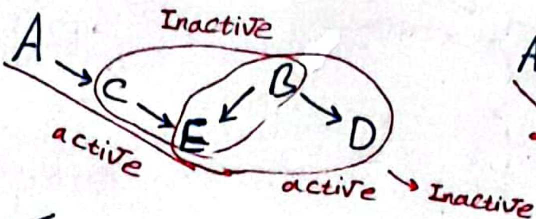
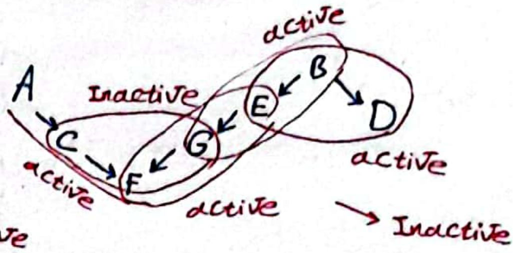
به نام خدا

علی قاسم‌زاده ۴۰۱۱۰۶۳۳۹

۱- ① $A \perp D$ ، غلط است زیرا اگر تمام مسیرهای بین A و D را در نظر بگیریم،

$A \to B \to D$ → active





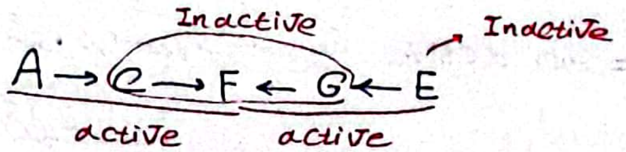
باید تمامی مسیرها Inactive باشند تا A و D مستقل باشند، در صورتی که همان مسیر $A \to B \to D$ active است ← A و D وابسته‌اند.

---

② $A \perp E | B$: غلط است زیرا اگر تمام مسیرهای بین A و E به شرطی که B را دیده باشیم را در نظر بگیریم، داریم:

$A \to \cancel{B} \to E$ → Inactive

$A \to C \to E$ → active



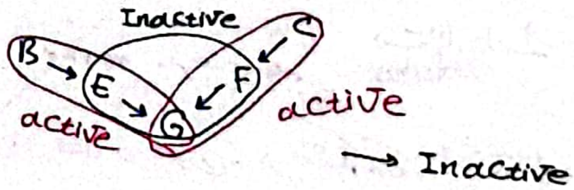
باید تمام مسیرهای Inactive باشند تا A, E به شرط B مستقل باشند، در صورتی که همان مسیر $A \to C \to E$ active است ← $A \perp E|B$ برقرار نیست.

---

③ $B \perp C | A$ ، درست است زیرا اگر تمام مسیرهای بین B و C به شرطی که A دیده باشیم را در نظر بگیریم، داریم:

$\underbrace{B \leftarrow \cancel{A} \to C}_{\text{Inactive}}$

$\underbrace{B \to E \leftarrow C}_{\text{Inactive}}$



⇐ تمامی مسیرها Inactive هستند ← B و C به شرط دیده شدن A، مستقل هستند:

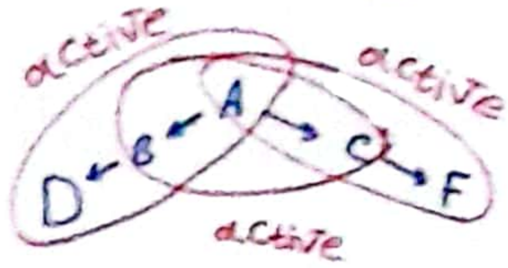
← $B \perp C | A$ برقرار است.

---

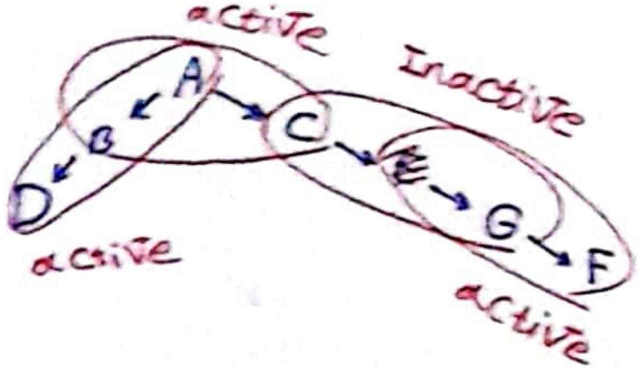
④ $B \perp C | A, E$: غلط است، مانند عبارت C است با این تفاوت که E دیده شده است ← $B \to E \leftarrow C$

یک active path است ← چون یک active path داریم بین B و C، وابسته‌اند که نیازی به بررسی مورد آخر هم نیست (البته مورد آخر Inactive است.) ← $B \perp C | A, E$ برقرار نیست

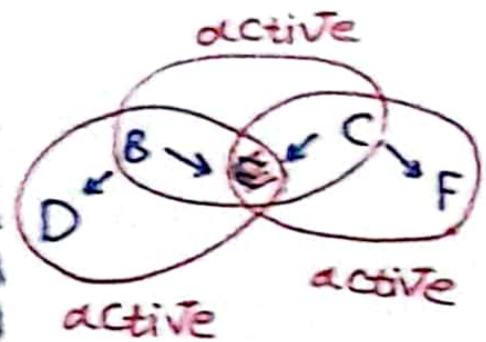
۵) $D \perp F \mid E$: غلط است، اگر تمام مسیرهای بین D و F را در نظر بگیرید داریم:



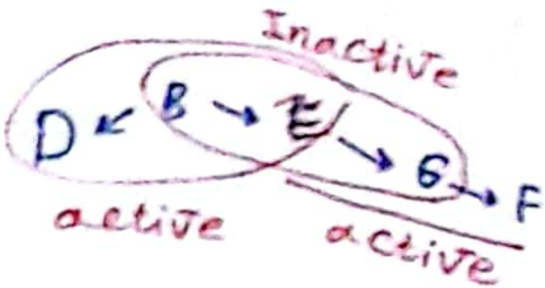
active path



Inactive path



active path



→ Inactive path

تمامی مسیرها باید Inactive باشند، در صورتی که مسیرهای active داریم ⟸ D و F به شرط دیدن E وابسته‌اند ⟸ $D \perp F \mid E$ برقرار نیست.

