CSC 412/2506: Probabilistic Learning and Reasoning

Week 8: Midterm review

Murat A. Erdogdu

University of Toronto

Midterm exam

- Exam will be held online on March 2nd, during lecture time.
- Students will be on the same zoom call during the exam; link to be shared via a quercus on the exam day. Pdf with the questions will be shared on the zoom call.
- Exams will be 100 points in total and 50 mins long, which includes the time you need to scan and upload your work to crowdmark. If you run into technical difficulties with crowdmark, you may submit your solutions to `csc412tas@cs.toronto.edu`.
- Students are required to take the exam with their enrolled sections (x = 12 or 13 depending on the section you are enrolled in).
  - Instructions x:00-x:10
  - Exam starts at x:10 and ends at (x+1):00.
  - The call ends at (x+1):00. 10 additional minutes are allowed for students to scan/upload any remaining work.
  - Late submissions after (x+1):10 will receive 2 points per late min penalty, strictly enforced with no exceptions.
- Exam covers all lectures (weeks 1-6), it is closed book/internet. You can use one optional A4 aid sheet - double-sided.
- You can produce your solutions any way you want (photo of handwriten solutions, ipad, latex etc). Latex is not recommended as it may take longer to type up. You do not have to print out the question pdf.

Overview of topics

- Exponential families formulation
- MLE derivations
- Decision theory
- Bayes nets: Implied conditional independence and factorization
- Markov Random Fields: Implied conditional independence and factorization
- Variable elimination: Complexity, order of elimination
- Message passing: Belief propagation, purpose, convergence properties on trees
- Sampling/MCMC: Sampling tools, how to use Simple Monte Carlo
- Variational Inference: Objectives, KL divergence, properties

Exponential families

- Density of a member of exponential families is of the form

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

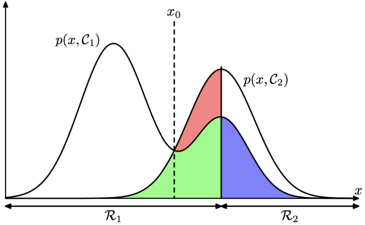
Here,

$$\begin{aligned} T(x) &: \text{Sufficient statistics} \\ \eta &: \text{Natural parameter} \\ A(\eta) &: \text{log-partition function} \\ h(x) &: \text{carrying density} \end{aligned}$$

- Examples of exponential familes are
  - Bernoulli, Gaussian, Gamma, exponential, Poisson etc.
  - defines a broad class of distributions
  - Moments of sufficient statistics can be found easily by differentiating the log-partition function.

Decision theory: Expected loss

- Minimizing the misclassification rate:



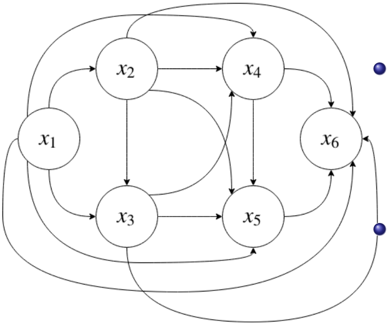
- We use a **loss function** to measure the loss incurred by taking any of the available decisions.
- Consider medical diagnosis example: example of a loss matrix:

Truth \ Decision	cancer	normal
cancer	0	1000
normal	1	0

Incorrectly classify as healthy (→ 1000)

Incorrectly classify as cancer (→ 1)

Directed Acyclic Graphical Models (Bayes' Nets)



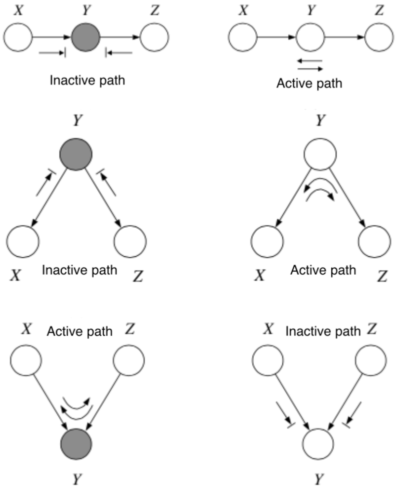
- A directed acyclic graphical model (DAGM) implies a factorization of the joint distribution.
- Variables are represented by nodes, and edges represent dependence.

