



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

$y_1 - y_2 - y_3 - y_4 - y_5 - y_6 - y_7 - y_8 - y$

- 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. $P(y^1|x) . P(y^2|x); \cdot P(y^k|x)$
- We cannot predict the whole sentence in one shot but need to decompose it into parts

Image captioning

Input: **x**

Predicted sequence: **y**

→

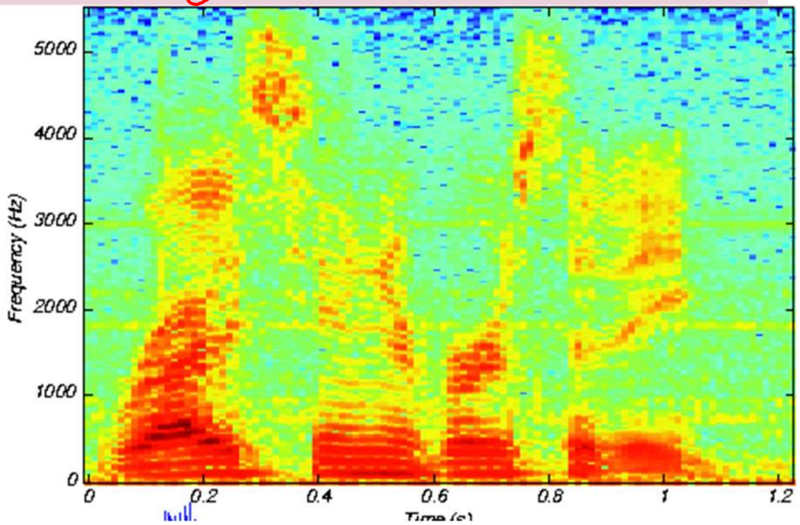
A person riding a motorcycle on a dirt road

$y_1 - y_2 - y_3 - y_4 \cdots - y$

Image from http://idealog.co.nz/tech/2014/11/googles-latest-auto-captioning-experiment-and-its-deep-fascination-artificial-intelligence

Speech recognition

Context: **x** **(Speech spectrogram)**



→

Output: **Y** **(Phoneme Sequence)**

Ri ce Uni ver si ty

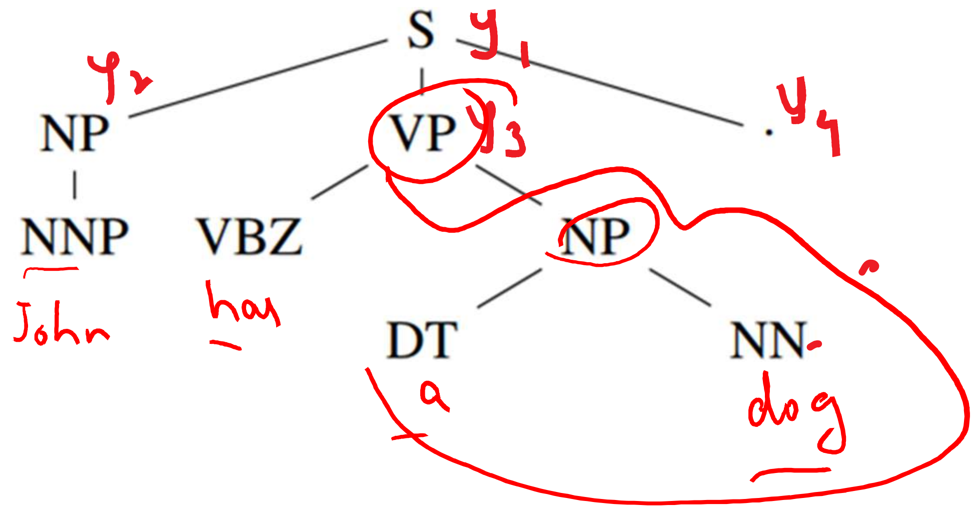
$y_1 - y_2 - y_3$ y_4

Syntactic parsing

Input: **x**

John has a dog . →

Predicted output: **y**



Show me my average SPI over the past 3 semesters → SQL → tree

ASC

Topics for Prediction

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

- Form of P_θ, how to factorize $P(y_1, \dots, y_n|x)$ into simpler parts : **Representation/Modeling**
- How to efficiently learn the parameters θ of the distribution: **Learning**
  - Training data: $D = \{(x^1, \boldsymbol{y}^1), \dots, (x^N, \boldsymbol{y}^N)\}$
- Given a x, how to efficiently find the most likely $y_1, \dots, y_n$: **Inference.**

Different task settings

- Prediction: 60
    - Two types of variables: input and output. Output comprises of multiple variables.
- Generation: Two types
    - All variables are observed during training, during inference we need to generate new samples.
    - Variables are partitioned into an observed and hidden set. During training samples contain only observed variables. We need to reason about hidden variables to understand the data or generate more data
- Density estimation:
    - What is the probability that a given sample is part of the training distribution
- Other forms of reasoning:
    - Causality, Counter-factual reasoning, recourse on predictions.