۲- در markov net ، جهت نداریم و استقلال ها به این صورت است که اگر مسیر unobserved بین دو نود نداشته باشیم، آن دو نود مستقل اند که به این ویژگی global markov property گفته می‌شود.

در bayes net ، جهت داریم و استقلال ها با توجه به d-separation هستند، وقتی که تمام مسیرهای بین دو نود Inactive باشند، آن دو مستقل اند.

حالا از آنجایی که اسکلت دو گراف یکسان است، در markov net هرجا بخواهیم استقلال داشته باشیم باید در bayes net هم استقلال داشته باشیم و برعکس.

حالا جفت نود هایی که مسیری بینشان نیست که مشکلی ایجاد نمی‌کند و استقلال شرطی شان یکسان است در هر دو شبکه، markov , bayes و مسئله نود هایی هستند که به هم مسیر دارند.

اگر دو نود مستقیم به هم وصل باشند که وابسته اند و در هر دو در اینطوری است و مشکلی ایجاد نمی‌شود.

مشکل جایی ایجاد می‌شود که مسیری بین دو نود باشد که علاوه بر دو نود، نود دیگر را نیز شامل شود. حالا هر نودی در وسط این مسیر بگیریم، اگه آنرا ببینیم، دو نود سر و ته مسیر به شرط این نود در markov مستقل اند. (به عبارتی کل path های بین دو نود را بگیریم و از هر کدام یه تعدادی را observe کنیم، این دو نود مستقل می‌شوند به شرط دیدن آن نود های وسطی در markov net)

پس در bayesnet هم باید همین استقلال اتفاق بیفتد، از آنجایی که ما هر ترکیبی از نود های وسط می‌توانیم برداریم و observe کنیم باید در bayesnet v-structure نداشته باشیم ($\circ \rightarrow \circ \leftarrow \circ$)

چون در این حالت با observe شدن نود Triple active تشکیل می‌شود و وابستگی ایجاد می‌کند

پس تمام سه تایی ها به صورت $\circ \leftarrow \circ \rightarrow \circ$ و $\circ \rightarrow \circ \rightarrow \circ$ باید باشند.

۳- جواب بله به صورتی، به ازای هر متغیر X ، parent های آن باید تمام همسایه هایش باشند، زیرا در markov هر نود به شرط همسایه هایش از سایر نود ها مستقل است و در bayes ، هر نود به شرط parent هایش

از سایر نودها مستقل است. همچنین G نباید دور داشته باشد، زیرا که یک شبکه‌ی bayes با دور جهت‌دار نمی‌تواند به یک شبکه‌ی markov تبدیل شود.

همچنین شرط‌های احتمالی در G باید با توزیع‌های شرطی محلی در H همخوانی داشته باشند. این همخوانی تضمین می‌کند که وابستگی‌های محلی در دو شبکه یکسان اند.

۳- با در نظر گرفتن $h_i \in H \; , \; v_i \in V$ به صورت باینری داریم که:

$$P(H_k = h_k | V) = \frac{P(H_k = h_k, V)}{P(V)} = \frac{P(H_k = h_k, V)}{\sum_{h_k} P(H_k = h_k, V)} = \frac{P(H_k = h_k, V)}{P(H_k = 0, V) + P(H_k = 1, V)}$$

که در فرمول داده شده عبارت صورت و دو عبارت مخرج به ترتیب در متغیر H_k تفاوت دارند یکی h_k، یکی ۰، یکی ۱ است و بقیه مقادیر با هم ساده می‌شوند پس داریم که:

$$P(H_k = h_k | V) = \frac{\exp(\alpha_k h_k + h_k \sum_i w_{ki} v_i) \times \cancel{t}}{\exp(0) \times \cancel{t} + \exp(\alpha_k + \sum_i w_{ki} v_i) \times \cancel{t}}$$

t متغیر مشترک هر سه عبارت است

$$\Rightarrow P(H_k = h_k | V) = \frac{\exp(\alpha_k h_k + h_k \sum_i w_{ki} v_i)}{1 + \exp(\alpha_k + \sum_i w_{ki} v_i)}$$

$$\Rightarrow P(h_j = 1 | V) = \frac{\exp(\alpha_k + \sum_i w_{ji} v_i)}{1 + \exp(\alpha_k + \sum_i w_{ji} v_i)} = \underbrace{\frac{1}{1 + \exp(-(\alpha_k + \sum_i w_{ji} v_i))}}_{\sigma(\text{—})}$$

$$\Rightarrow \boxed{P(h_j = 1 | V) = \sigma\left(\alpha_j + \sum_i w_{ji} v_i\right)} \; ,$$

$$\boxed{P(h_j = 0 | V) = 1 - \sigma\left(\alpha_j + \sum_i w_{ji} v_i\right)}$$

با روشی دیگر و بدون از استقلال هم در ادامه این رابطه را نشان می‌دهیم

حالا باید نشان دهیم که $P(H|V) = \prod_j P(H_j|V)$: طبق اسلاید ها داشتیم که اگر مسیر unobserved بین دو متغیر داشته باشیم، آن دو وابسته‌اند ~~...~~ اگر نداشته باشیم مستقل اند، در سوال دوهم توضیح دادیم، حالا هر H_i به تمام v_i ها وصل است ← اگر تمام v_i ها observed شده باشند H_i جمع تمام همسایه‌هایش v_i ها هستند، به شرط V از سایرین مستقل است، به عبارتی داشتیم که هر نود به شرط همسایه‌هایش از سایرین مستقل است ← هر H_i به شرط V، از تمام H_j ها دیگر مستقل است.

