UMC203: AI and ML

Naman Mishra

January 2024

Contents

0.1 Restricted Boltzmann Machines

Lecture 14.
Tue 26 Mar '24

$$\Pr(S = s) = \frac{e^{-E(s)/T}}{Z}$$

where T is the temperature. We will fix $T = 1$.

$$\Pr(S = s) = \frac{e^{-E(s)}}{Z}$$

Suppose $w_{ii} = 0$ and (w_{ij}) is symmetric.

$$\begin{aligned} E(s) &= -\frac{1}{2}\sum_{i,j} w_{ij}s_i s_j - \sum_i b_i s_i \\ &= \sum_i s_i\Big(\sum_{j\geq i} w_{ij}s_j + b_i\Big) \\ &= \sum_i s_i\Big(\sum_{j>i} w_{ij}s_j + b_i\Big) \end{aligned}$$

since $w_{ii} = 0$. Suppose further that $w_{12} = 0$. Then

$$E(s) = s_1\Big(\sum_{j>2} w_{1j}s_j + b_1\Big) + s_2\Big(\sum_{j>2} w_{2j}s_j + b_2\Big) + K$$

where K depends only on $s_3, \ldots, s_n$. Thus conditioned on $S_{3:n}$, S_1 and S_2 are conditionally independent.

0.1.1 A real life example

You feel sick. You go to the doctor. The doctor asks you a series of questions, perhaps about the weather, your kids, your job, your symptoms. The doctor then diagnoses you. The doctor is a restricted Boltzmann machine?

The doctor has a knowledge base

$$\Pr(S = s \mid D_1 = d_1, \ldots, D_m = d_m).$$

They figure out the inverse of this,

$$\Pr(D = d \mid S = s)$$

using some Bayesion wizardry.

Now that we have touched some grass, let's go back to the math.

Split $S = (V, H)$ where V are the "visible" symptoms and H are the "hidden" symptoms.

$$V = \begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix}, \quad H = \begin{pmatrix} S_{m+1} \\ \vdots \\ S_d \end{pmatrix}$$

Let

$$\mathcal{D} = \left\{v^{(1)}, \ldots, v^{(N)}\right\}$$

We apply the latent variable model.

$$\log \Pr(V = v) = \log \sum_h \Pr(V = v, H = h)$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr\big(V = v^{(i)}\big)$$

and we employ the algorithm

$$\theta \leftarrow \theta + \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

where η is the learning rate. This is called *gradient ascent*.

Consider the baby case $N = 1$.

$$\begin{aligned}
\mathcal{L} &= \log \sum_h \Pr(V = v, H = h) \\
&= \log \sum_h e^{-E(s)} - \log Z \\
\implies \frac{\partial \mathcal{L}}{\partial w_{ij}} &= \frac{1}{\sum_h e^{-E(s)}} \sum_h e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial w_{ij}} \\
&= \frac{1}{\sum_h e^{-E(s)}} \sum_h e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} + \frac{1}{Z} \sum_s e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} \\
&= \sum_h \frac{e^{-E(s)}}{\sum_h e^{-E(s)}} s_i s_j - \sum_s \frac{e^{-E(s)}}{Z} s_i s_j \\
&= \sum_h \Pr(H = h \mid V = v) s_i s_j - \sum_s \Pr(S = s) s_i s_j \\
&= \mathop{\mathbf{E}}_{H|V}[s_i s_j] - \mathop{\mathbf{E}}_{S}[s_i s_j]
\end{aligned}$$

Chapter 1

Maximum Likelihood Estimation

Lecture 18. Tuesday 26 Mar '24

Definition 1.0.1. Let $X_1, X_2, \ldots$ be i.i.d. random variables drawn from a distribution $\Pr_\theta$, where $\theta \in \Theta$. The *likelihood function* is defined as

$$L_n(\theta) = \prod_{i=1}^{n} \Pr_\theta(X_i),$$

which of course motivates the *log-likelihood function*

$$\ell_n(\theta) = \sum_{i=1}^{n} \log \Pr_\theta(X_i).$$

The *maximum likelihood estimator* (MLE) is defined as

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \ell_n(\theta).$$

Definition 1.0.2 (KL divergence)**.** The *Kullback-Leibler divergence* of the distribution P from the distribution Q is defined as

$$D_{KL}(P \parallel Q) = \mathop{\mathbf{E}}_{P}\left[\log \frac{P(X)}{Q(X)}\right].$$

For discrete distributions, this is

$$D_{KL}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}.$$

Lemma 1.0.3. *For all $x \in \mathbb{R}_+$,*

$$\log x \leq x - 1$$

Proof. log is convex down, so the tangent at $x = 1$ always lies above the curve. □

