دانشگاه آزاد اسلامی واحد تبریز

نام درس: یادگیری ماشین

بخش: شبکه های باور بیزی

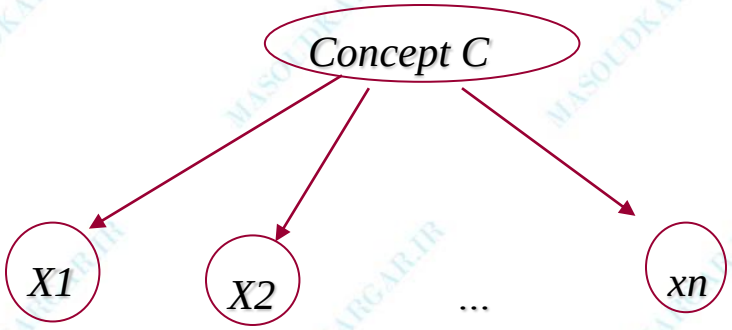
نام استاد: دکتر مسعود کارگر

مقدمه

- در عمل پیاده‌سازی Bayes Optimal Classifier بسیار پرهزینه است.
- همانگونه که دیدیم دسته‌بندی‌کننده Naive Bayes Classifier بر این اصل استوار بود که مقادیر ویژگی‌ها مستقل شرطی باشند .اما این یک شرط بسیار محدود کننده است که غالبا برآورده نمی‌شود.
- شبکه‌های باور بیزی یا Bayesian Belief Networks که Bayes Netsهم نامیده می‌شود روشی است برای توصیف توزیع احتمال توام مجموعه ای از متغیرها.
- BBN استقلال شرطی زیر مجموعه‌ای از متغیرها را قابل توصیف کرده و امکان ترکیب دانش قبلی درباره وابستگی متغیرها را با داده‌های آموزشی فراهم می‌اورد.