Scanned with CamScanner

در واقع داریم (طبق chain rule)

$$P(H|V) = P(H_1|V)\,\underbrace{P(H_2|H_1,V)}_{P(H_2|V)} \cdots \underbrace{P(H_n|H_1,H_2,\dots,H_{n-1},V)}_{P(H_n|V)}$$

$$\Rightarrow \underline{P(H|V) = \prod_j P(H_j|V)}$$

حالا برای هر H_j داریم که:

$$P(H_j|V) = \sigma\left(\alpha_j + \sum_i w_{ji} v_i\right)^{h_j} \left(1 - \sigma\left(\alpha_j + \sum w_{ji} v_i\right)\right)^{1-h_j}$$

$$\Rightarrow \boxed{P(H|V) = \prod_j \sigma\left(\alpha_j + \sum_i w_{ji} v_i\right)^{h_j} \left(1 - \sigma\left(\alpha_j + \sum w_{ji} v_i\right)\right)^{1-h_j}}$$

ب) دوباره از chain rule استفاده می‌کنیم

$$P(V|H) = P(V_1 \_\_\_ V_m|H) = P(V_1|H)\,P(V_2|V_1,H) \cdots P(V_m|V_1 \dots V_{m-1},H)$$

که به طور مشابه اگر تمام H_iها observe باشند هر V_K به شرط تمام H_iها از سایر V_jها مستقل است دقیقاً مطابق توضیحات بخش الف فقط اینبار جای V_i و H_i عوض شده است.

$$\Rightarrow P(V|H) = \prod P(V_j|H)$$

$$P(V_j = v_j|H) = \frac{P(V_j = v_j, H)}{P(H)} = \frac{P(V_j = v_j, H)}{P(V_j = 1, H) + P(V_j = 0, H)}$$

به طور مشابه با الف صورت و مخرج شامل عبارت مشترک هستند داریم:

t مقدار مشترک است

$$\Rightarrow P(V_j = v_j|H) = \frac{\exp\left(b_j v_j + v_j \sum_i w_{ij} h_i\right) \times \cancel{t}}{\exp(0) \times \cancel{t} + \exp\left(b_j + \sum_i w_{ij} h_i\right) \times \cancel{t}} = \frac{\exp\left(b_j v_j + v_j \sum_i w_{ij} h_i\right)}{1 + \exp\left(b_j + \sum_i w_{ij} h_i\right)}$$

$\to \sigma(\text{—})$

$$\Rightarrow P(v_j=1|H) = \frac{\exp(b_j + \sum_i w_{ij}h_i)}{1+\exp(b_j+\sum_i w_{ij}h_i)} = \frac{1}{1+\exp(-(b_j+\sum_i w_{ij}h_i))}$$

$$\Rightarrow P(v_j=1|H) = \sigma\left(b_j + \sum_i w_{ij}h_i\right)$$

$$P(v_j=0|H) = 1-\sigma\left(b_j + \sum_i w_{ij}h_i\right)$$

همچنین داریم که ،

$$P(v_j|H) = \sigma\left(b_j+\sum_i w_{ij}h_i\right)^{v_j}\left(1-\sigma\left(b_j+\sum_i w_{ij}h_i\right)\right)^{1-v_j}$$

$$\Longrightarrow P(v|H) = \prod_j \sigma\left(b_j+\sum_i w_{ij}h_i\right)^{v_j}\left(1-\sigma\left(b_j+\sum_i w_{ij}h_i\right)\right)^{1-v_j}$$

ب)

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

$$\Rightarrow P(H) = \frac{\frac{1}{Z}\exp\left(\sum_i a_ih_i + \sum_j b_jv_j + \sum_{i,j} w_{ij}h_iv_j\right)}{\prod_j \sigma\left(b_j+\sum_i w_{ij}h_i\right)^{v_j}\left(1-\sigma\left(b_j+\sum_i w_{ij}h_i\right)\right)^{1-v_j}}$$

راه دیگر

می‌توانیم تعریف کنیم:

$$E(v,h) = -\sum_i a_ih_i - \sum_j b_jv_j - \sum w_{ij}h_iv_j$$

$$\Rightarrow E(v,h) = -(a^Th + b^Tv + h^TWv)$$

$$\Rightarrow P(v,h) = \frac{1}{Z}e^{-E(v,h)}$$

$$P(h) = \frac{1}{Z}\sum_{\{v\}} e^{-E(v,h)}$$

در هر دو حالت نمی‌توانیم توزیع P(H) را روی Hها تجزیه کنیم. دلیل محکم‌تری نیز داریم که نمی‌توانیم P(H) را تجزیه کنیم.

$$P(H) = \sum_{V} P(V,H)$$

در هر سری ما یک سری V_i را داریم و به ازای آن سری V_i، Hها مستقل می‌شوند، در واقع پس از حساب کردن این Σ، H_jها یه سری ارتباط با هم دارند و مستقل نیستند، بخاطر جمع روی مقادیر مختلف برای V_iها، حالا چون مستقل نیستند و با هم ارتباط‌هایی دارند نمی‌توانیم P(H) را روی H_iها بشکنیم.

شهود بیشتر: هر H_j با یه V_i از طریق w_{ij} مرتبط است، وقتی جمع می‌بندیم روی تمام حالت‌های V تا P(H) را بدست آوریم تمام hidden unitها مرتبط می‌شوند، چون احتمال joint به تمام configurationهای V و ارتباطات (Wها) بستگی داره.