DAGM induces the following factorization of the joint distribution of random variables $x_1, x_2, \ldots, x_N$, we can write:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}) = \prod_{i=1}^{N} p(x_i | \text{parents}(x_i))$$

where parents(x_i) is the set of nodes with edges pointing to x_i.

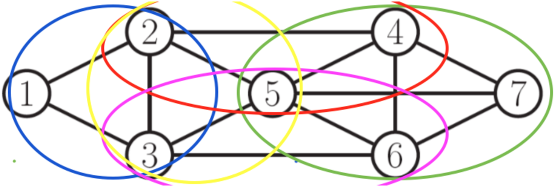
Bayes Ball: Rules for active/inactive triples



- Arrows: paths the balls can travel
- Arrows with bars: paths the balls cannot travel
- Notice balls can travel opposite to edge directions!

Markov Random Fields

- Markov random fields (MRFs), are a set of random variables where the dependencies are described by an undirected graph.



Lets see how to factorize the undirected graph of our running example:

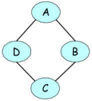
$$p(x) \propto \psi_{1,2,3}(x_1, x_2, x_3)\psi_{2,3,5}(x_2, x_3, x_5)\psi_{2,4,5}(x_2, x_4, x_5)\psi_{3,5,6}(x_3, x_5, x_6) \times \psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

Representing potentials

If the variables are discrete, we can represent the potential functions as tables of (non-negative) numbers

$$p(A, B, C, D) = \frac{1}{Z}\psi_{A,B}(A, B)\psi_{B,C}(B, C)\psi_{C,D}(C, D)\psi_{A,D}(A, D)$$

where



$\psi_{AB}[A, B]$			$\psi_{BC}[B, C]$			$\psi_{CD}[C, D]$			$\psi_{AD}[D, A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Note that these potentials are not probabilities, but instead encode relative affinities between the different assignments. For example, in the above table, a^0, b^0 is taken to be 30X more likely than a^1, b^0.

Variable elimination

Order which variables are marginalized affects the computational cost!

Main tool in exact inference is **variable elimination**:

- A simple and general **exact inference** algorithm in any probabilistic graphical model (DAGMs or MRFs).
- Has computational complexity that depends on the graph structure of the model.
- Sum-product is used to obtain marginals.

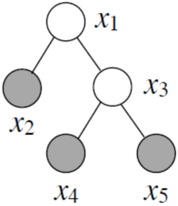
Complexity of Variable Elimination Ordering

- Different elimination orderings will involve different number of variables appearing inside each sum.
- The complexity of the VE algorithm is

$$O(mk^{N_{\max}})$$

  where
  - m is the number of initial factors.
  - k is the number of states each random variable takes (assumed to be equal here).
  - N_i is the number of random variables inside each sum $\sum_i$.
  - $N_{\max} = \max_i N_i$ is the number of variables inside the largest sum.

Inference in Trees



- Joint distribution is

$$p(x_{1:n}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \psi(x_i) \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j).$$

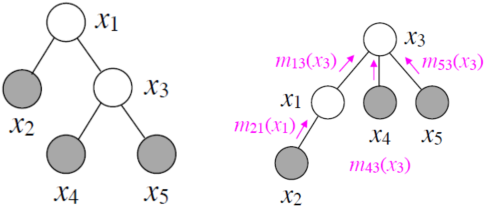
- Want to compute $p(x_3 | \bar{x}_2, \bar{x}_4, \bar{x}_5)$.
- We have

$$p(x_3 | \bar{x}_2, \bar{x}_4, \bar{x}_5) \propto p(x_3, \bar{x}_2, \bar{x}_4, \bar{x}_5).$$

$$p(x_3 \mid \bar{x}_2, \bar{x}_4, \bar{x}_5) = \frac{1}{Z^E} \sum_{x_1} \psi_1(x_1)\psi_3(x_3)\psi_2(\bar{x}_2)\psi_4(\bar{x}_4)\psi_5(\bar{x}_5)\psi_{12}(\bar{x}_2, x_1)\psi_{34}(\bar{x}_4, x_3)\psi_{35}(\bar{x}_5, x_3)\psi_{13}(x_1, x_3)$$

- Let's write the variable elimination algorithm.