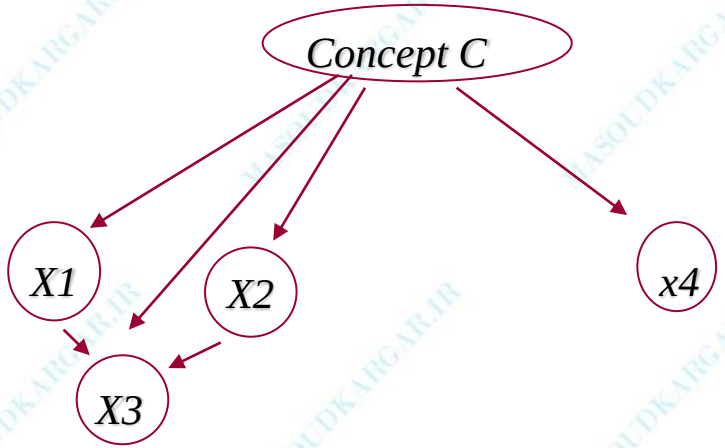
مقدمه

دسته‌بندی‌کننده ساده بیزی



$$P(x1,x2,...xn,c) = P(c)\ P(x1|c)\ P(x2|c)\ ...\ P(xn|c)$$

شبکه باور بیزی



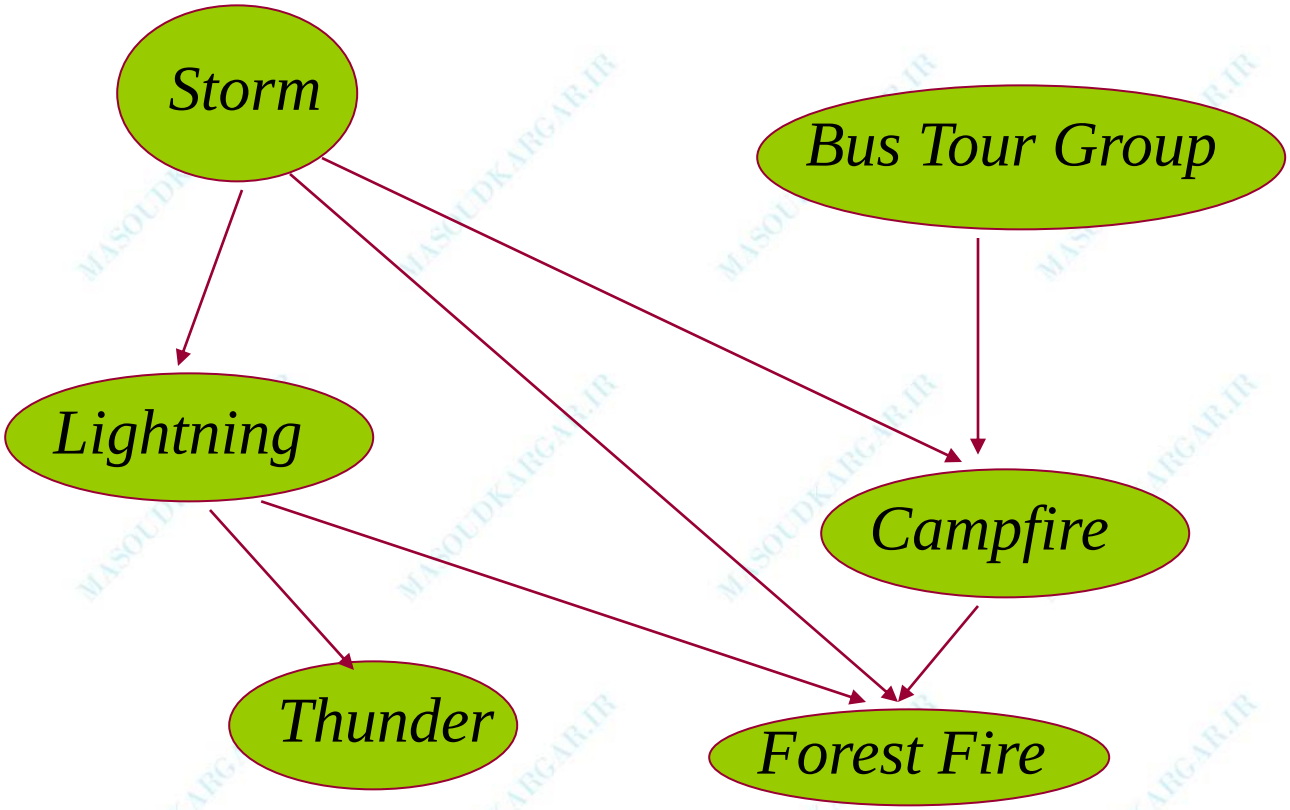
$$P(x1,x2,...xn,c) = P(c)\ P(x1|c)\ P(x2|c)\ P(x3|x1,x2,c)P(x4,c)$$

مقدمه

- اگر $a_1,a_2,...a_n$ مجموعه‌ای از ویژگیها و یا متغیرها *باشند* BN می‌تواند احتمال هر ترکیبی از آنان را بیان کند.

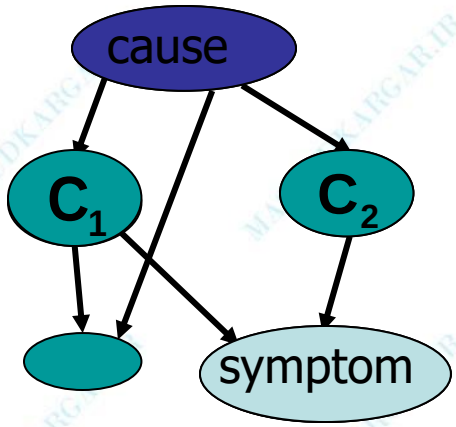
- اگرچه در هنگام استفاده از BN هم ناگزیر به استفاده از شروط استقلال خواهیم بود اما BN راه حل میانه‌تری است که محدودیت کمتری دارد.

مثال



$$P\,(Campfire=True \mid Storm=True, BusTourGroup=True\,) = ?$$

کاربرد BN



- تشخیص $P(cause|symptom)=?$
- پیش‌بینی $P(symptom|cause)=?$
- دسته‌بندی $\max_{class} P(class|data)$
- تصمیم‌گیری (در صورت وجود تابع ارزش)

- مثال:

Speech recognition, Stock market, Text Classification, Computer troubleshooting, medical diagnostic systems, real-time weapons scheduling, Intel processor fault diagnosis (Intel). generator monitoring expert system (General Electric) troubleshooting (Microsoft)

کاربرد BN

- خواستگاه BN به ترکیب احتمال با سیستم‌های خبره بر می‌گردد و این زمینه یکی از کاربردهای مهم آنرا تشکیل می‌دهد.
- BN را می‌توان در بسیاری از کاربردهائی که سیستم‌های مبتنی بر دانش متداول مورد استفاده هستند، به کار برد.
- BN در مقایسه با شبکه‌های عصبی دارای مزایای زیر است:
  - می‌توان از اطلاعات افراد خبره در ساخت BN استفاده کرد.
  - فهم و توسعه ساختار BN ساده‌تر است.
  - BN می‌تواند با داده‌های ناقص نیز کار کند.

استقلال شرطی

- تعریف استقلال شرطی
- اگر X,Y,Z سه متغیر تصادفی با مقادیر گسسته باشند می‌گوئیم که X با دانستن Z بطور شرطی از Y مستقل است اگر با دانستن Z توزیع احتمال X مستقل از مقدار Y باشد.