د) استقلال‌های شرطی که در گراف هستند، در توزیع هم هستند، معمولاً در RBN این اتفاق نمی‌افتد که استقلال شرطی در گراف باشد ولی در توزیع نباشد و هر دوتا hidden به شرط تمام V ها مستقل‌اند و هر دوتا V_i به شرط تمام Hها مستقل‌اند.

ولی ممکن است استقلال شرطی در توزیع داشته باشیم که در گراف نباشد، برای مثال اگر یه سری وزن‌ها صفر باشند یه استقلالی در تابع توزیع احتمالمون وجود دارد در صورتی که در گراف این استقلال دیده نمی‌شود، و فقط مسیر unobserved اگر داشتیم می‌گوییم دو متغیر وابسته‌اند، در صورتی که در این مسیر ممکن است یکی از وزن‌ها صفر باشه و آن دو متغیر مستقل باشند بدون نیاز به دیده شدن نودی در این مسیر.

۲. راه حلی دیگر:

$$P(H|V) = \frac{P(V,H)}{P(V)} = \frac{P(V,H)}{\sum_H P(V,H)} = \frac{\cancel{\frac{1}{Z}} \exp\left(\sum_i a_i h_i + \cancel{\sum_j b_j v_j} + \sum_{i,j} w_{ij} h_i v_j\right)}{\cancel{\frac{1}{Z}} \sum_H \exp\left(\sum_i a_i h_i + \underbrace{\cancel{\sum_j b_j v_j}} + \sum_{i,j} w_{ij} h_i v_j\right)}$$

این در صورت و تمامی عبارت های مخرج وجود دارد، ثابت است.

$$\Rightarrow P(H|V) = \frac{\exp\left(\sum_i a_i h_i + \sum_{i,j} w_{ij} h_i v_j\right)}{\sum_H \exp\left(\sum_i a_i h_i + \sum_{i,j} w_{ij} h_i v_j\right)} = \frac{\prod_i \exp\left(a_i h_i + h_i \left(\sum_j w_{ij} v_j\right)\right)}{\sum_H \underbrace{\prod_i \exp\left(a_i h_i + h_i \left(\sum_j w_{ij} v_j\right)\right)}_{1+\exp(a_1 + \sum w_{1j} v_j) + \cdots}}$$

جمع تمام حالت های صفر و یک

جمع کل حالت ها برای H ها

$$\Rightarrow P(H|V) = \frac{\prod_i \exp\left(a_i h_i + h_i \left(\sum_j w_{ij} v_j\right)\right)}{\prod_i \left(1 + \exp\left(a_i + \left(\sum_j w_{ij} v_j\right)\right)\right)}$$

که این عبارت برابر است با:
با توجه به باینری بودن h_i ها

$$\prod_i \left(1 + \exp\left(a_i + \left(\sum_j w_{ij} v_j\right)\right)\right)$$

$$\Rightarrow P(H|V) = \prod_i \underbrace{\frac{\exp\left(a_i h_i + h_i \sum w_{ij} v_j\right)}{1 + \exp\left(a_i + \sum_j w_{ij} v_j\right)}}_{P(H_i|V)} = \prod_i P(H_i|V)$$

همچنین داریم که

$$\boxed{P(H_i = 1|V) = \sigma\left(a_i + \sum_j w_{ij} v_j\right)} \leftarrow P(H_i = 1|V) = \frac{\exp\left(a_i + \sum_j w_{ij} v_j\right)}{1 + \exp\left(a_i + \sum_j w_{ij} v_j\right)}$$

ب) می‌دانیم که V و H متقارن اند، طبق شکل مال تمام H ها به تمام V وصل اند و تمام V ها به تمام H ها وصل اند. مشابه الف با نوشتن روابط بدست می‌آید که:

$$\boxed{P(V|H) = \prod_i P(v_i|H), \quad P(v_i = 1|H) = \sigma\left(b_j + \sum_i w_{ij} h_i\right)}$$

$$\Rightarrow \begin{cases} P(H|V) = \prod_j \sigma\left(a_j + \sum_i w_{ji} v_i\right)^{h_j} \left(1 - \sigma\left(a_j + \sum_i w_{ji} v_i\right)\right)^{1-h_j} \\ P(V|H) = \prod_j \sigma\left(b_j + \sum_i w_{ij} h_i\right)^{v_j} \left(1 - \sigma\left(b_j + \sum_i w_{ij} h_i\right)\right)^{1-v_j} \end{cases}$$

۴- Causality باعث افزایش تفسیرپذیری مدل می‌شود چراکه اهمیت هر component مدل را می‌تواند توضیح دهد. خیلی از مدل های جدید مثل شبکه های عمیق یا روش‌های سنتی machine learning قابل تفسیر نیستند و برای دیدن تأثیر هر لایه یا component، می‌توانیم مدل را یکبار بدون آن لایه یا component دیگبار با آن train کنیم و تأثیرش را ببینیم ولی چنین کاری عملی نیست چراکه هزینه محاسباتی بالایی دارد.

حالا ایده اصلی این است که ساختار یک DNN (شبکه عصبی عمیق) را با یک SCM مدل کنیم و causal effect های هر component روی خروجی را توسط causal reasoning تخمین بزنیم.

در واقع causal structure ما را قادر می‌کند که به سوال های what-if جنس برای تأثیر هر component روی خروجی پاسخ دهیم.

SCM هر لایه یا unit در DNN به یک node در causal graph تبدیل می‌کند و یال های جهت دار طوریکه ارتباط بین این نود ها را نشان می‌دهد، causal graph هم به ما این امکان را می‌دهد تا causality در شبکه را بفهمیم و مدل را بهتر تفسیر کنیم.

