Machine Learning for Data Science (CS 4786)

Lecture 23: Approximate Inference

The text in black outlines main ideas to retain from the lecture. The text in blue give a deeper understanding of how we "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested. But is not crucial for understanding the basic workings of the method.

1 Approximate Inference

Variable elimination algorithm is an exact inference technique but can be computationally very expensive. Belief propagation algorithm is guaranteed to work only on trees. Belief propagation on trees can be extended to general graphs by something called the junction tree algorithm. That is an algorithm where we form a tree-like structure out of the graphical model by a process called tree decomposition. But, the bottom line is that exact inference is known to be provably computationally hard! So naturally the intuitive thing to do is to perform approximate inference. In approximate inference, we aim at approximately computing the probabilities we are aiming to infer. The approximation could be additive in that we aim to get an estimate $\hat{P}$ such that,

$$\left|\widehat{P}(\text{Latent}|\text{Observed}) - P(\text{Latent}|\text{Observed})\right| \leq \epsilon$$

for some small $\epsilon > 0$. Or the approximation can be multiplicative:

$$(1-\epsilon)P(\text{Latent}|\text{Observed}) \leq \widehat{P}(\text{Latent}|\text{Observed}) \leq (1+\epsilon)P(\text{Latent}|\text{Observed})$$

Broadly, there are (at least) two strategies for approximate inference. The first is a sampling based approach where we approximately infer probabilities by sampling from the desired distribution and taking some form of empirical estimate to approximate the true probabilities. The next approach is a more functional form where we approximate the probability to be inferred by a simpler distributions that are progressively easier compute. A prototypical methods that uses the second approach are variational approaches.

2 Inference via sampling

Before we proceed to inference via sampling in general, we start with a simple observation that for Bayesian networks, sampling from the joint distribution can be done efficiently given the parameters. How?

Well note that Bayesian networks are generative models where we are given the distribution of each variable given its parents. Hence we can topologically sort variables and then starting from parents going down the network, we can one by one sample first parent variables from marginal,

then each child given its parent and so on. This way, since the joint distribution factorizes over the graph as $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\text{Parent}(X_i))$ we can easily obtain samples from the joint distribution.

2.1 Rejection Sampling

Now lets move to the more interesting question of inference in general. Say we wanted to infer the probability

$$P(X_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, X_{\text{Latent}_m} = x_{\text{Latent}_1} | X_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X_{\text{Observed}_n} = x_{\text{Observed}_n})$$

The idea behind rejection sampling is simple: since we know how to sample from joint distribution efficiently, we draw multiple samples from the joint distribution. We retain only those samples that match observations and throw away the remaining. Finally we approximate $P(X_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, X_{\text{Latent}_m} = x_{\text{Latent}_m} | X_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X_{\text{Observed}_n} = x_{\text{Observed}_n})$ as

$$\begin{aligned}
&\widehat{P}(X_{\text{Latent}_1} = x_{\text{Latent}_m}, \ldots, X_{\text{Latent}_m} = x_{\text{Latent}_m} | X_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X_{\text{Observed}_n} = x_{\text{Observed}_n}) \\
&= \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{\tilde{X}^r_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, \tilde{X}^r_{\text{Latent}_m} = x_{\text{Latent}_m}\}
\end{aligned}$$

where R is the total number of samples after rejecting ones that don't match observation and $\tilde{X}^1, \ldots, \tilde{X}^R$ are used to represent the non-rejected samples, that is the subset of samples $X^1, \ldots, X^T$ that match observation.