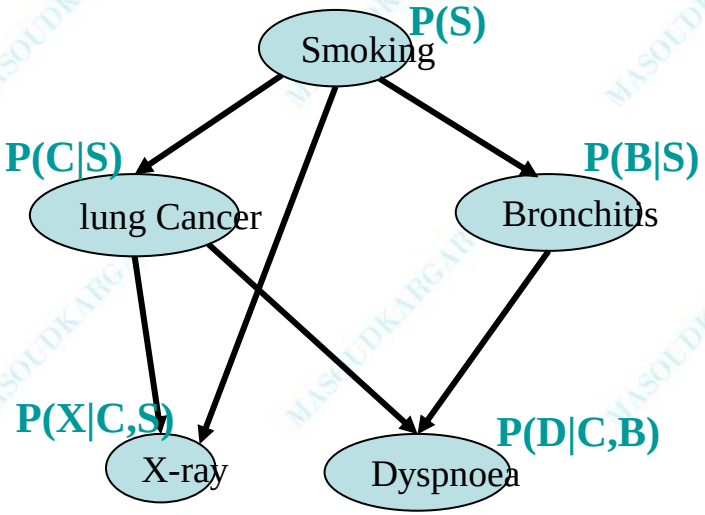
$$(\forall x_i, y_j, z_k)\, P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

- تعریف مشابهی را می‌توان برای مجموعه‌ای از متغیرها بکار برد:

$$P(X_1 \ldots X_L \mid Y_1 \ldots Y_m, Z_1 \ldots Z_n) = P(X_1 \ldots X_L \mid Z_1 \ldots Z_n)$$

