5.6 Bayesovská klasifikace

Metody bayesovské klasifikace vycházejí z Bayesovy věty o podmíněných pravděpodobnostech. Ačkoliv se tedy jedná o metody pravděpodobnostní, jsou intenzivně studovány v souvislosti se strojovým učením a uplatňují se rovněž v systémech pro dobývání znalostí.

5.6.1 Základní pojmy

Bayesův vztah pro výpočet podmíněné pravděpodobnosti že platí hypotéza H, pokud pozorujeme evidenci E má podobu:

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

Apriorní pravděpodobnost hypotézy $P(H)$ odpovídá znalostem o zastoupení jednotlivých hypotéz (tříd) bez ohledu na nějaké další informace. Podmíněná pravděpodobnost $P(H \mid E)$, též nazývaná *aposteriorní*, vyjadřuje, jak se změní pravděpodobnost hypotézy, pokud víme, že nastalo E. $P(E)$ vyjadřuje pravděpodobnost evidence (pozorování).

Hypotéz mezi kterými se rozhodujeme bývá obvykle více[1]. Nás bude zajímat pro danou evidenci ta nejpravděpodobnější. Pro každou hypotézu H_t $t=1,...,T$ můžeme spočítat $P(H_t \mid E)$ a z nich vybrat hypotézu H_{MAP}, která má největší aposteriorní pravděpodobnost (maximum aposteriori probability)

$$H_{MAP} = H_J \text{ právě když } P(H_J \mid E) = \max_t \frac{P(E \mid H_t) \times P(H_t)}{P(E)} .$$

Vzhledem k tomu, že nás zajímá pouze to, pro které H_t je hodnota aposteriorní pravděpodobnosti maximální, ale už nás nezajímá konkrétní hodnota, můžeme výpočet poněkud zjednodušit zanedbáním jmenovatele.

$$H_{MAP} = H_J \text{ právě když } P(E \mid H_J) \times P(H_J) = \max_t (P(E \mid H_t) \times P(H_t))$$

Ve zanedbávání můžeme jít ještě dále tak, že budeme předpokládat, že všechny hypotézy jsou stejně pravděpodobné (a tedy že na $P(H)$ nezáleží). Nalezneme tak hypotézu, která má největší věrohodnost (maximum likelihood):

$$H_{ML} = H_J \text{ právě když } P(E \mid H_J) = \max_t P(E \mid H_t)) .$$

Pro ilustraci tohoto postupu vezměme opět úlohu poskytování úvěru, tentokrát ale pouze na základě výše příjmu. Přepokládejme, že banka vyhoví u 2/3 žádosti o úvěr; tedy apriorní pravděpodobnosti budou P(půjčit)=0.667 a P(nepůjčit)=0.333. Dále předpokládejme, že vysoký příjem mělo 91% klientů, kterým banka půjčila a nízký příjem mělo 88% klientů, kterým banka nepůjčila. Tedy

$$P(\text{vysoký\_příjem} \mid \text{půjčit}) = 0.91 \qquad P(\text{nízký\_příjem} \mid \text{půjčit}) = 0.09$$

$$P(\text{vysoký\_příjem} \mid \text{nepůjčit}) = 0.12 \qquad P(\text{nízký\_příjem} \mid \text{nepůjčit}) = 0.88.$$

[1] V tomto případě se $P(E)$ ve jmenovateli Bayesova vztahu obvykle vyjadřuje jako $\sum_t P(E \mid H_t) P(H_t)$.

Předpokádejme, že posuzujeme klienta s vysokým příjmem. Bude větší pravděpodobnost, že banka půjčí nebo že nepůjčí? Podle Bayesovy věty spočítáme

$$P(\text{vysoký\_příjem} \mid \text{půjčit}) \times P(\text{půjčit}) = 0.607$$

$$P(\text{vysoký\_příjem} \mid \text{nepůjčit}) \times P(\text{nepůjčit}) = 0.040$$

Tedy H_{MAP} = půjčit. Víme tedy, která hypotéza je pravděpodobnější, přestože jsme přímo nespočítali aposteriorní pravděpodobnosti obou hypotéz. Tyto pravděpodobnosti získáme, pokud budeme uvedené hodnoty normovat tak, aby jejich součet byl roven 1.

Výhodou bayesovských metod je právě tato schopnost klasifikovat příklady do tříd s určitou pravděpodobností. Tuto pravděpodobnost můžeme interpretovat jako spolehlivost rozhodnutí.

S Bayesovým teorémem úzce souvisí tzv. *princip minimální deskripční délky* (minimum description length principle, MDL [Rissanen, 1978]). Tento princip interpretuje definici hypotézy s největší aposteriorní pravděpodobností ve světle teorie informace. Původní formulaci nalezení H_{MAP}

$$H_{MAP} = H_J \text{ právě když } P(E \mid H_J) \times P(H_J) = \max_t (P(E \mid H_t) \times P(H_t))$$

lze ekvivalentně zapsat jako

$$H_{MAP} = H_J \text{ právě když } \log_2 P(E \mid H_J) + \log_2 P(H_J) = \max_t (\log_2 P(E \mid H_t) + \log_2 P(H_t))$$

resp.

$$H_{MAP} = H_J \text{ právě když } -\log_2 P(E \mid H_J) - \log_2 P(H_J) = \min_t (-\log_2 P(E \mid H_t) - \log_2 P(H_t))$$