Inference in Trees



$$
\begin{aligned}
p(x_3 \mid \bar{x}_2, \bar{x}_4, \bar{x}_5) &= \frac{1}{Z^E} \sum_{x_1} \psi_1(x_1)\psi_3(x_3)\psi_2(\bar{x}_2)\psi_4(\bar{x}_4)\psi_5(\bar{x}_5)\psi_{12}(\bar{x}_2, x_1)\psi_{34}(\bar{x}_4, x_3)\psi_{35}(\bar{x}_5, x_3)\psi_{13}(x_1, x_3) \\
&= \frac{1}{Z^E} \underbrace{\psi_4(\bar{x}_4)\psi_{34}(\bar{x}_4, x_3)}_{m_{43}(x_3)} \underbrace{\psi_5(\bar{x}_5)\psi_{35}(\bar{x}_5, x_3)}_{m_{53}(x_3)} \psi_3(x_3) \sum_{x_1} \psi_1(x_1)\psi_{13}(x_1, x_3) \underbrace{\psi_2(\bar{x}_2)\psi_{12}(\bar{x}_2, x_1)}_{m_{21}(x_1)} \\
&= \frac{1}{Z^E} \psi_3(x_3)m_{43}(x_3)m_{53}(x_3) \underbrace{\sum_{x_1} \psi_1(x_1)\psi_{13}(x_1, x_3)m_{21}(x_1)}_{m_{13}(x_3)} \\
&= \frac{1}{Z^E} \psi_3(x_3)m_{43}(x_3)m_{53}(x_3)m_{13}(x_3) = \frac{\psi_3(x_3)m_{43}(x_3)m_{53}(x_3)m_{13}(x_3)}{\sum_{x_3} \psi_3(x_3)m_{43}(x_3)m_{53}(x_3)m_{13}(x_3)}
\end{aligned}
$$

Slide credit: S. Ermon

Message Passing on Trees

- The message sent from variable j to $i \in N(j)$ is

$$m_{j \to i}(x_i) = \sum_{x_j} \psi_j(x_j)\psi_{ij}(x_i, x_j) \prod_{k \in N(j)/i} m_{k \to j}(x_j)$$

  - If x_j is observed, the message is

$$m_{j \to i}(x_i) = \psi_j(\bar{x}_j)\psi_{ij}(x_i, \bar{x}_j) \prod_{k \in N(j)/i} m_{k \to j}(\bar{x}_j)$$

- In trees, if the marginal we want to compute is chosen as the root node, a single pass from leaves to root is enough.
- To compute all marginals, two passes are needed: one from leaves to root, one from root to leaves.
- Once the message passing stage is complete, compute beliefs

$$b(x_i) \propto \psi_i(x_i) \prod_{j \in N(i)} m_{j \to i}(x_i).$$

- If it is not a tree, run loopy BP.

Sum-product vs. Max-product

- The algorithm we learned is called **sum-product BP** and approximately computes the **marginals** at each node.
- For MAP inference, we maximize over x_j instead of summing over them. This is called **max-product BP**.
- BP updates take the form

$$m_{j\to i}(x_i) = \max_{x_j} \psi_j(x_j)\psi_{ij}(x_i, x_j) \prod_{k\in N(j)\neq i} m_{k\to j}(x_j)$$

- MAP inference:

$$\hat{x}_i = \arg\max_{x_i} b(x_i).$$

Estimation method: Simple Monte Carlo

Estimation problem using simple Monte Carlo:

- **Simple Monte Carlo**: Given $\{x^{(r)}\}_{r=1}^{R} \sim p(x)$ we can estimate the expectation $\mathop{\mathbb{E}}_{x \sim p(x)}[\phi(x)]$ using the estimator $\hat{\Phi}$:

$$\Phi := \mathop{\mathbb{E}}_{x \sim p(x)}[\phi(x)] \approx \frac{1}{R}\sum_{r=1}^{R} \phi(x^{(r)}) := \hat{\Phi}$$

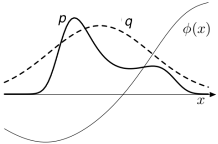
- The fact that $\hat{\Phi}$ is a consistent estimator of Φ follows from the Law of Large Numbers (LLN).
- Easy to design estimators using simple Monte Carlo, e.g. practice midterm.