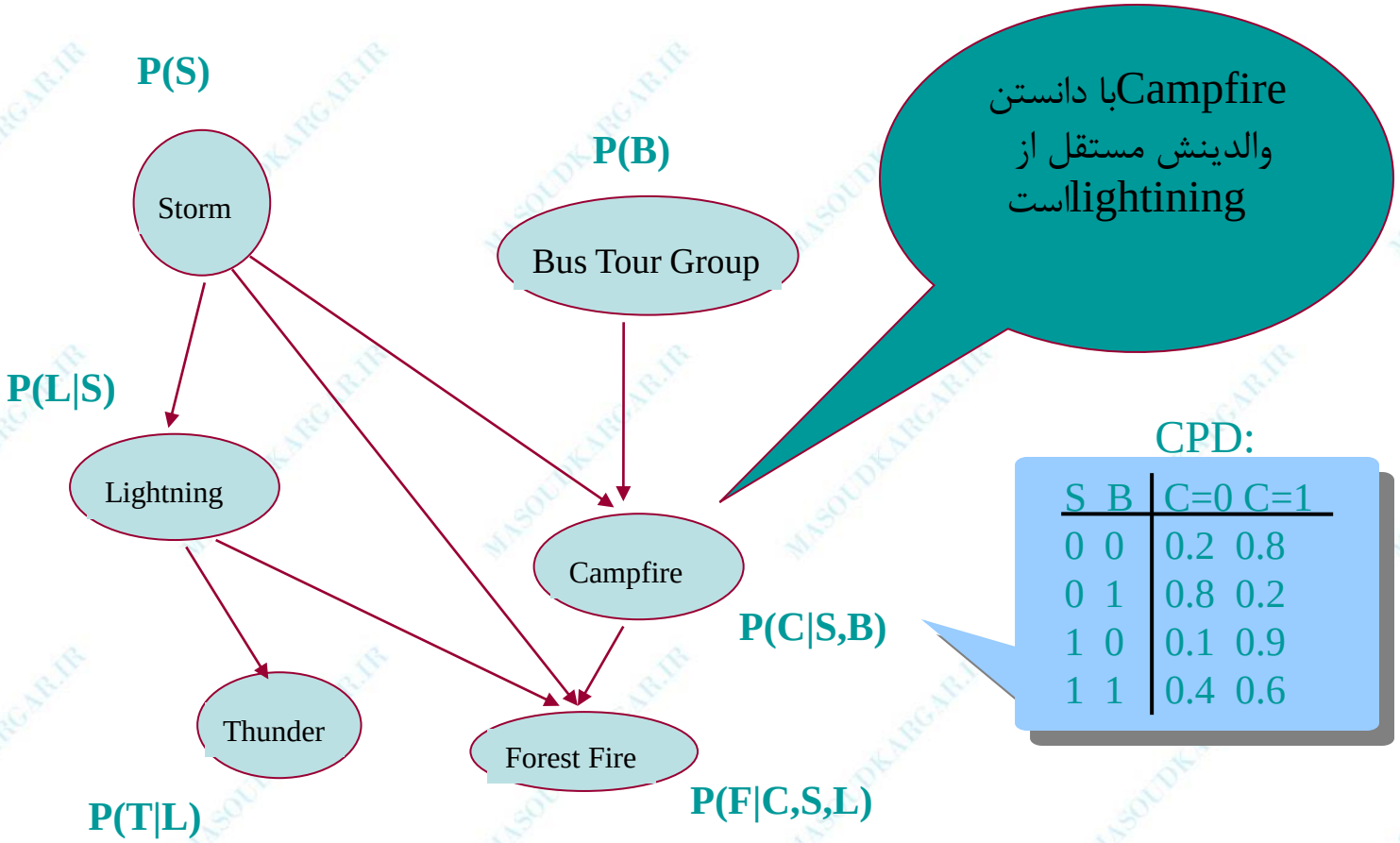
نمایش BN

- یک BN مدلی گرافیکی برای نشان دادن توزیع احتمال توام مجموعه‌ای از متغیرها است. دانش به دست آمده برای یک مسئله به صورت اطلاعات کمی و کیفی در این گراف مدل می‌شود.
- اینکار با مشخص کردن مجموعه‌ای از فرضیات استقلال شرطی توسط کمانهای گراف، همراه با ذکر مقادیر احتمال شرطی گره‌ها انجام می‌شود.
- هر متغیری از فضای توام به صورت یک گره در BN نمایش داده شده و برای هر متغیر دو نوع اطلاعات ارائه می‌گردد:



- کمانهای شبکه برای نشان دادن رابطه استقلال شرطی بکار می‌رود: یک متغیر با دانستن والدین آن از گره‌های غیر فرزند آن مستقل است.
- جدولی نیز ارائه می‌گردد که توزیع احتمال هرگره برای والدین بلافصل آنرا مشخص می‌کند .

نمایش BN



با دانستن Storm و BusTourGroup متغیرهای Lightning و Thunder اطلاعات اضافی دیگری برای Campfireارائه نمی‌دهند.

توزیع احتمال توام
joint probability distribution

- در BN برای محاسبه توزیع احتمال توام مجموعه‌ای از متغیرها از رابطه زیر استفاده می‌شود:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))$$

- که والدین یک گره، گره‌های قبلی بلافصل آن می‌باشند.
- مثال

P(Campfire, Storm, BusGroupTour, Lightning, Thunder, ForestFire)?

P(Storm)P(BusTourGroup)P(Campfire|Storm,BusTourGroup)
P(Lightning|Storm)P(Thunder|Lightning)
P(ForestFire|Lightning,Storm,Campfire).