همچنین SCM ما را قادر می‌کند تا ACE (Average causal effect) را حساب کنیم برای هر نورون، فیلتر، لایه و خروجی. برای مثال اگر متغیر مشخصی به یک نورون بدهیم، می‌توانیم تغییرات خروجی پیش بینی شده را با توجه به این تغییر حساب کنیم.

SCM به ما این امکان را می‌دهد که counterfactual analysis انجام دهیم یا به سری سوالات فرضی که مثلاً اگر فلان مقدار را به فلان نود بدهیم چه اتفاقی می‌افتد ــــــ

مدل های پیچیده DNN را با یه سری abstraction، ساده می‌کند و کمک می‌کند بهتر مدل را تفسیر کنیم.

با انتساب یه سری latent feature توسط causal concept، فیچرهایی که برای انسان قابل فهم تر اند و پنهانی بین لایه های مدل پدید می‌آورد و بهتر می‌توانیم مدل را بفهمیم و تفسیر کنیم.

به طور خلاصه SCM تفسیر پذیری DNN ها را با formalize کردن یه سری ارتباط causal افزایش می‌دهد

طبق اسلایدها داشتیم که ،

$$A_{ji} = P(Z_{n} = i \mid Z_{n-1} = j)$$

$$\pi_i = P(Z_1 = i)$$

$$\phi_{ik} = P(X_n = k \mid Z_n = i)$$

که X ها و Z ها به ترتیب متغیرهای observed و unobserved هستند.

حالا گام های الگوریتم را می نویسیم:

E-step: Use x and current θ^t to calculate $P(Z \mid X, \theta^t)$

M-step: $\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\ E_{Z \sim P(Z \mid X, \theta^t)}\left[\log P(X, Z \mid \theta)\right]$

حالا $P(X,Z\mid\theta)$ را محاسبه می کنیم، داریم که :

$$P(X,Z\mid\theta) = P(Z_1\mid\pi)\prod_{n=2}^{N} P(Z_n\mid Z_{n-1},A)\prod_{n=1}^{N} P(X_n\mid Z_n,\phi)$$

$$P(Z_1\mid\pi_K) = \begin{cases} 0 & \text{if } Z_1 \neq k \\ \pi_K & \text{else} \end{cases} \longrightarrow \boxed{P(Z_1\mid\pi) = \prod_{k=1}^{K} \pi_k^{\mathbb{1}(Z_1=k)}}$$

بردار $(1 \text{—} K)$ داریم که Z_1 یکی از آنها است به احتمال π_i هر کدام از $\{1,\dots,K\}$ $i \in \{1,\dots,K\}$ می تواند باشد ← توزیع احتمال آن به صورت بالا است.

$$P(Z_n\mid Z_{n-1},A) \rightsquigarrow P(Z_n = k\mid Z_{n-1} = j, A) = \begin{cases} 0 & \text{if } Z_n \neq k \text{ or } Z_{n-1} \neq j \\ A_{jk} & \text{else} \end{cases}$$

Z_n هر مقداری در $\{1,\dots,K\}$ می تواند باشد، Z_{n-1} هم همین طور و احتمال ~~...~~ A_{jk} هر کدام از $\{1,\dots,K\}$ و $k \in \{1,\dots,K\}$ و j می تواند باشد ← توزیع احتمال آن به صورت زیر است:

$$\boxed{P(Z_n\mid Z_{n-1},A) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{\mathbb{1}(Z_n=k)\mathbb{1}(Z_{n-1}=j)}}$$

همون حالت برنولی که می نوشتیم $\theta^x(1-\theta)^{1-x}$ را تعمیم دادیم.

$\cancel{P(X_n\mid Z_n,\phi)} = P(X_n = k\mid Z_n = i, \phi) = \begin{cases} 0 & \text{if } Z_n \neq i \\ P(X_n = k\mid\phi) & \text{else} \end{cases}$

$P(X_n = k\mid\phi) \longrightarrow P(X_n\mid\phi_k)$

حالا برای بدست آوردن توزیع $P(x_n|z_n,\phi)$ داریم که

$$P(x_n|z_n,\phi) = \prod_{k=1}^{K} P(x_n|\phi_k)^{\mathbb{1}(z_n=k)}$$

حالا از احتمال $P(x,z|\theta)$ لگاریتم می‌گیریم:

$$\log P(x,z|\theta) = \sum_{k=1}^{K} \mathbb{1}(z_1=k)\log \pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \underbrace{\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)}_{\mathbb{1}(z_{n-1}=j,\, z_n=k)} \log A_{jk} + \sum_{n=1}^{K}\sum_{k=1}^{K} \mathbb{1}(z_n=k)\log P(x_n|\phi_k)$$

حالا روی $Z \sim P(z|\theta,x)$ ، Expected می‌گیریم، داریم که:

$$E(\log P(x,z|\theta)) = \sum_{k=1}^{K} E(\mathbb{1}(z_1=k))\log \pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} E(\mathbb{1}(z_n=k)\,\mathbb{1}(z_{n-1}=j)) \log A_{jk} + \sum_{n=1}^{K}\sum_{k=1}^{K} E(\mathbb{1}(z_n=k))\log P(x_n|\phi_k)$$

حالا تعریف می‌کنیم: $\gamma(z_n) = P(z_n|x,\theta)$, $\xi(z_{n-1},z_n) = P(z_n,z_{n-1}|x,\theta)$

از آنجایی که $\mathbb{1}(z_n=k)$, $\mathbb{1}(z_{n-1}=j)$, $\mathbb{1}(z_1=k)$ ، متغیرهای باینری هستند $\leftarrow$