Estimation tool: Importance Sampling

- Target $p(x)$ can be evaluated up to normalizing constant $\tilde{p}(x)$
- There is a simpler density, $q(x)$ from which it is easy to sample from and can evaluate up to normalizing constant $\tilde{q}(x)$

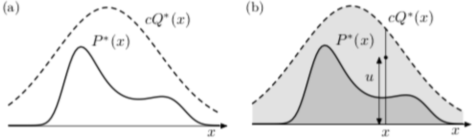
$$\text{Sample:} \quad x^{(r)} \sim q(x) = \tilde{q}(x)/Z_q$$

Importance sampling: estimate the expectation of a function $\phi(x)$.



- Introduce weights: $\tilde{w}_r = \frac{\tilde{p}(x^{(r)})}{\tilde{q}(x^{(r)})}$
- The importance weighted estimator $\hat{\Phi}_{iw} = \sum_{r=1}^{R} \phi(x^{(r)}) \cdot w_r$ where $w_r = \frac{\tilde{w}_r}{\sum_{r=1}^{R} \tilde{w}_r}$

Sampling tool: Rejection sampling



The procedure is as follows:

1. Generate two random numbers.
   1.1 The first, x, is generated from the proposal density $q(x)$.
   1.2 The second, u is generated uniformly from the interval $[0, c\tilde{q}(x)]$ (see figure (b) above: book's notation $P^* = \tilde{p}$, $Q^* = \tilde{q}$).
2. Accept or reject the sample x by comparing the value of u with the value of $\tilde{p}(x)$
   2.1 If $u > \tilde{p}(x)$, then x is rejected
   2.2 Otherwise x is accepted; x is added to our set of samples $\{x^{(r)}\}$ and the value of u discarded.

Hidden Markov Models

- Important DAGMs to simplify the joint distribution.
- Posterior inference takes the special form:

$$\begin{aligned} p(z_t|x_{1:T}) &\propto p(z_t, x_{1:t})p(x_{t+1:T}|z_t) \\ &\propto (\text{Forward Recursion})(\text{Backward Recursion}) \end{aligned}$$

- **Forward-backward algorithm** to compute $p(z_t|x_{1:T})$
- **Viterbi algorithm** to compute the most probable sequence.

$$\hat{z} = \arg\max_{z_{1:T}} \; p(z_{1:T}|x_{1:T})$$

Variational Inference: KL divergence

We will measure the difference between q and p using the **Kullback-Leibler divergence**

$$KL(q(z)||p(z)) = \int q(z) \log \frac{q(z)}{p(z)} dz$$
$$\text{or} = \sum_z q(z) \log \frac{q(z)}{p(z)}$$

Properties of the KL Divergence

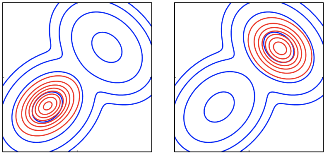
- $KL(q||p) \geq 0$
- $KL(q||p) = 0 \Leftrightarrow q = p$
- $KL(q||p) \neq KL(p||q)$
- KL divergence is not a metric, since it's not symmetric

Information (I-)Projection:

I-projection: $q^* = \arg\min_{q \in Q} KL(q||p) = \mathbb{E}_{x \sim q(x)} \log \frac{q(x)}{p(x)}$:

- $p \approx q \implies KL(q||p)$ small
- I-projection underestimates support, and does not yield the correct moments.
- $KL(q||p)$ penalizes q having mass where p has none.

$p(x)$ is mixture of two 2D Gaussians and Q is the set of all 2D Gaussian distributions (with arbitrary covariance matrices)



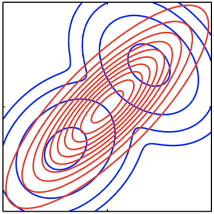
p=Blue, q^*=Red (two equivalently good solutions!)

Moment (M-)projection

M-projection: $q^* = \arg\min_{q \in Q} KL(p||q) = \mathbb{E}_{x \sim p(x)} \log \frac{p(x)}{q(x)}$:

- $p \approx q \implies KL(p||q)$ small
- $KL(p||q)$ penalizes q missing mass where p has some.
- M-projection yields a distribution $q(x)$ with the correct mean and covariance.

$p(x)$ is mixture of two 2D Gaussians and Q is the set of all 2D Gaussian distributions (with arbitrary covariance matrices)



p=Blue, q^*=Red

Summary

- Review lectures.
- Solve the practice midterm.
- Good luck!