Why does this work? Note that by drawing from the joint distribution and rejecting, we are basically only retaining those samples such that match observation. Imagine we drew a very large number of samples T from the joint distribution. Out of this, R, is the number of ones that matched the observation. Hence R/T is an approximation for $P(\text{Observation})$. Also note that total number of samples out of the T samples that both match the observations and are such that $X_{\text{Latent}_1} = x_{\text{Latent}_m}, \ldots, X_{\text{Latent}_m} = x_{\text{Latent}_m}$ is the quantity $\mathbf{1}\{X^r_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, X^r_{\text{Latent}_m}$. This quantity approximates $P(\text{Observation}, \text{Latent})$. Since

$$P(\text{Latent}|\text{Observation}) = \frac{P(\text{Observation}, \text{Latent})}{P(\text{Observation})}$$

we can conclude that

$$\begin{aligned}
&\frac{\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{X^t_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, X^t_{\text{Latent}_m = x_{\text{Latent}_m}, X^t_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X^t_{\text{Observed}_n} = x_{\text{Observed}_n}}\}}{\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{X^t_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X^t_{\text{Observed}_n} = x_{\text{Observed}_n}\}} \\
&= \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{\tilde{X}^r_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, \tilde{X}^r_{\text{Latent}_m} = x_{\text{Latent}_m}\}
\end{aligned}$$

Drawback at a high level: The drawback at a high level for this approach is that we could reject a lot of samples. Especially is we have multiple variables that are observed, then since most draws from joint distribution wont match observation we reject most of the total T number of draws.

Why it works (detailed): Let us try to understand how well this approach works. Note that R/T is the empirical estimate of $P(\text{Observations})$. From law of large numbers we can conclude that with high probability over samples, R/T is close to $P(\text{Observations})$. Specifically, via Hoeffding/Bernstein inequality we can in fact conclude that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\left| P(\text{Observations}) - \frac{R}{T} \right| \leq \sqrt{\frac{2 \log\left(\frac{2}{\delta}\right)}{T}}$$

Similarly, we also simultaneously have that with probability at least $1 - \delta$

$$\left| P(\text{Observations}, \text{Latent}) - \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{X^t_{\text{Latent}} = x_{\text{Latent}}, X^t_{\text{Observed}} = x_{\text{Observed}}\} \right| \leq \sqrt{\frac{2 \log\left(\frac{2}{\delta}\right)}{T}}$$

where in the above I have used $X^t_{\text{Latent}} = x_{\text{Latent}}$ to mean $X^t_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, X^t_{\text{Latent}_m} = x_{\text{Latent}_m}$ and similarly $X^t_{\text{Observed}} = x_{\text{Observed}}$ to mean $X^t_{\text{Observed}_1} = x_{\text{Observed}_1}, \ldots, X^t_{\text{Observed}_n} = x_{\text{Observed}_n}$.
Now based on these two inequalities we see that,

$$\frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{\tilde{X}^r_{\text{Latent}_1} = x_{\text{Latent}_1}, \ldots, \tilde{X}^r_{\text{Latent}_m} = x_{\text{Latent}_m}\} = \frac{\frac{1}{T} \sum_{t=1}^{T} \sum_{t=1}^{T} \mathbf{1}\{X^t_{\text{Latent}} = x_{\text{Latent}}, X^t_{\text{Observed}} = x_{\text{Observed}}\}}{\frac{R}{T}}$$

and so,

$$\frac{P(\text{Observations}, \text{Latent}) - \sqrt{\frac{2\log(2/\delta)}{T}}}{P(\text{Observations}) + \sqrt{\frac{2\log(2/\delta)}{T}}} \leq \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\{\tilde{X}^r_{\text{Latent}} = x_{\text{Latent}}\} \leq \frac{P(\text{Observations}, \text{Latent}) + \sqrt{\frac{2\log(2/\delta)}{T}}}{P(\text{Observations}) - \sqrt{\frac{2\log(2/\delta)}{T}}}$$

Notice that this first of all means that as T becomes larger and larger we get arbitrarily close to $P(\text{Observations}, \text{Latent})/P(\text{Observations}) = P(\text{Latent}|\text{Observations})$. More specifically, notice that if T is large enough so that $\sqrt{\frac{2\log(2/\delta)}{T}} \leq \epsilon P(\text{Observations}, \text{Latent})$, that is

$$T > \frac{2\log(2/\delta)}{(\epsilon P(\text{Observations}, \text{Latent}))^2}$$

then we can conclude that

$$(1 - \epsilon) P(\text{Latent}|\text{Observation}) \leq \widehat{P}(\text{Latent}|\text{Observation}) \leq (1 + \epsilon) P(\text{Latent}|\text{Observation})$$

Clearly we have the issue that number of samples T we need grows inversely with $P(\text{Observations}, \text{Latent}))^2$ which when small is problematic.