مثال

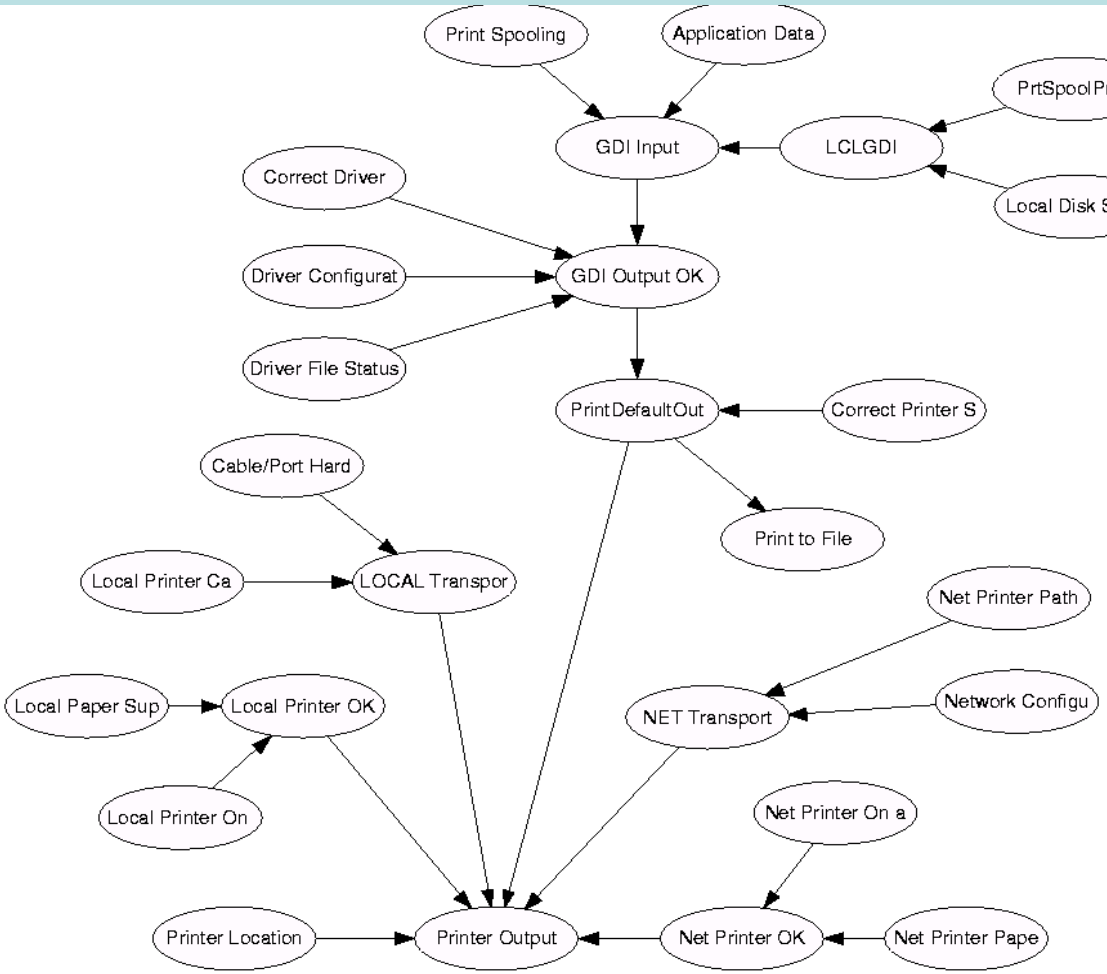
- احتمال شرطی

	S,B	*S,~B*	*~S,B*	*~S,~B*
C	*0.4*	*0.1*	*0.8*	*0.2*
~C	*0.6*	*0.9*	*0.2*	*0.8*

$$P(Campfire=true|Storm=true,BusTourGroup=true) = 0.4$$

مثال

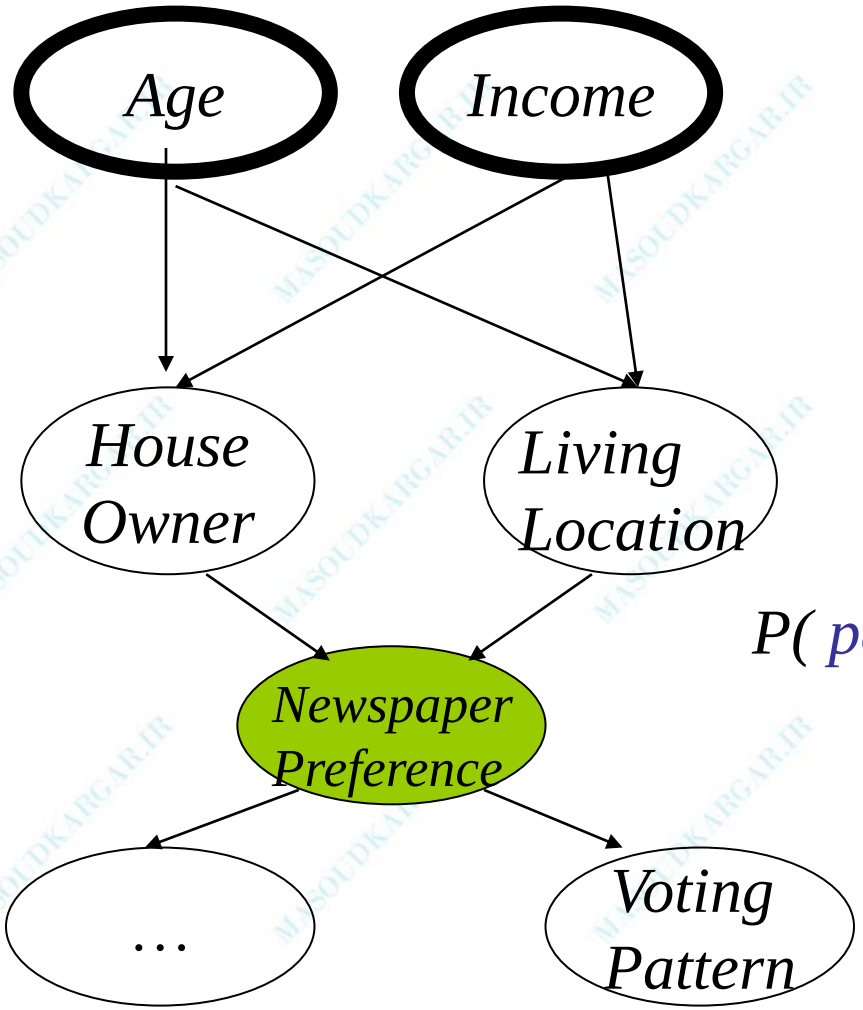
- عیب‌یابی پرینتر



نحوه ساخت BN

- به صورت دستی، توسط یک خبره.
- به صورت اتوماتیک، توسط روشهای یادگیری ماشین.

استنتاج توسط BN



How likely are elderly rich people to buy DallasNews?

$$P(\text{paper} = \text{DallasNews} \mid \text{Age}>60, \text{Income} > 60k)$$

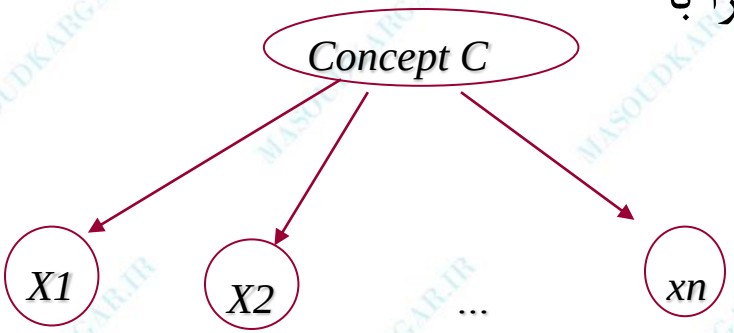
استنتاج

ارتباط بین BN و دسته‌بندی:

