LECTURE 1 · Intro - Course Overview

Molecular Biology Review

DNA is a polymer of : adenine (A), thymine (T), guanine (G), and cytosine (C).

DNA is replicated and recombined.

5' (upstream) 3' (downstream)

RNA polymerase reads the anti-sense strand (template) —produces→ coding (sense) strand.

mRNA → codes for proteins
tRNA → translation
rRNA → ribosome structure
long intergenic noncoding RNAs (lincRNAs) → regulation? (e.g. XIST lncRNA silences X chromosome in women).
miRNAs → bind to RNA (post-transcriptional regulation)
snRNAs → help guide chemical modification of other RNAs.

Genes begin at TSS. The coding sequences: 5' UTR → 3' UTR { exons, introns ← alternate splice isoforms

Regulatory regions { upstream TSS → promoter (proximal reg. regions)
distal reg. → enhancers → conformational DNA changes
↳ insulators

Epigenetics → { DNA methylation → CpG → repression
Histone modification → acetylation
⋮

Probability Review

Discrete random variable X can take multiple values from a sample space S.

$P(X=x)$ → Probability mass function: $P(x)$

$$0 \leq P(X=x) \leq 1$$

$$\sum_{x \in S} P(X=x) = \sum_{x} P(x) = 1$$

* $P(A,B) = P(A \cap B)$

Conditional Probability: $P(A|B) = \frac{P(A,B)}{P(B)}$

or $P(S|M)$ → S given M

* $P(M|S)$ → given S what is the prob that it came from M expressing it could also come from B.

Chain Rule: $P(A,B,C,\ldots) = P(A) \cdot P(B|A) \cdot P(C|A,B) \cdots$

Bayes's Rule: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Independence: $P(A|B) = P(A)$

Conditional independence: $P(A|B,C) = P(A|C)$

↳ A & B are only conditionally independent given C if of, given knowledge C occurs, knowledge of whether A occurs provides no info of whether B occurs and viceversa.

Inference problems:

LEARNING: updating posterior probabilities, we want to compute $P(x|y)$. From Bayes', we'll need $P(y|x)$. From chain rule we can factor joint distribution of x, y. Then use conditional independence to simplify.

EST. MOST LIKELY: maximum a posterior estimate $\longrightarrow \arg\max_x (P(x|y))$.

Probability distributions

Discrete: take specific countable values (e.g. rolling a dice $\{1,2,3,4,5,6\}$).

Continuous: take any value within a range

For any: $P(x) \geq 0$

Discrete: $\sum P(x) = 1$

Continuous: $\int P(x)\,dx = 1$

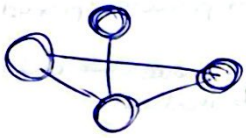
Gaussian distribution (Continuous)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ = mean
σ^2 = variance

Graphical Probabilistic Models (GPMs)

Nodes: random variable
Edge: dependencies / relationship
Structure: directed or undirected



Bayesian Networks

→ edges represent conditional dependencies (Directed acyclic graphs or DAGs)

→ encodes the joint probability

$$P(x_1, x_2, x_3, \dots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{Parents}(x_i))$$

E.g.: R(rain) → W(wet)
S(sprinkler) → W(wet)

$$P(R,S,W) = P(R) \cdot P(S) \cdot P(W|S,R)$$

Markov Networks

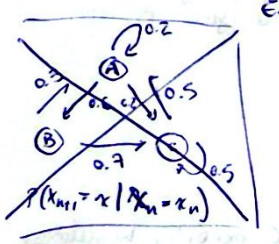
→ undirected; edges represent pairwise relationships 

→ encode joint probability distributions as products of potential functions over cliques (fully connected subsets of nodes)

$$P(x_1, x_2, \dots, x_n) \propto \prod_{\text{cliques}} \phi(c),$$

where $\phi(C)$ is the potential function for clique C

E.g.: social networks



$A \leftrightarrow B$, $B \leftrightarrow C$, $A \leftrightarrow C$

$$P(A,B,C) \propto \phi(A,B) \cdot \phi(B,C) \cdot \phi(A,C)$$

Normalization:

$$P(A,B,C) = \frac{1}{Z}\, \phi(A,B) \cdot \phi(B,C) \cdot \phi(A,C)$$

A & B tend to agree
B & C "
A & C are directly influenced

A	B	$\phi(A,B)$
0	0	0.8
1	0	0.2
0	1	0.2
1	1	0.8

Bayes' Rule

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E)}$$

- $P(H|E)$: probability of hypothesis after evidence — POSTERIOR
- $P(H)$: p. of hypothesis — PRIOR
- $P(E|H)$: how likely E is if hypothesis is true — LIKELIHOOD
- $P(E)$: ensures valid probability — NORMALIZATION

Ex:

$P(D) = 0.01$

$P(P|D) = 0.95$ (true positive)

$P(P|N) = 0.05$ (false positive)

$P(N) = 0.99$

$$P(D|P) = \frac{P(P|D)\cdot P(D)}{P(P|D)\cdot P(D) + P(P|N)\cdot P(N)} = \frac{0.95 \cdot 0.01}{0.95\cdot 0.01 + 0.05\cdot 0.99} = \boxed{0.16}$$

Markov Chains

Probability of the next state depends only on the current state:

$$P(X_{t+1}|X_t, X_{t-1}, \dots) = P(X_{t+1}|X_t)$$

Components

States: $S = \{s_1, s_2, s_3, \dots, s_n\}$

Transition prob (T): $T_{i,j} = P(X_{t+1} = s_j \mid X_t = s_i)$, $T_{ij} \geq 0$ and $\sum_j T_{ij} = 1$.

Ex:

s_1 = sunny

s_2 = rainy

$$T = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$$

T_{11} = if today sunny => 80% also sunny

T_{12} = if today sunny => 20% rainy.

T_{21} = if today rain => 40% sunny

T_{22} = if today rain => 60% rain

Hidden Markov Models (HMMs)

1. Hidden states (S): true underlying states we aim to ~~observe~~ infer.
2. observations (O): symbols emitted by S according to a probability dist.
3. Emission probabilities (E) likelihood of each O given S.

Assuming that the Markov property holds and observations are conditionally independent given S.

Example: detecting DNA units in a sequence

S { unit (U) / background (B) }, $O = \{A, C, T, G\}$,

$T_{U\to U} = 0.9$, $T_{U\to B} = 0.1$,

$E_U(A) = 0.4$, $E_U(C) = 0.3$, $E_U(G) = 0.2$, $E_U(T) = 0.1$

To infer hidden states we can use:

Forward-Backward Algorithm

Viterbi Algorithm

Maximum likelihood and Max A post Estimates

- **MLE**

  Find θ that max. likelihood

  $$\hat{\theta}_{MLE} = \arg\max_{\theta} P(data|\theta)$$

- **MAP**

  $$\hat{\theta}_{MAP} = \arg\max_{\theta} P(data|\theta)\, P(\theta)$$