$\leftarrow$ Expected هر متغیر باینری برابر است! $P(z=1)$ $\leftarrow$

$$E(\mathbb{1}(z_n=k)) = \gamma(\mathbb{1}(z_n=1))$$

$$E(\mathbb{1}(z_n=k)\,\mathbb{1}(z_{n-1}=j)) = \xi(\mathbb{1}(z_n=k)\,\mathbb{1}(z_{n-1}=j))$$

$$\rightsquigarrow \sum_{k=1}^{K} \gamma(\mathbb{1}(z_1=k))\log \pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \xi(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k))\log A_{jk} + \sum_{n=2}^{N}\sum_{k=1}^{K} \gamma(\mathbb{1}(z_n=k))\log P(x_n|\phi_k)$$

حالا برای مشتق گیری کردن روی A_{ij} با یک قید ← for each j : $\sum_{k}^{K} A_{jk} = 1$

از تکنیک ضرایب لاگرانژ استفاده می کنیم:

$$\mathcal{L} = \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right) \log A_{jk} + \sum_{j=1}^{K} \lambda_j \left(1 - \sum_{k=1}^{K} A_{jk}\right)$$

$$\frac{\partial \mathcal{L}}{\partial A_{jk}} = \frac{\xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right)}{A_{jk}} - \lambda_j = 0 \rightsquigarrow A_{jk} = \frac{\xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right)}{\lambda_j},$$

$$\sum_{k=1}^{K} A_{jk} = \underbrace{\sum_{k=1}^{K} \frac{\xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right)}{\lambda_j} = 1}_{\lambda_j = \sum_{k=1}^{K} \xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right)} \longrightarrow \hat{A}_{jk} = \frac{\xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=k)\right)}{\sum_{l=1}^{K} \xi\left(\mathbb{1}(z_{n-1}=j)\,\mathbb{1}(z_n=l)\right)}$$

مطابق بالا برای π_k و ϕ_{ik} هم همین روابط را می نویسیم و خواهیم داشت که

$$\pi_k = \frac{\gamma\left(\mathbb{1}(z_1=k)\right)}{\sum_{j=1}^{K} \gamma\left(\mathbb{1}(z_1=j)\right)}, \quad \phi_{ik} = \frac{\sum_{n=1}^{N} \gamma\left(\mathbb{1}(z_n=k)\,\mathbb{1}(x_n=i)\right)}{\sum_{n=1}^{N} \gamma\left(\mathbb{1}(z_n=k)\right)}$$

بر طولانی تر نشدن محاسبات آنها را نمی آوریم.

۶- از Laplace smoothing استفاده می‌کنیم تا نتایج بهتر شوند و از بروز مشکلاتی مثل صفر ~~بودن~~ شدن یکی از احتمال‌ها بدلیل تعداد کمی دیتا جلوگیری شود حالا داریم:

۶ در مخرج تعداد کلمات unique است.

$$P(\text{work-related}) = \frac{۳}{۴}, \quad P(\text{not-work-related}) = \frac{۱}{۴}.$$

$$P(\text{schedule}|\text{work-related}) = \frac{۴+۱}{۸+۶} = \frac{۵}{۱۴}, \quad P(\text{schedule}|\text{not-work-related}) = \frac{۱+۱}{۳+۶} = \frac{۲}{۹}$$

$$P(\text{event}|\text{work-related}) = \frac{۲+۱}{۸+۶} = \frac{۳}{۱۴}, \quad P(\text{event}|\text{not-work-related}) = \frac{۰+۱}{۳+۶} = \frac{۱}{۹}$$

$$P(\text{meeting}|\text{work-related}) = \frac{۱+۱}{۸+۶} = \frac{۲}{۱۴}, \quad P(\text{meeting}|\text{not-work-related}) = \frac{۰+۱}{۳+۶} = \frac{۱}{۹}$$

$$P(\text{deadline}|\text{work-related}) = \frac{۱+۱}{۸+۶} = \frac{۲}{۱۴}, \quad P(\text{deadline}|\text{not-work-related}) = \frac{۰+۱}{۳+۶} = \frac{۱}{۹}$$

$$P(\text{movie}|\text{work-related}) = \frac{۰+۱}{۸+۶} = \frac{۱}{۱۴}, \quad P(\text{movie}|\text{not-work-related}) = \frac{۱+۱}{۳+۶} = \frac{۲}{۹}$$

$$P(\text{lunch}|\text{work-related}) = \frac{۰+۱}{۸+۶} = \frac{۱}{۱۴}, \quad P(\text{lunch}|\text{not-work-related}) = \frac{۱+۱}{۳+۶} = \frac{۲}{۹}.$$

حالا به محاسبه احتمال work-related , not-work-related می‌پردازیم:

$$P(\text{work-related}|\text{email}) = \frac{P(\text{email}|\text{work-related})\,P(\text{work-related})}{P(\text{email})}$$

$$P(\text{not-work-related}|\text{email}) = \frac{P(\text{email}|\text{not-work-related})\,P(\text{not-work-related})}{P(\text{email})}$$

P(email) در هردو مشترک است پس می‌توانیم از آن صرف نظر کنیم و بقیه دو قسمت را با هم مقایسه کنیم.

$$P(\text{work-related}|\text{email}) = P(\text{work-related})\,P(\text{schedule}|\text{work-related})^{۲}\,P(\text{movie}|\text{work-related})\,P(\text{lunch}|\text{work-related}) / P(\text{email})$$

$$\Rightarrow P(\text{email}|\text{work-related}) = \frac{۳}{۴} \times \frac{۱}{۱۴} \times \frac{۱}{۱۴} \;/\; P(\text{email})$$

$$P(\text{not-work-related}|\text{email}) = P(\text{not-work-related})\,P(\text{schedule}|\text{not-work-related})^{۲}\,P(\text{movie}|\text{not-work-related})\,P(\text{lunch}|\text{not-work-related}) / P(\text{email})$$

$$\Rightarrow P(\text{not-work-related}|\text{email}) = \frac{۱}{۴} \times \left(\frac{۲}{۹}\right)^{۲} \times \frac{۲}{۹} \times \frac{۲}{۹} \;/\; P(\text{email})$$

⇐ با توجه به مقادیر بدست آمده، $P(\text{not work-related} \mid \text{email}) < P(\text{work-related} \mid \text{email})$

پس به احتمال خوبی ایمیل داده شده مربوط به کار است.