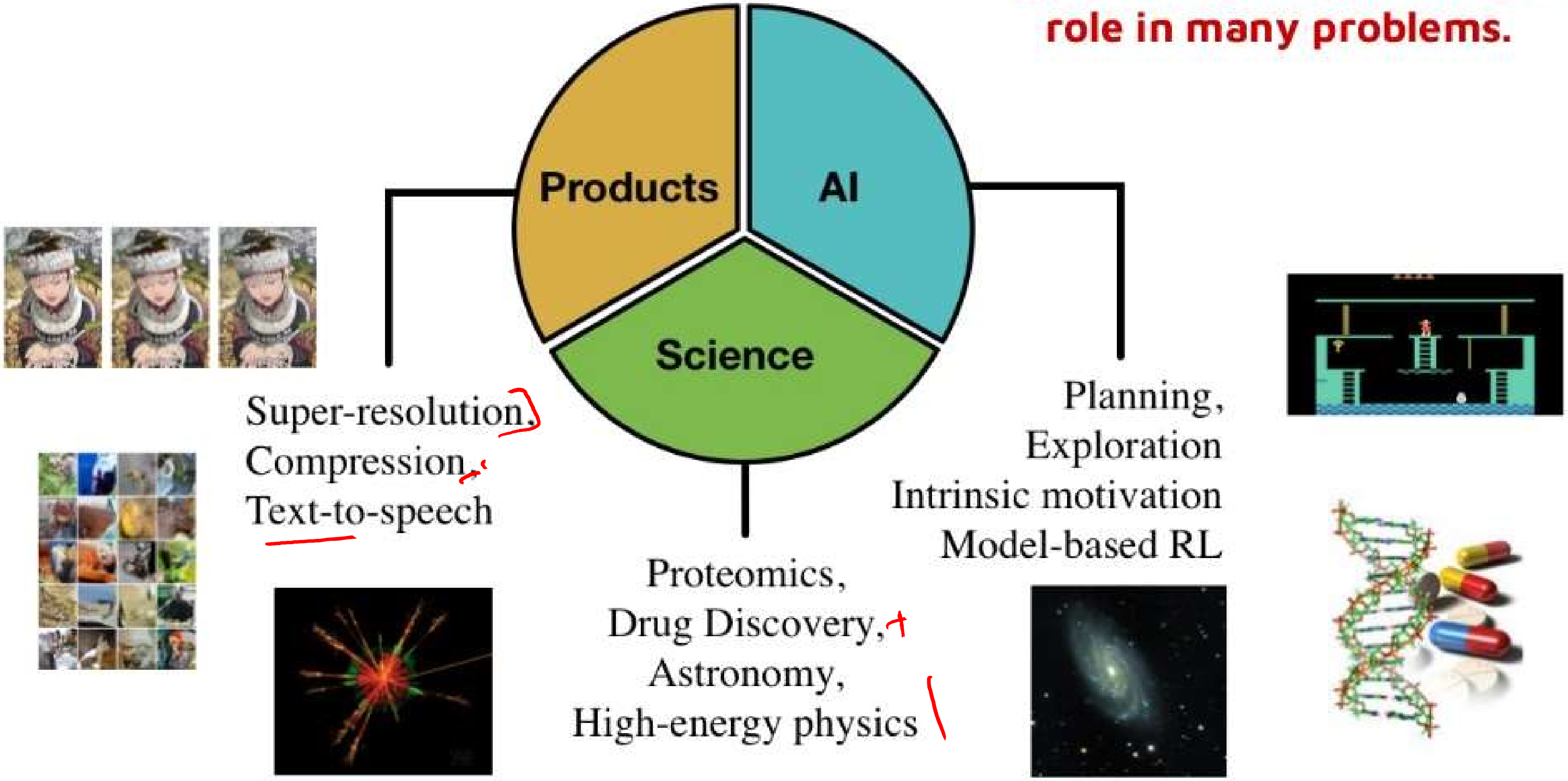
Generation

Given $D = \{x^1, x^2, ..., x^N\}$ learn a $P(x)$ from which it is easy to generate samples

Example: generate images, music, dialogue, text-to-speech, augment training data

Why Generative Models

Generative models have a role in many problems.