2.2 Importance Sampling

The idea behind importance sampling or importance weighted estimates can be seen fairly generally beyond just for inference. Say we want to estimate for some function $f(x)$ the expected value under draws from distribution P. Computing such expectations for arbitrary functions is often hard. If we were able to sample from P, then one way to approximate the expectation could be to simply

draw multiple samples from P and take the average value of f under the samples drawn. However, what if we didn't know how to sample from P? Say we have access or can compute $P(X = x)$ easily but we cannot compute expectation of f under P (since integral of arbitrary function f under P can be hard). But say we do know some distribution Q from which both we know how to sample from and can compute $Q(X = x)$ easily. In this case, note that

$$\mathbb{E}_{X\sim P}[f(X)] = \sum_x P(X = x)f(x) = \sum_x Q(X = x)\frac{P(X = x)}{Q(X = x)}f(x) = \mathbb{E}_{X\sim Q}\left[\frac{P(X)}{Q(X)}f(X)\right]$$

Thus we see that expected value of f under P can be seen as expected value of $\frac{P}{Q}f$ under Q. Now we can simply approximate this expectation by sampling from Q and taking average under the sample of function $\frac{P}{Q}f$. That is, specifically,

$$\mathbb{E}_{X\sim P}[f(X)] \approx \frac{1}{T}\sum_{t=1}^{T}\frac{P(X = X^t)}{Q(X = X^t)}f(X^t)$$

Now coming back to our inference problem, note that

$$P(X_{\text{Latent}} = x_{\text{Latent}}|X_{\text{Observed}} = x_{\text{observed}}) = \mathbb{E}_{X_{\text{Latent}}\sim P(\text{Latent}|X_{\text{Observed}}=x_{\text{observed}})}\left[\mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}\right]$$

Hence we can view function $f(X_{\text{Latent}}) = \mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}$ as the function we are interested in. The target distribution we want to get expectation over as $P(\text{Latent}|X_{\text{Observed}} = x_{\text{observed}})$.

Now for inference problem, the distribution we instead sample from is:

$$\begin{aligned}
&\propto \prod_{i=1}^{m} P(X_{\text{Latent}_i}|\text{Parent}(X_{\text{Latent}_i})) \\
&= \frac{\prod_{i=1}^{m} P(X_{\text{Latent}_i}|\text{Parent}(X_{\text{Latent}_i}))}{\sum_{\text{value}-\text{of}-\text{latent}}\prod_{i=1}^{m} P(X_{\text{Latent}_i}|\text{Parent}(X_{\text{Latent}_i}))}
\end{aligned}$$

where for $\text{Parent}(X_{\text{Latent}_i})$, if a parent is unobserved we use the value sampled earlier (in the topological sort order) or if parent is observed, then we use the observed value. Hence notice that

