Inference & Representation
Recitation #2

Today's Agenda:

1. Maximum Entropy Distribution
2. Intro to Graphs
3. Hidden Markov Models (HMM)

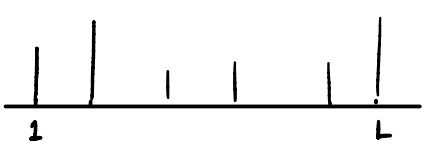
1.1 Variational Principle

$\mathcal{X}$: domain of the data

$\mathcal{P}(\mathcal{X})$: space of probability distribution over $\mathcal{X}$

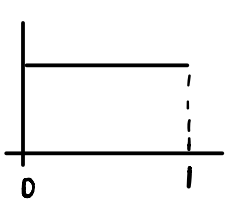
How to define a notion of smoothness for $p \in \mathcal{P}(\mathcal{X})$?

Ex. $\mathcal{X}$ discrete set $\mathcal{X} = \{1, \ldots, L\}$



→ Uniform distribution maximizes "smoothness"

Ex. $\mathcal{X} = [0,1]$



1.2 Entropy

We can quantity the smoothness of a probability distribution with its entropy:

$$\mathcal{H}_p = -\sum_{i=1}^{L} p_i \log p_i$$

and under appropriate assumption

$$\mathcal{H}_p = -\int_x p(x) \log p(x) dx \qquad \text{for } \underbrace{x \in \mathbb{R}^d}$$

Entropy measures uncertainty

1.3 Lagrange Multiplier

We can use entropy to characterize distributions from a given set of constraints

$$(*) \quad \max_{p \in P(\mathcal{X})} \mathcal{H}(p) \quad \text{subject to} \quad \begin{cases} \mathbb{E}_{x \sim p}\, x = \mu_0 \\ \mathbb{E}_{x \sim p}\, (x-\mu_0)(x-\mu_0)^T = \Sigma \end{cases}$$

Remark. This constrained optimization problem is not always well-defined.

It depends on the domain $\mathcal{X}$ as well as the constraints.

Ex. $\mathcal{X} = \mathbb{R}$

$$\max_{p:\mathbb{R}\to\mathbb{R}} -\int p(x)\log p(x)dx \iff \min_{p} \int p(x)\log p(x)dx$$

$$\text{s.t.} \begin{cases} p(x) \geq 0 & \forall x \\ \int p(x)dx = 1 \\ \int x\,p(x)dx = \mu \\ \int (x-\mu)^2 p(x)dx = \sigma^2 \end{cases}$$

→ Constrained optimization problem : we introduce **Lagrange multipliers** :

$$\mathcal{L}[p] = \int p(x)\log p(x)dx + \lambda\left(\int p(x)dx - 1\right) + \beta\left(\int x\,p(x)dx - \mu\right) + \gamma\left(\int (x-\mu)^2 p(x)dx - \sigma^2\right)$$

$$\frac{\partial \mathcal{L}}{\partial p(x)} = 0 = \log p(x) + 1 + \lambda + \beta x + \gamma (x-\mu)^2$$

$$\log p(x) = a - bx - cx^2$$

$$p(x) \propto e^{-bx-cx^2} \quad \text{with } b, c \text{ adjusted s.t.}$$

$$\mathbb{E}_p x = \mu, \quad \mathbb{E}_p (x-\mu)^2 = \sigma^2$$

Same thing happens in d-dimensions :

the distribution of max entropy with $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ is $\mathcal{N}(\mu, \Sigma)$

1.4 Some Exercises :

Let $\mathcal{X} = \{x_1, \cdots, x_N\}$ be a discrete set and let $p \in P(\mathcal{X})$ be the space of probability distributions defined over $\mathcal{X}$. Define the entropy :

$$\mathcal{H}(p) := -\sum_{i=1}^{n} p_i \log(p_i)$$

Ex1. We want to show its maximum is attained at uniform distribution, with maximum entropy $\log N$.

Lagrange Multipliers : ...