Super-resolution,
Compression,
Text-to-speech

Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

Planning,
Exploration
Intrinsic motivation
Model-based RL

Topics on Generation

- How to model $\boxed{P(x)}$: **Representation**
- How to learn parameters of $P_{\theta}(x)$ from training samples given D = $\{x^1, x^2, \ldots.,x^N\}$: **Learning/Estimation**
- How to generate examples from estimated model: **Sampling**
    - Generate examples that are representative of the distribution rather than the highly probability example in Prediction

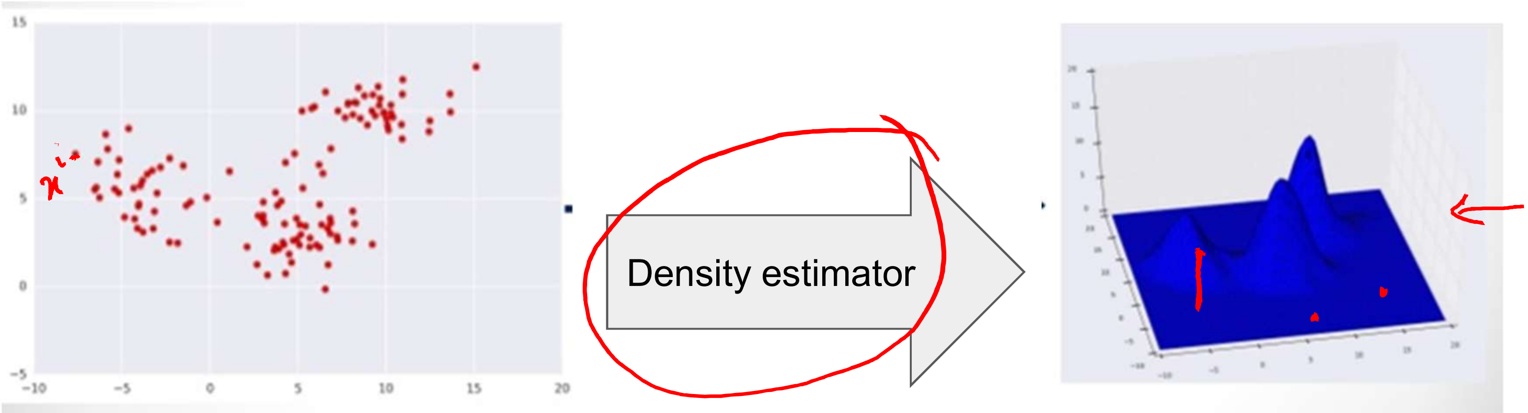
Different task settings

- Prediction:
  - Two types of variables: input and output. Output comprises of multiple variables.
- Generation: Two types
  - All variables are observed during training, during inference we need to generate new samples.
  - Variables are partitioned into an observed and hidden set. During training samples contain only observed variables. We need to reason about hidden variables to understand the data or generate more data
- Density estimation:
  - What is the probability that a given sample is part of the training distribution
- Other forms of reasoning:
  - Causality, Counter-factual reasoning, recourse on predictions.

Density estimation

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

Applications: Out of distribution detection, outlier detection, classification



Density estimator

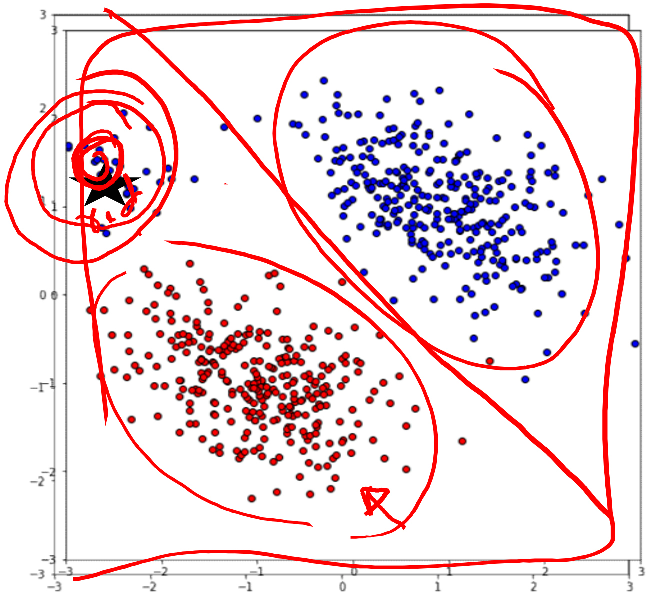
THE ART OF KNOWING IS KNOWING WHAT TO IGNORE.