Proposition 1.0.4. *For all distributions P and Q,*

$$D_{KL}(P \parallel Q) \geq 0.$$

Proof.

$$\begin{aligned} -D_{KL}(P \parallel Q) &= \mathop{\mathbf{E}}_{P}\left[\log \frac{Q(X)}{P(X)}\right] \\ &\leq \mathop{\mathbf{E}}_{P}\left[\frac{Q(X)}{P(X)} - 1\right] \\ &= \int \frac{Q(x)}{P(x)} P(x) \mathrm{d}x - 1 \\ &= 0. \end{aligned}$$

For equality to hold, $P(X) = Q(X)$ with probability 1. □

Exercise 1.0.5. *Find the MLE of $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(\theta)$.*

Solution. The log-likelihood function is

$$\begin{aligned} \ell_n(\hat{\theta}) &= \sum_{i=1}^{n} \log \Pr_{\hat{\theta}}(X_i) \\ &= \sum_{i=1}^{n} X_i \log \hat{\theta} + (1 - X_i) \log (1 - \hat{\theta}) \\ &= n\overline{X}_n \log \hat{\theta} + n(1 - \overline{X}_n) \log (1 - \hat{\theta}), \end{aligned}$$

where $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then for any $\hat{\theta}, \theta' \in [0, 1]$,

$$\frac{\ell_n(\hat{\theta}) - \ell_n(\theta')}{n} = \overline{X}_n \log \frac{\hat{\theta}}{\theta'} + (1 - \overline{X}_n) \log \frac{1 - \hat{\theta}}{1 - \theta'}.$$

For $\hat{\theta} = \overline{X}_n$, this is precisely

$$\frac{\ell_n(\hat{\theta}) - \ell_n(\theta')}{n} = D_{KL}(\mathrm{Ber}\, \hat{\theta} \parallel \mathrm{Ber}\, \theta') \geq 0.$$

Thus the MLE is $\widehat{\theta}_n = \overline{X}_n$. ■

Definition 1.0.6 (Entropy)**.** The *entropy* of a distribution P is defined as

$$H(P) = -\mathop{\mathbf{E}}_{P} \log P(X).$$

Exercise 1.0.7. *Find MLE of* $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_0, \Sigma_0)$.

Solution. The log-likelihood function is

$$\ell_n(\mu, \Sigma) = -\frac{1}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{n}\|X_i - \mu\|^2_{\Sigma^{-1}} + \text{constant}.$$

Let us first fix Σ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then ignoring the constant terms, we have to maximize

$$\begin{aligned}
&\sum_{i=1}^{n}\big\|X_i - \overline{X}_n + \overline{X}_n - \mu\big\|^2_{\Sigma^{-1}} \\
&= \sum_{i=1}^{n}\left(\big\|X_i - \overline{X}_n\big\|^2_{\Sigma^{-1}} + 2\langle X_i - \overline{X}_n,\, \overline{X}_n - \mu\rangle_{\Sigma^{-1}} + \big\|\overline{X}_n - \mu\big\|^2_{\Sigma^{-1}}\right) \\
&= \sum_{i=1}^{n}\big\|X_i - \overline{X}_n\big\|^2_{\Sigma^{-1}} + 2\Big\langle \sum_{i=1}^{n} X_i - n\overline{X}_n \,\Big|\, \Sigma^{-1} \,\Big|\, \overline{X}_n - \mu \Big\rangle + n\big\|\overline{X}_n - \mu\big\|^2_{\Sigma^{-1}} \qquad \left(\sum_{i=1}^{n} X_i - n\overline{X}_n = 0\right) \\
&= \sum_{i=1}^{n}\big\|X_i - \overline{X}_n\big\|^2_{\Sigma^{-1}} + n\big\|\overline{X}_n - \mu\big\|^2_{\Sigma^{-1}}.
\end{aligned}$$

Thus for any value of Σ, $\ell(\mu, \Sigma)$ is maximized when $\mu = \overline{X}_n$.

Now let us fix $\mu = \overline{X}_n$. Let $S = \sum_{i=1}^n |X_i - \mu\rangle\langle X_i - \mu|$. Then

$$\begin{aligned}
\sum \|X_i - \mu\|^2_{\Sigma^{-1}} &= \sum \langle X_i - \mu \,|\, \Sigma^{-1} \,|\, X_i - \mu\rangle \\
&= \sum \operatorname{Tr}\big(\Sigma^{-1}|X_i - \mu\rangle\langle X_i - \mu|\big) \\
&= \operatorname{Tr}\big(\Sigma^{-1}S\big).
\end{aligned}$$

Then

$$\begin{aligned}
\ell_n(\mu, \Sigma) - \ell_n(\mu, S) &\propto -\log\det\Sigma - \operatorname{Tr}\big(\Sigma^{-1}S\big) + \log\det S + \operatorname{Tr}\big(S^{-1}S\big) \\
&= \log\det\big(\Sigma^{-1}S\big) - \operatorname{Tr}\big(\Sigma^{-1}S\big) + n \\
&= \sum \log\lambda_i - \sum \lambda_i + n,
\end{aligned}$$

where λs are the eigenvalues of $\Sigma^{-1}S$.

$$\begin{aligned}
&= \sum (\log\lambda_i - (1 - \lambda_i)) \\
&\leq 0
\end{aligned}$$

with equality iff each $\lambda_i = 1$, that is, $\Sigma^{-1}S = I \iff \Sigma = S$.