- می‌توان از BN استفاده نموده و مقدار یک متغیر را در صورت مشاهده مقادیر سایر متغیرها استنتاج نمود. البته معمولا امکان بدست آوردن یک مقدار وجود نداشته و به جای آن یک توزیع احتمال محاسبه می‌شود.
- اگر مقادیر همه متغیرها از پیش معلوم باشد انجام چنین استنتاجی ساده است ولی معمولا فقط مقدار بخشی از متغیرها مشاهده می‌شود. مثلا ممکن است بخواهیم با مشاهده Thunder , BusTourGroup در مورد Forestfire نتیجه‌گیری کنیم.

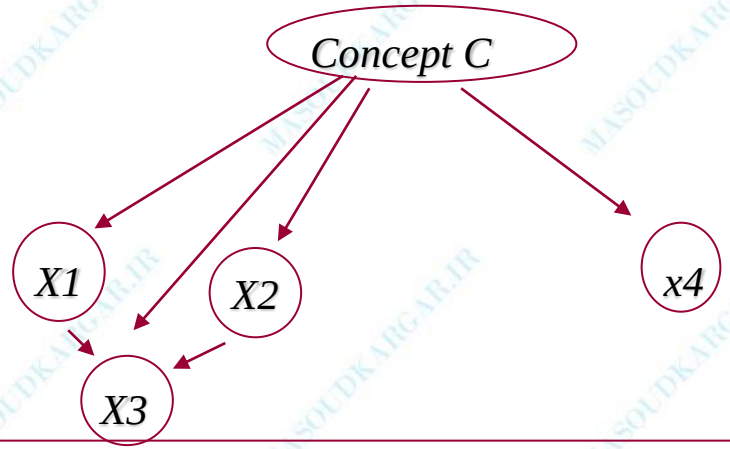
حالت کلی

اگر یکی از متغیرها مقدار هدف بوده و بخواهیم آنرا با داشتن مقدار سایر متغیرها محاسبه کنیم



$$P(x1,x2,...xn,c) = P(c)\ P(x1|c)\ P(x2|c)\ ...\ P(xn|c)$$

دسته‌بندی‌کننده بیزی ساده



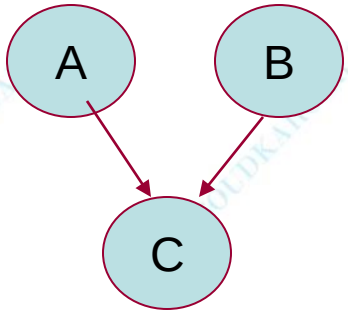
شبکه باور بیزی

$$P(x1,x2,...xn,c) = P(c)\ P(x1|c)\ P(x2|c)\ P(x3|x1,x2,c)P(x4,c)$$

استنتاج

- روشهای مختلف:
  - روشهای دقیق (NP-Hard)
  - روشهای تقریبی Monte Carlo
  - Dynamic programming
  - Variable elimination

مثال



CPD:

A	B	C=0	C=1
0	0	0.5	0.5
0	1	0.5	0.5
1	0	0.6	0.4
1	1	0.8	0.2

با دانستن احتمال A , B می‌توان احتمال درستی C را محاسبه نمود.

***A**: p(A) = 0.1 p(~A) = 0.9*

***B**: p(B) = 0.4 p(~B) = 0.6*

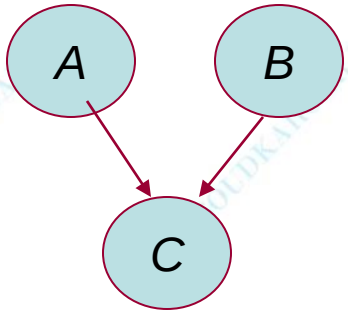
*The **initialized** Probability of C:*

$$p(C) = p(CAB) + p(C\sim AB) + p(CA\sim B) + p(C\sim A\sim B)$$

$$= p(C \mid AB) * p(AB) + p(C \mid \sim AB) * p(\sim AB) + p(C \mid A\sim B) * p(A\sim B) + p(C \mid \sim A\sim B) * p(\sim A\sim B)$$

$$= p(C \mid AB) * p(A) * p(B) + p(C \mid \sim AB) * p(\sim A) * p(B) + p(C \mid A\sim B) * p(A) * p(\sim B) + p(C \mid \sim A\sim B) * p(\sim A) * p(\sim B) = ۰.۵۱۸$$

مثال



CPD:

A	B	C=0	C=1
0	0	0.5	0.5
0	1	0.5	0.5
1	0	0.6	0.4
1	1	0.8	0.2

در صورتیکه بدانیم C درست است می‌توان با استفاده از تئوری بیز و احتمال اولیه C احتمال اینکه کدامیک از A یا B علت وقوع آن بوده را محاسبه نمائیم.