the weight

$$
\begin{aligned}
&\frac{P(X^t_{\text{Latent}} = x_{\text{Latent}})}{Q(X^t_{\text{Latent}} = x_{\text{Latent}})} \\
&= \frac{P(X^t_{\text{Latent}} = x_{\text{Latent}} | X^t_{\text{Observed}} = x_{\text{observed}})}{\prod_{i=1}^m P(X_{\text{Latent}_i} = x_{\text{Latent}_i} | \text{Parent}(X_{\text{Latent}_i}))} \times \sum_{\text{value}-\text{of}-\text{latent}} \prod_{i=1}^m P(X_{\text{Latent}_i} | \text{Parent}(X_{\text{Latent}_i})) \\
&= \frac{P(X^t_{\text{Latent}} = x_{\text{Latent}}, X^t_{\text{Observed}} = x_{\text{observed}})}{P(X^t_{\text{Observed}} = x_{\text{observed}}) \prod_{i=1}^m P(X^t_{\text{Latent}_i} = x_{\text{Latent}_i} | \text{Parent}(X^t_{\text{Latent}_i}))} \\
&\qquad \times \sum_{\text{value}-\text{of}-\text{latent}} \prod_{i=1}^m P(X_{\text{Latent}_i} | \text{Parent}(X_{\text{Latent}_i})) \\
&= \frac{\prod_{i=1}^m P(X^t_{\text{Latent}_i} = x_{\text{Latent}_i} | \text{Parent}(X^t_{\text{Latent}_i} = x_{\text{Latent}_i})) \times \prod_{j=1}^n P(X^t_{\text{Observed}_j} = x_{\text{Observed}_j} | \text{Parent}(X^t_{\text{Observed}_j}))}{P(X^t_{\text{Observed}} = x_{\text{observed}}) \prod_{i=1}^m P(X^t_{\text{Latent}_i} = x_{\text{Latent}_i} | \text{Parent}(X^t_{\text{Latent}_i}))} \\
&\qquad \times \sum_{\text{value}-\text{of}-\text{latent}} \prod_{i=1}^m P(X_{\text{Latent}_i} | \text{Parent}(X_{\text{Latent}_i})) \\
&= \prod_{j=1}^n P(X^t_{\text{Observed}_j} = x_{\text{Observed}_j} | \text{Parent}(X^t_{\text{Observed}_j})) \times \frac{\sum_{\text{value}-\text{of}-\text{latent}} \prod_{i=1}^m P(X_{\text{Latent}_i} | \text{Parent}(X_{\text{Latent}_i}))}{P(X^t_{\text{Observed}} = x_{\text{observed}})} \\
&\propto \prod_{j=1}^n P(X^t_{\text{Observed}_j} = x_{\text{Observed}_j} | \text{Parent}(X^t_{\text{Observed}_j})) = w^t
\end{aligned}
$$

Thus we see that

$$w^t \propto \frac{P(X^t_{\text{Latent}} = x_{\text{Latent}})}{Q(X^t_{\text{Latent}} = x_{\text{Latent}})}$$

So for instance, say $\frac{P(X^t_{\text{Latent}} = x_{\text{Latent}})}{Q(X^t_{\text{Latent}} = x_{\text{Latent}})} = \frac{w^t}{C}$ where C is the normalizing constant. We have that,

$$1 = \mathbb{E}_{X^t_{\text{Latent}} \sim P}[1] = \mathbb{E}_{X^t_{\text{Latent}} \sim Q}\left[\frac{P(X^t_{\text{Latent}})}{Q(X^t_{\text{Latent}})}\right] = \mathbb{E}_{X^t_{\text{Latent}} \sim Q}\left[\frac{w^t}{C}\right] = \frac{\mathbb{E}_{X^t_{\text{Latent}} \sim Q}[w^t]}{C}$$

Thus we see that $C = \mathbb{E}_{X^t_{\text{Latent}} \sim Q}[w^t]$ and so,

$$\frac{P(X^t_{\text{Latent}} = x_{\text{Latent}})}{Q(X^t_{\text{Latent}} = x_{\text{Latent}})} = \frac{w^t}{\mathbb{E}_{X^t_{\text{Latent}} \sim Q}[w^t]}$$

where $w^t = \prod_{j=1}^n P(X^t_{\text{Observed}_j} = x_{\text{Observed}_j} | \text{Parent}(X^t_{\text{Observed}_j}))$. Thus, our estimate for under importance sampling with $f(X_{\text{Latent}}) = \mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}$ is given by

$$
\begin{aligned}
\tilde{P}(X_{\text{Latent}} = x_{\text{Latent}} | \text{Obervations}) &= \frac{1}{T} \sum_{t=1}^T \frac{w^t}{\mathbb{E}_{X_{\text{Latent}} \sim Q}[w(X_{\text{Latent}})]} \mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\} \\
&= \frac{1}{\mathbb{E}_{X_{\text{Latent}} \sim Q}[w(X_{\text{Latent}})]} \frac{1}{T} \sum_{t=1}^T w^t \mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}
\end{aligned}
$$