□

Now let us move from a discrete to continuous domain, by setting $\mathcal{X} = [0, R]$ Let $P(\mathcal{X})$ denote the space of probability distribution admitting a density $p(x)$, $x \in \mathcal{X}$, and define the Shannon entropy as

$$\mathcal{H}(p) = -\int_{\mathcal{X}} p(x) \log p(x) \, dx$$

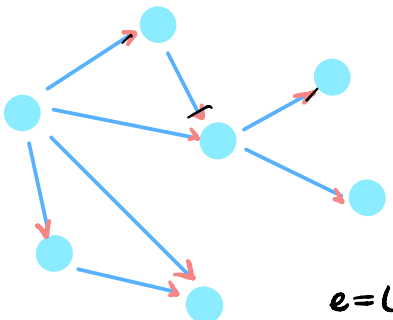
Ex2. Show that the maximum entropy distribution in $P(\mathcal{X})$ is the uniform measure in $\mathcal{X}$, with entropy $\log R$.

Similarly to discrete domain,

Lagrange multipliers, ...

□

2. Graph Theory



$G = (V, E)$

$V = \{v_1, \dots, v_n\}$ set of vertices

$E = \{e_1, \dots, e_m\}$ set of edges

$v_i \sim v_j$

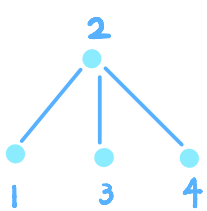
$e = (v_i, v_j) \in E$

Two types of graphs:

- generic : directed $E: \{(i,j),\ i,j \in V\}$ ($i \to j$ ordered)
- symmetric/ undirected : $(v_i, v_j) \in E \iff (v_j, v_i) \in E$

Adjacency matrix : $A_{ij} = \begin{cases} 1 & \text{if } v_i \to v_j \\ 0 & \text{o.w.} \end{cases}$, $A \in \{0,1\}^{n \times n}$

Ex.



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Weighted graph : graph with weights on edges

$W = (w_{ij}) \in \mathbb{R}^{n \times n}$

Some concepts in graph theory

* path $v_1, v_2, \ldots, v_k \in V : v_i \sim v_{i+1}, v_i \neq v_j$

* cycle path with $v_k = v_1$

* connected A subset of vertices is connected
  , if $\exists$ path between each pair of vertices.

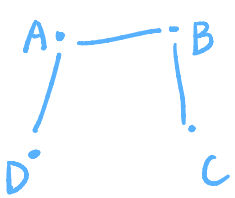
* connected component maximal connected subset

* tree symmetric + connected + acyclic

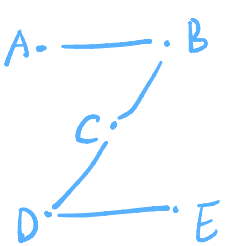
  (*) if G is a tree. $|E| = N-1$

* spanning tree subgraph $T \subseteq G : V(T) = V(G)$ + T is a tree.

Ex. Tree



Spanning Tree :



* Clique $H \subseteq G$: there is an edge between each pair of vertices in H

- maximal clique There is no other clique containing it.

- clique number of G $c(G) = \max_{\text{clique } C \subseteq G} |C|$

* Complementary notion

- independent set : no two nodes share an edge

  $\Leftrightarrow$ a clique of G^c

- independence number size of maximal independent set

  The independence number of G^c = clique # of G

---

3. Hidden Markov Model (HMM)

$$\begin{array}{ccccccccc} X_1 & \rightarrow & X_2 & \rightarrow & X_3 & \rightarrow & \cdots & \rightarrow & X_n \\ \downarrow & & \downarrow & & \downarrow & & & & \downarrow \\ Y_1 & & Y_2 & & Y_3 & & & & Y_n \end{array}$$

Hidden variables

Observed variables

Ex1. Hidden : events in the atmosphere
Observation : daily weather

Ex2. Part-of-Speech tagging

Y_i : words
X_i : tags (n. / verb. / adj)

Ex3. Speech recognition

Y_i : sounds, recordings, audio
X_i : words, letters

Densities of HMM

$$P(X,Y) = P(X_1) \prod_{i=2}^{n} P(X_i | X_{i-1}) \prod_{i=1}^{n} P(Y_i | X_i)$$

$P(X_i | X_{i-1})$: parent of X_i
$P(Y_i | X_i)$: parent of Y_i

Conditional Independence implied by HMM

$$X_i \perp X_j \mid X_{i-1} \qquad \forall j < i-1$$ ← Markov property

$$Y_i \perp \{X_1, \cdots X_{i-1}, X_{i+1}, \cdots, X_n, Y_1 \cdots Y_{i-1}, Y_{i+1}, \cdots Y_n\} \mid X_i$$

Why "Markov"? Future ⊥ Past | Present

Ex. Compute Probabilities in HMM

TBC...