A: $p(A) = 0.1 \quad p(\sim A) = 0.9$

B: $p(B) = 0.4 \quad p(\sim B) = 0.6$

C: $p(c) = 0.518$

$$p(B \mid C) = (p(C \mid B) * p(B)) / p(C) = ((p(C \mid AB) * p(A) + p(C \mid \sim AB) * p(\sim A)) * p(B)) / p(C) = ((0.8 * 0.1 + 0.5 * 0.9) * 0.4) / 0.518 = 0.409$$

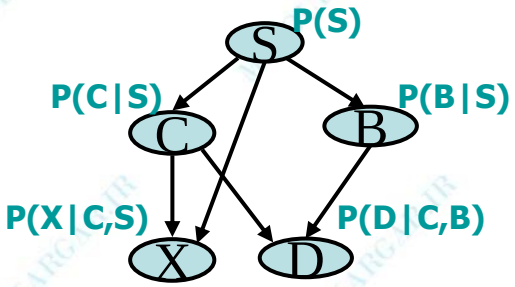
$$p(A \mid C) = (p(C \mid A) * p(A)) / p(C) = ((p(C \mid AB) * p(B) + p(C \mid A\sim B) * p(\sim B)) * p(A)) / p(C) = ((0.8 * 0.4 + 0.6 * 0.6) * 0.1) / 0.518 = 0.131$$

لذا در صورت صحیح بودن C می‌توان چنین گفت که احتمال اینکه B عامل آن بوده باشد بیشتر است.

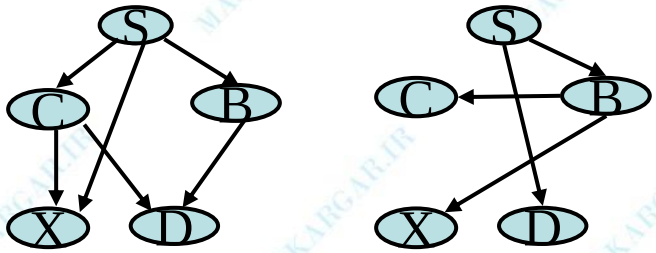
یادگیری یک BN

چگونه می‌توان یک BN را از روی داده‌های آموزشی یاد گرفت؟

1. ممکن است ساختار شبکه از قبل معلوم باشد،
   و بخواهیم پارامترهای آن یعنی مقادیر احتمال شرطی متغیرها را یاد بگیریم



2. ممکن است ساختار شبکه از قبل معلوم باشد،
   ولی فقط برخی از متغیرها (در مثالهای آموزشی) قابل مشاهده باشند.



3. ممکن است ساختار شبکه معلوم نبوده و
   مجبور باشیم ساختار گراف و پارامترهای آنرا از داده‌های آموزشی استنتاج کنیم.

$$\hat{G} = \arg\max_{G} Score(G)$$

یادگیری یک BN

1. در صورتی که ساختار شبکه معلوم بوده و تمامی متغیرها قابل مشاهده باشند می‌توان براحتی جدول احتمال شرطی را از روی داده‌های آموزشی یاد گرفت (مشابه یادگیری بیزی ساده).

2. در صورتیکه ساختار شبکه از قبل معلوم بوده ولی فقط برخی از مقادیر متغیرها قابل مشاهده باشند، یادگیری مشکل‌تر خواهد بود. اینکار شبیه یادگیری وزنهای لایه مخفی شبکه‌های عصبی است.

3. در صورتیکه ساختار شبکه معلوم نباشد یادگیری مشکل بوده و از روشهای جستجوئی نظیر K2 برای جستجو در فضای ساختارهای ممکن استفاده می‌شود.

روشهای یادگیری BN

- روشهای یادگیری BN بسته به اینکه ساختار شبکه از قبل معلوم باشد و همچنین قابلیت مشاهده داده به صورت زیر است:

Structure	***Observability***	***Method***
Known	***Full***	***Maximum Likelihood Estimation***
Known	***Partial***	***EM (or gradient ascent)***
Unknown	***Full***	***Search through model space***
Unknown	***Partial***	***EM + search through model space***

روشهای یادگیری ساختار BN

1. روشهای Scoring-based که در آنها بهترین BN ساختاری است که بیشترین تطابق با داده را داشته باشد. این روشها بدنبال ماکزیمم کردن MDL(Minimum Description Length) بیزی و یا تابع آنتروپی KL (Kullback-Leibler) هستند.

2. روشهای Constraint-based که در آن ساختار BN با مشخص کردن رابطه استقلال شرطی بین گره‌ها بدست می‌آید.

- نشان داده‌اند که روش اول کارائی کمتری داشته و دسته‌بندی‌کننده ضعیفی را بدست می‌دهد.