Now note that we can use sampling again to approximate $\mathbb{E}_{X_{\text{Latent}} \sim Q}[w(X_{\text{Latent}})]$ as average over samples drawn from Q as $\frac{1}{T} \sum_{t=1}^T w^t$. Thus our final estimate is given by

$$\widehat{P}(X_{\text{Latent}} = x_{\text{Latent}} | \text{Obervations}) = \frac{\sum_{t=1}^T w^t \mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}}{\sum_{t=1}^T w^t}$$

Why importance sampling works (detailed version): To see that the importance sampling estimate works we will first show that $P(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$ is close to $\tilde{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$ and then we will show that $\tilde{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$ is close to $\widehat{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$. To this end, note that

$$\begin{aligned} P(X_{\text{Latent}} = x_{\text{Latent}}|\text{Observation}) &= \mathbb{E}_{X_{\text{Latent}}\sim Q}\left[\frac{w(X_{\text{Latent}})}{\mathbb{E}_{X_{\text{Latent}}\sim Q}[w(X_{\text{Latent}})]}\mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}\right] \\ &= \frac{1}{\mathbb{E}_{X_{\text{Latent}}\sim Q}[w(X_{\text{Latent}})]}\mathbb{E}_{X_{\text{Latent}}\sim Q}\left[w(X_{\text{Latent}})\mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}\right] \end{aligned}$$

Now note that by law of large numbers we have that $\mathbb{E}_{X_{\text{Latent}}\sim Q}\left[w(X_{\text{Latent}})\mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}\right]$ is with high probability over sample close to $\frac{1}{T}\sum_{t=1}^{T} w^t\mathbf{1}\{X^t_{\text{Latent}} = x_{\text{Latent}}\}$. Specifically, using Hoeffding inequality we can conclude that for any δ with probability $1-\delta$,

$$\left|\frac{1}{T}\sum_{t=1}^{T} w^t\mathbf{1}\{X^t_{\text{Latent}} = x_{\text{Latent}}\} - \mathbb{E}_{X_{\text{Latent}}\sim Q}\left[w(X_{\text{Latent}})\mathbf{1}\{X_{\text{Latent}} = x_{\text{Latent}}\}\right]\right| \leq \sqrt{\frac{2\log(2/\delta)}{T}}$$

Hence we have that with probability $1-\delta$,

$$\left|\tilde{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations}) - P(X_{\text{Latent}} = x_{\text{Latent}}|\text{Observation})\right| \leq \sqrt{\frac{2\log(2/\delta)}{T}}$$

Hence, picking $T > 2\log(2/\delta)/\epsilon^2$ we get additive approximation of ϵ and picking $T > 2\log(2/\delta)/\epsilon^2(P(X_{\text{Latent}} = x_{\text{Latent}}|\text{Observation}))^2$ we get the multiplicative approximation. Now to show that $\tilde{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$ is close to $\widehat{P}(X_{\text{Latent}} = x_{\text{Latent}}|\text{Obervations})$. We note that $\mathbb{E}_{X_{\text{Latent}}\sim Q}[w(X_{\text{Latent}})]$ is close to $\frac{1}{T}\sum_{t=1}^{T} w^t$ by Hoeffding bound again. and this way we can show that our final estimate $\widehat{P}$ is close to our target inference probability.

The analysis here does not show the difference between number of samples required for rejection versus importance sampling. To understand this difference mathematically, one has to look more closely at the variance of the two estimates.