تفاوت ⓓ naive با discriminative، در naive bayes ما توزیع $P(X,Y)$ را مدل می‌کنیم و سپس با استفاده از قانون بیز $P(Y|X)$ را بدست می‌آوریم و فرض می‌کنیم که بین فیچرها استقلال داریم. در مدل discriminative ما مستقیم $P(Y|X)$ را مدل می‌کنیم و هدفمان پیدا کردن یه کران تصمیم‌گیری هست که کلاس‌ها را جدا کند. مثل logistic reg یا svm و معمولاً این مدل‌ها بهتر عمل می‌کنند، زیرا که استقلال بین فیچرها را فرض نمی‌کند و مدل دقت بیشتری خواهد داشت. (فیچرها الزاماً مستقل نیستند ولی ما در naive بیز مستقل فرضشون می‌کنیم و برای همین بدتر عمل می‌کنند)

$$q(x|x^s) = Exp(x|\lambda = x^s+1)$$

آ) از آنجایی که $x^s=1$ است $q = Exp(2) \leftarrow$

$$\boxed{E(q) = \frac{1}{\lambda} \longrightarrow E(q) = 0/5}$$

گفته شده که نمونه‌ی سمپل برداری شده برابر است با میانگین توزیع $\leftarrow x^p$ که همان نمونه سمپل گیری شده است، برابر است با $\frac{1}{2}$.

ب) $\alpha = \frac{\tilde{p}(x^p)}{\tilde{p}(x^s)}$ طبق نمودار احتمال در نقاط ۰/۵ و ۱ برابر است $\leftarrow$ $\alpha=1$ است.

پ) از آنجایی که q یک توزیع نمایی بود، داریم که،

$$q(x^s|x^p) = Exp(x^s|\lambda = 1+x^p) = exp(1|\lambda = 1/5) = \underline{1/5e^{-1/5}}$$

$$q(x^p|x^s) = Exp(x^p|\lambda = 1+x^s) = exp(0/5|\lambda=2) = 2e^{-2\times 0/5} = 2e^{-1}$$

$$\Rightarrow \frac{q(x^s|x^p)}{q(x^p|x^s)} = \frac{1/5e^{-1/5}}{2e^{-1}} = \frac{3}{4}e^{-0/5} \longrightarrow \text{Hastings correlation} < 1$$

ت) $\alpha_1 = 1$ ، $\alpha_2 = \frac{\tilde{p}(x^p)\,q(x^s|x^p)}{\tilde{p}(x^s)\,q(x^p|x^s)} \leftarrow$

$$\alpha_2 = \underbrace{\frac{\tilde{p}(x^p)}{\tilde{p}(x^s)}}_{\alpha_1} \times \underbrace{\frac{q(x^s|x^p)}{q(x^p|x^s)}}_{\text{Hastings correlation}}$$

$$\Rightarrow \alpha_2 < \alpha_1$$

Hastings-correlation باعث کوچکتر شدن الفا می‌شود $\leftarrow$ در الگوریتم Hasting، به احتمال کمتری این proposed move انجام می‌شود، این باعث می‌شود بازه‌هایی با احتمال کمتر در $\tilde{P}_\theta(x)$ شانس کمتری داشته باشیم تا برویم به آنجا.

= جوابی مناقشه کارانه‌تر عمل می‌کند با وجود Hastings correlation و به نواحی با احتمال ~~...~~ کمتر می‌رود.

$$p(\mu,\tau|x) = \frac{p(\mu,\tau,x)}{p(x)} \quad , \quad p(\mu,\tau,x) = p(x|\mu,\tau)\,p(\mu|\tau)\,p(\tau)$$

$$\rightarrow q(\mu,\tau) = q(\mu)\,q(\tau)$$

$$\rightarrow \ln q(\mu) = E_{q(\tau)}\left[\ln q(\mu,\tau)\right] = E_{q(\tau)}\left[\ln\left(\frac{p(\mu,\tau,x)}{p(x)}\right)\right]$$

$$= E_{q(\tau)}\left[\ln(p(\mu,\tau,x))\right] - \underbrace{E_{q(\tau)}\left[\ln p(x)\right]}_{\text{constant}}$$

$$\Rightarrow \ln q(\mu) \propto E_{q(\tau)}\left[\ln p(x|\mu,\tau) + \ln p(\mu|\tau) + \ln p(\tau)\right]$$

$$\ln p(x|\mu,\tau) \overset{i.i.d}{=} \sum_{i=1}^{N} \ln\left(\frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(x_i-\mu)^2\right)\right) = \frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{N}{2}\ln(2\pi)$$

$$\ln p(\mu|\tau) = \ln\left(\frac{\sqrt{\tau.\tau_0}}{\sqrt{2\pi}} \exp\left(-\frac{\tau.\tau_0}{2}(\mu-\mu_0)^2\right)\right) = \frac{1}{2}\ln(\tau.\tau_0) - \frac{\tau.\tau_0}{2}(\mu-\mu_0)^2 - \frac{1}{2}\ln(2\pi)$$

$$\Rightarrow \ln q(\mu) = E_{q(\tau)}\left[\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\ln(\tau.\tau_0) - \frac{\tau.\tau_0}{2}(\mu-\mu_0)^2\right] + \text{constant}$$

حالا قرار می دهیم $E_{q(\tau)}[\tau] = \tilde{\tau}$ داریم:

$$\Rightarrow \ln q(\mu) = -\frac{\tilde{\tau}}{2}\left(\sum_{i=1}^{N}(x_i-\mu)^2 + \tau_0(\mu-\mu_0)^2\right) + \text{const} \rightarrow$$ نسبت به μ.