RUMI

OOD detection

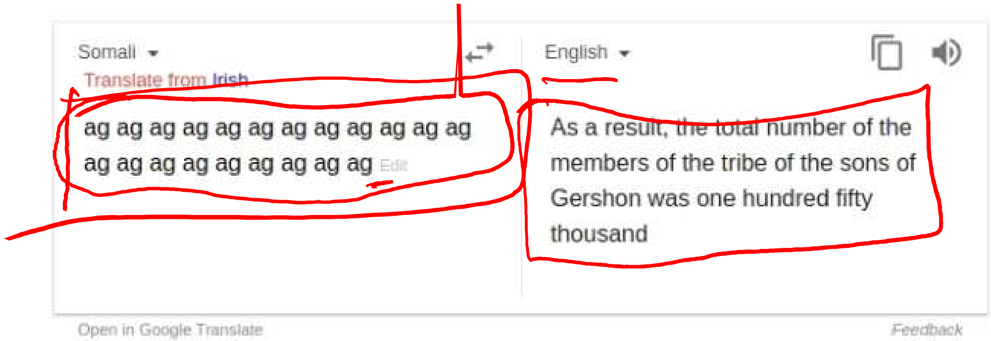
- OOD examples could be assigned wrong labels with high confidence.

- Ideally should model P(x) and reject examples which are outside P(x).
- Challenge: P(x) is hard to model for higher dimension data



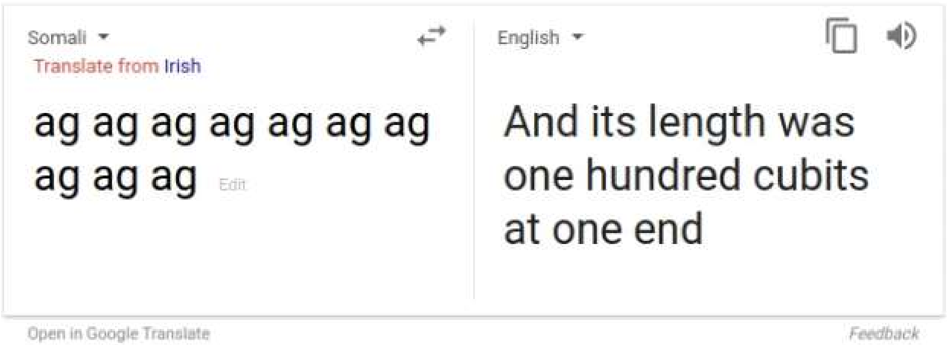
Importance of OOD detection

MOTHERBOARD



ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag

As a result, the total number of the members of the tribe of the sons of Gershon was one hundred fifty thousand



ag ag ag ag ag ag ag ag ag ag

And its length was one hundred cubits at one end

Different task settings

- Prediction:
  - Two types of variables: input and output. Output comprises of multiple variables.

- Generation: Two types
  - All variables are observed during training, during inference we need to generate new samples.
  - Variables are partitioned into an observed and hidden set. During training samples contain only observed variables. We need to reason about hidden variables to understand the data or generate more data

- Density estimation:
  - What is the probability that a given sample is part of the training distribution

- Other forms of reasoning:
  - Causality, Counter-factual reasoning, recourse on predictions.

What is causal inference?

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

Examples:

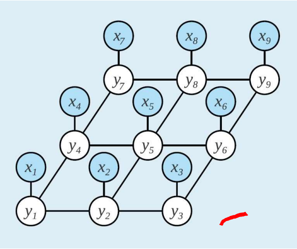
- Effect of treatment on a disease
- Effect of climate change policy on emissions
- Effect of social media on mental health
- Many more (effect of X on Y)
- Effect of Interest rate on inflation
- Effect of air pollution on cancer

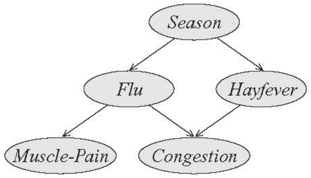
Counterfactual reasoning

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
  - Deep Learning models:
    - Any general network, Bayesian Neural networks, VAEs, GANs, Normative flows, Energy-based models

Deep networks have high capacity but unless you factorize the distribution well you might overfit, have an intractable inference algorithm, may not generalize to new scenarios, and may not be explainable.

Course contents

- Learning
    - Conditional Random Fields
    - Generalized Expectation Maximization
    - Advanced topics from deep learning:
        - Learning with discrete latent variables
        - Energy-based Models for structured learning
        - Meta-learning: many related distributions, e.g. speech recognition/NLP models for multiple languages

Course contents

Inference.

- Sum-product and max-product Inference in Graphical Models
  - Junction tree and Variational methods.
- 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
- 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 θ
- 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

- All classes online. Each week:
    - 1 hour lecture for each of Mon/Thur
    - 25 minutes problem solving that could be interleaved with the lectures or held separately
    - 20 minutes quiz, 5 minutes discussion of solution.
- SAFE/Moodle quiz on the material covered in the prior week
    - Quiz will be done in groups of size 2. A student must partner with 5 different students through the semester.
- All materials will be uploaded on moodle, announcements via Moodle, questions on moodle or cs726@googlegroups.com
- Course calendar https://www.cse.iitb.ac.in/~sunita/cs726/