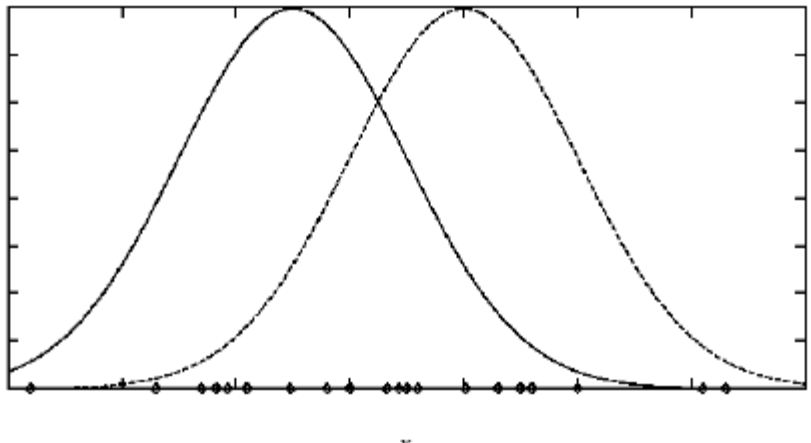
الگوریتم EM : یادگیری با داده‌های غیرقابل مشاهده

- در برخی از کاربردها فقط برخی از ویژگیها قابل مشاهده هستند. راههای متفاوتی برای یادگیری در چنین مواقعی ارائه شده است.
- در شرایطی که شکل کلی توزیع احتمال متغیری معلوم باشد، می‌توان از الگوریتم EM (Expectation-Maximization)برای یادگیری متغیری استفاده نمود که بطور کامل مشاهده نشده است.
- این الگوریتم برای آموزش BN بکار می‌رود. علاوه بر آن در الگوریتمهای دسته‌بندی بدون ناظر و یادگیری HMM نیز کاربرد دارد.

تخمین میانگین k تابع گاوسی

مثال

- یک مجموعه آموزشی D را در نظر بگیرید که دارای نمونه‌هائی باشد که مخلوطی از k توزیع نرمال مجزا باشد. در شکل زیر $k=2$ است.



- برای تولید هر نمونه:

1. یکی از k توزیع احتمال گاوسی با احتمال یکنواختی انتخاب می‌شود.
2. داده به صورت تصادفی و با استفاده از توزیع فوق انتخاب می‌گردد.

الگوریتم EM برای تخمین میانگین k تابع گاوسی

- ورودی:

نمونه‌های x که به صورت مخلوطی از k توزیع گاوسی درست شده است

- اطلاعاتی که نداریم:
  - مقادیر k میانگین این توزیعها $<\mu_1,\ldots \mu_k>$ (واریانس همه توزیعها برابر و معلوم است.)
  - اینکه کدام نمونه توسط کدام توزیع تولید شده است. (اطلاعات ناقص)
- هدف:

یافتن مقدار فرضیه ML (حداکثر شباهت از روی حداکثرکردن $p(D|h)$)
برای تخمین هر یک از مقادیر میانگین $h=<\mu_1,\ldots \mu_k>$

اطلاعات ناقص

- اگر نمایش کامل هر نمونه را به صورت زیر نشان دهیم:

$$Y_i = \langle x_i, z_{i1}, z_{i2} \rangle$$

که در آن z_{ij} برابر با یک است اگر نمونه x_i توسط توزیع jام درست شده باشد.

x_i نمونه مشاهده شده است و z_{i1}, z_{i2} متغیرهای مخفی هستند.

– سوال: اگر نمونه‌های کامل (حاوی z_{i1}, z_{i2}) را داشتیم چگونه می‌شد $\mu_1, \ldots \mu_k$ را محاسبه کرد؟

$$\mu_{ML} = \arg\min_{\mu} \sum_{i=1}^{m} (x_i - \mu) \quad , \quad \mu_{ML} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

تخمین پارامتر

EM for Estimating k Means

EM Algorithm: Pick random initial $h = \langle \mu_1, \mu_2 \rangle$, then iterate

- E step: Calculate the expected value $E[z_{ij}]$ of each hidden variable z_{ij}, assuming the current hypothesis $h = \langle \mu_1, \mu_2 \rangle$ holds.
- M step: Calculate a new maximum likelihood hypothesis $h' = \langle \mu_1', \mu_2' \rangle$, assuming the value taken on by each hidden variable z_{ij} is its expected value $E[z_{ij}]$ calculated above. Replace $h = \langle \mu_1, \mu_2 \rangle$ by $h' = \langle \mu_1', \mu_2' \rangle$.

تخمین پارامتر

E Step For *k* Means

$$E[z_{ij}] = p(x=x_i|\mu=\mu_j) \; / \; \Sigma_{n=1}^{2} p(x=x_i|\mu=\mu_n)$$

$$p(x=x_i|\mu=\mu_j) = \exp(-1/(2\sigma^2)(x_i - \mu_j)^2)$$

Derived via PDF for Gaussians and Bayes rule