تنها تابع درجه 2 در توان exp دارد، گاوسی است:

$$\rightarrow q(\mu) \sim N(\mu; \tilde{\mu}, \tilde{\sigma}^2)$$

$$\tilde{\mu} = \frac{\tilde{\tau}\left(\sum_{i=1}^{N} x_i + \tau_0\mu_0\right)}{N+\tau_0} \quad , \quad \tilde{\sigma}^2 = \frac{1}{\tilde{\tau}(N+\tau_0)}$$

$$\Rightarrow q(\mu) = N\left(\mu;\ \frac{\tilde{\tau}\left(\sum_{i=1}^{N} x_i + \tau_0 \mu_0\right)}{N+\tau_0},\ \frac{1}{\tilde{\tau}(N+\tau_0)}\right)$$

حالا به محاسبه $q(\tau)$ می پردازیم، خواهیم داشت:

$$\ln q(\tau) = E_{q(\mu)}\left[\ln q(\mu,\tau)\right] = E_{q(\mu)}\left[\ln\left(p(x,\mu,\tau)/p(x)\right)\right] =$$

$$E_{q(\mu)}\left[\ln p(x,\mu,\tau)\right] - \underbrace{E_{q(\mu)}\left[\ln p(x)\right]}_{\text{constant}}$$

$$\Rightarrow \ln q(\tau) \propto E_{q(\mu)}\left[\ln p(x|\mu,\tau) + \ln p(\mu|\tau) + \ln p(\tau)\right] \Rightarrow$$

$$\ln q(\tau) \propto E_{q(\mu)}\left[\frac{N}{2}\ln\tau - \frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \underbrace{\frac{1}{2}\ln(\tau\tau_0)}_{\frac{1}{2}\ln\tau_0 + \frac{1}{2}\ln\tau} - \frac{\tau\tau_0}{2}(\mu-\mu_0)\right] + \overset{\ln p(\tau)}{\cancel{\text{constant}}}$$

$$(\alpha-1)\ln\tau - \beta\tau - \ln\Gamma(\alpha) + \alpha\ln\beta$$

$\cancel{E_{q(\tau)} \propto}$

$$\Rightarrow E_{q(\tau)} \propto \left(\frac{N}{2}+\frac{1}{2}\right)\ln\tau - \frac{\tau}{2}E_{q(\mu)}\left[\sum_{i=1}^{N}(x_i-\mu)^2 + \tau_0(\mu-\mu_0)^2\right] + \ln p(\tau) + \text{constant}$$

$$\Rightarrow E_{q(\tau)} \propto \left(\frac{N}{2}+\frac{1}{2}\right)\ln\tau - \frac{\tau}{2}\left(\sum_{i=1}^{N} E_{q(\mu)}\left[(x_i-\mu)^2\right] + \tau_0\left(E_{q(\mu)}\left[(\mu-\mu_0)^2\right]\right)\right) + \ln p(\tau)$$

$\boxed{\tilde{\mu} = E_{q(\mu)}[\mu]}$ در نظر می گیریم:

$\sigma_\mu^2 = V_{q(\mu)}(\mu)$

$$\longrightarrow E_{q(\mu)}\left[(x_i-\mu)^2\right] = (x_i-\tilde{\mu})^2 + \sigma_\mu^2,\quad E_{q(\mu)}\left[(\mu-\mu_0)^2\right] = (\tilde{\mu}-\mu_0)^2 + \sigma_\mu^2$$

$$\rightarrow E_{q(\mu)}\left[\frac{\tau}{2}\sum_{i=1}^{N}(x_i-\mu)^2\right] = \frac{\tau}{2}\sum_{i=1}^{N}\left((x_i-\tilde{\mu})^2 + \sigma_\mu^2\right)$$

$$E_{q(\mu)}\left[\frac{\tau\tau_0}{2}(\mu-\mu_0)^2\right] = \frac{\tau\tau_0}{2}\left((\tilde{\mu}-\mu_0)^2 + \sigma_\mu^2\right)$$

$$\Rightarrow \ln q(\tau) \propto \left(\alpha + \frac{N+1}{2} - 1\right) \ln \tau - \tau\left(\beta + \frac{1}{2}\sum_{i=1}^{N}\left((x_i-\bar{\mu})^2 + \sigma_\mu^2\right) + \frac{\tau_0}{2}\left((\bar{\mu}-\mu_0)^2 + \sigma_\mu^2\right)\right)$$

$$q(\tau) \propto \tau \, e^{\left(\alpha + \frac{N+1}{2} - 1\right)} \Big/ e^{\tau\left(\beta + \frac{1}{2}\sum_{i=1}^{N}\left((x_i-\bar{\mu})^2 + \sigma_\mu^2\right) + \frac{\tau_0}{2}\left((\bar{\mu}-\mu_0)^2 + \sigma_\mu^2\right)\right)}$$

$$\rightsquigarrow q(\tau) = Gamma(\tau, \alpha', \beta') \,, \quad \alpha' = \alpha + \frac{N+1}{2}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{N}\left((x_i-\bar{\mu})^2 + \sigma_\mu^2\right) + \frac{\tau_0}{2}\left((\bar{\mu}-\mu_0)^2 + \sigma_\mu^2\right)$$