Thus the MLE is

$$\widehat{\mu}_n = \overline{X}_n, \quad \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n\rangle\langle X_i - \overline{X}_n|.$$

■

Lecture 19. Thursday 14 Mar '24

Definition 1.0.8 (Consistency)**.** A sequence of estimators $\widehat{\theta}_n$ for a parameter θ is said to be *consistent* if

$$\widehat{\theta}_n \xrightarrow{\text{P}} \theta.$$

That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\widehat{\theta}_n - \theta\right| < \epsilon\right) = 1.$$

Theorem 1.0.9 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Then*

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Let X_is be distributed according to the pmf/pdf $f_{\theta_0}(x)$, where f can be parameterized by θ and θ_0 is the true parameter.

Recall the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

By the law of large numbers,

$$\frac{1}{n}\ell(\theta) \xrightarrow{\text{P}} \mathbf{E}[\log f_\theta(X_i)] \quad \text{if it exists.}$$

This expectation is under the true parameter θ_0.

Thus

$$\begin{aligned}\frac{1}{n}\ell(\theta) - \frac{1}{n}\ell(\theta_0) &\xrightarrow{\text{P}} \mathop{\mathbf{E}}_{\theta_0}\left[\log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right] \\ &= D_{KL}(f_{\theta_0} \parallel f_\theta) \\ &\geq 0,\end{aligned}$$

where the equality holds iff $f_{\theta_0}(X) = f_\theta(X)$ almost everywhere. Thus

$$\operatorname*{argmax}_\theta \ell(\theta) \xrightarrow{\text{P}} \theta_0.$$

1.1 How Good is This Convergence?

In this section we redefine

Definition 1.1.1 (Log-likelihood)**.** The *log-likelihood* function is

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i).$$

for convenience.

Definition 1.1.2 (Score function)**.** Given a probability mass/density function $f_\theta(x)$, where $\theta \in \mathbb{R}^d$ is the parameter, the *score function* is

$$z_\theta(x) = \frac{\partial}{\partial\theta}\log f_\theta(x) = \frac{1}{f_\theta(x)}\frac{\partial f_\theta(x)}{\partial\theta}.$$

Let $Z_x(\theta) = z_\theta(x)$ be thrice differentiable in θ. Then

$$Z_x(\theta) = Z_x(\theta_0) + Z_x'(\theta_0)(\theta-\theta_0) + \frac{1}{2}Z_x''(\theta_0)(\theta-\theta_0)^2$$

Lecture 15. Thu 04 Apr '24

Let $(Z_t)_{t\in\mathbb{N}}$ be a Markov chain with state space $S = \{s_1, \dots, s_n\}$ and transition matrix $A = (a_{ij})_{i,j=1}^n$. That is,

$$\Pr(Z_{t+1} = s_j \mid Z_t = s_i) = a_{ij}.$$

If $Z_0 \sim \mu^\top$, then the distribution of Z_n is $\mu^\top A^n$.

The Ising model is incredibly hard to compute. Instead of computing

$$\mathbf{E}[S_i] = \sum_s s_i \Pr(S = s),$$

we can sample

$$\mathbf{E}[S_i] \approx \frac{1}{m}\sum_{i=1}^{m} s_i^*,$$

where s_i^*s are sampled according to the distribution $\Pr(S = \cdot)$.

This is done via the Metropolis algorithm.

Chapter 2

Graphical Models

Let $\{X_i\}_{i=1}^d$ be a set of random variables. To each X_i assign a vertex i, and let the vertex set be $[d]$. Edges will model dependencies between the random variables.

We first review some definitions from graph theory.

2.1 Definitions

Definition 2.1.1. A *graph* $G = (V, E)$ consists of a set of vertices V and a set of edges $E \subseteq V \times V$.

A graph is *undirected* if $(u, v) \in E$ implies $(v, u) \in E$. Otherwise, it is *directed*.

A vertex v is *adjacent* to u if $(u, v) \in E$. u is said to be the *parent* of v.

The *neighbourhood* of v is the set of vertices adjacent to v.

$$N(v) = \{u \in V \mid (u, v) \in E\}.$$

Definition 2.1.2 (Paths)**.** A *path* in a graph $G = (V, E)$ is a sequence of vertices $(v_1, \ldots, v_k)$ such that $(v_i, v_{i+1}) \in E$ for all i.

A path is *simple* if all vertices are distinct. A path is *closed* if $v_1 = v_k$.

A *cycle* is a closed path with no repeated vertices (except for the first and last).

Definition 2.1.3 (Separation)**.** Let $A, B, C \subseteq V$ be disjoint. A and B are *separated* by C if every path from A to B contains a vertex in C.

We now study two instances of graphical models:

- Bayesian networks
- Markov networks

2.1.1 Bayesian Networks

Definition 2.1.4. Let $G = ([n], E)$ be a directed acyclic graph. Then $(G, X_1, \ldots, X_n)$ is a *Bayesian network* if

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i \mid \text{parent}(i))$$

where $\text{parent}(i)$ is the set of parents of i.

For convenience, we will relabel the vertices in a topological sort. Then for all i,

$$\text{parent}(i) \subseteq [i-1]$$

2.2 A real life example

Let N, T, L, X be random variables representing the following:

- N represents whether a particular patient has pneumonia.
- T represents whether they have tuberculosis.
- L represents whether they have observable lung abnormalities.
- X represents whether they have a positive X-ray.

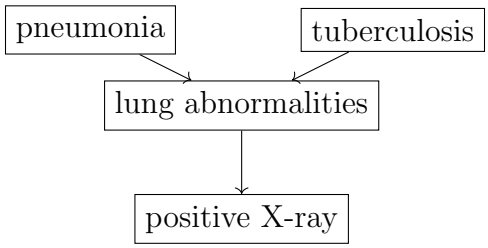
Then

$$\begin{aligned}&\Pr(N = n, T = t, L = l, X = x)\\ &\qquad = \Pr(N = n)\Pr(T = t)\Pr(L = l \mid N = n, T = t)\Pr(X = x \mid L = l).\end{aligned}$$

We will shorten such equations to

$$\Pr(N, T, L, X) = \Pr(N)\Pr(T)\Pr(L \mid N, T)\Pr(X \mid L).$$

This can be represented by the following Bayesian network:



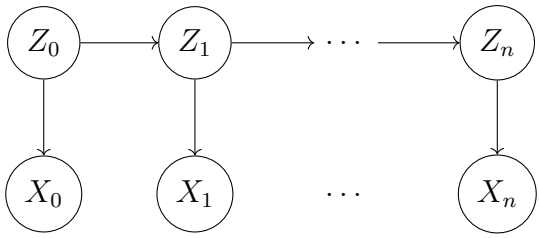
The *inference problem* is to compute

$$\Pr(N = 1 \mid X = x).$$

Suppose the X-ray machines are awesome, so that $L = X$ with probability 1.

2.2.1 HMMs

Consider the following Bayesian network:



The total probability is

$$\Pr(X_0, Z_0, \ldots, X_n, Z_n) = \Pr(Z_0) \prod_{i=1}^{n} \Pr(Z_i \mid Z_{i-1}) \prod_{i=0}^{n} \Pr(X_i \mid Z_i).$$

This is the hidden Markov model!

2.3 Markov Networks

Definition 2.3.1 (Global Markov property)**.** Let $G = ([n], E)$ be undirected. Then $(G, X_1, \ldots, X_n)$ satisfies the *global Markov property* if for all $A, B, C \subseteq [n]$ such that A and B are separated by C,

$$X_A \perp\!\!\!\perp X_B \mid X_C,$$

where $X_S = \{X_i\}_{i \in S}$.

Theorem 2.3.2 (Hammersly-Clifford theorem)**.** *If $(G, X_1, \ldots, X_n)$ satisfies the global Markov property, and $P(X_1, \ldots, X_n) > 0$, then the joint distribution of $X_1, \ldots, X_n$ factorizes over G. That is,*

$$\Pr(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{C \in cliques(G)} \psi_C(X_C),$$

where Z is a normalizing constant and ψ_C is a potential function.

Chapter 3

Principal Component Analysis

Lecture 16.
Mon 08 Apr '24

Let the data

$$\mathcal{D} = \left\{x^{(1)}, \ldots, x^{(n)}\right\} \subseteq \mathbb{R}^d$$

be drawn i.i.d. from a distribution P.

By Cauchy-Shwartz,

$$\langle u,\, v\rangle \leq \|u\|\|v\|,$$

with equality achieved when $v = \lambda u$.

THEREFORE, the maximum value of $u^\top Cu$ is achieved... when $Cu = \lambda u$.

This reminds me of

$$E = mc^2$$
$$E = \frac{hc}{\lambda}$$
$$\lambda = \frac{h}{mv}$$