Z teorie informace je známo, že množství informace obsažené ve zprávě, která má pravděpodobnost výskytu p je $-\log_2 p$. S množstvím informace úzce souvisí délka kódu potřebná pro přenos. Můžeme tedy výraz $-\log_2 P(H_t)$ chápat jako vyjádření délky optimálního zakódování hypotézy H_t a výraz $-\log_2 P(E \mid H_t)$ jako vyjádření délky optimálního zakódování evidence E, platí-li H_t. V tomto smyslu je H_{MAP} příkladem použití principu MDL, který preferuje model M minimalizující délku $L(D)$ kódu potřebného pro přenos dat D. Zakódování dat je přitom rozděleno do dvou částí: zakódování modelu (kódem délky $L(M)$) a zakódování dat, která jsou modelem chybně klasifikována[2] (kódem délky $L(err \mid M)$):

$$M_{MDL} = M_J \text{ právě když } L(M_J) + L(\text{err} \mid M_J) = \min_i (L(M_i) + L(\text{err} \mid M_i))$$

Bayesova věta dává návod jak stanovit vliv jedné evidence na uvažovanou hypotézu. Jak ale postupovat, pokud je evidencí více? Tedy, jak stanovit aposteriorní pravděpodobnost $P(H \mid E_1, \ldots, E_K)$? Následující dvě podkapitoly ukazují dva možné přístupy.

5.6.2 Naivní bayesovský klasifikátor

Naivní bayesovský klasifikátor vychází z předpokladu, že jednotlivé evidence $E_1, \ldots, E_K$ jsou podmíněně nezávislé při platnosti hypotézy H [Duda, Hart, 1973]. Tento zjednodušující předpoklad umožňuje spočítat aposteriorní pravděpodobnost hypotézy při platnosti všech evidencí

$$P(H \mid E_1, \ldots, E_K) = \frac{P(E_1, \ldots, E_K \mid H) \times P(H)}{P(E_1, \ldots, E_K)}$$

jako

[2] Data klasifikovaná správně nemusíme přenášet, pro ně nám stačí model.

$$P(H \mid E_1,\ldots,E_K) = \frac{P(H)}{P(E_1,\ldots,E_K)} \times \prod_{k=1}^{K} P(E_k \mid H) .$$

V případě klasifikace pomocí naivního bayesovského klasifikátoru tedy budeme hledat hypotézu s největší aposteriorní pravděpodobností H_{MAP}

$$H_{MAP} = H_J \text{ právě když } P(H_j) \times \prod_{k=1}^{K} P(E_k \mid H_j) = \max_t \left(P(H_t) \times \prod_{k=1}^{K} P(E_k \mid H_t)\right)$$

Abychom mohli tento způsob klasifikace použít, potřebujeme znát hodnoty $P(H_t)$ a $P(E_k \mid H_t)$. V kontextu dobývání znalostí z databází můžeme tyto hodnoty získat z trénovacích dat ve fázi učení [3]. Evidence E_k jsou pak hodnoty jednotlivých vstupních atributů (tedy $E_k = A_j(v_k)$) a hypotézy H_t jsou cílové třídy (tedy $H_t = C(v_t)$).

Na rozdíl od jiných algoritmů učení (rozhodovací stromy, rozhodovací pravidla) se zde neprovádí prohledávání prostoru hypotéz. Stačí jen spočítat příslušné pravděpodobnosti na základě četnosti výskytů hodnot jednotlivých atributů [4]. Tedy pravděpodobnosti $P(H_t)$ a $P(E_k \mid H_t)$ lze spočítat jako

$$P(H_t) = P(C(v_t)) = \frac{n_t}{n}$$

$$P(E_k \mid H_t) = P(A_j(v_k) \mid C(v_t)) = \frac{n_t(A_j(v_k))}{n_t}$$

klient	příjem	konto	pohlaví	nezaměstnaný	úvěr
k1	vysoký	vysoké	žena	ne	ano
k2	vysoký	vysoké	muž	ne	ano
k3	nízký	nízké	muž	ne	ne
k4	nízký	vysoké	žena	ano	ano
k5	nízký	vysoké	muž	ano	ano
k6	nízký	nízké	žena	ano	ne
k7	vysoký	nízké	muž	ne	ano
k8	vysoký	nízké	žena	ano	ano
k9	nízký	střední	muž	ano	ne
k10	vysoký	střední	žena	ne	ano
k11	nízký	střední	žena	ano	ne
k12	nízký	střední	muž	ne	ano

Tab. 1 Trénovací data pro bayesovský klasifikátor

Pro náš příklad trénovacích dat z Tab. 1 budou apriorní pravděpodobnosti různých hodnot cílového atributu *úvěr*:

$$P(\text{úvěr(ano)}) = 8/12 = 0.667$$

$$P(\text{úvěr(ne)}) = 4/12 = 0.333$$

Podobně spočítáme podmíněné pravděpodobnosti $P(A_j(v_k) \mid C(v_t))$, např.

[3] Jedná se tedy opět o učení s učitelem.

[4] Vzhledem k tomu se bayesovský klasifikátor hodí pro veliké datové soubory.

P(konto(střední) | úvěr(ano)) = 2/8 = 0.25

P(konto(střední) | úvěr(ne)) = 2/4 = 0.5

P(nezaměstnaný(ne) | úvěr(ano)) = 5/8 = 0.625

P(nezaměstnaný(ne) | úvěr(ne)) = 1/4= 0.25

. . .

Pro uchazeče o úvěr, který má střední konto a není nezaměstnaný spočítáme

P(úvěr(ano)) P(konto(střední) | úvěr(ano)) P(nezaměstnaný(ne) | úvěr(ano)) = 0.1042

P(úvěr(ne)) P(konto(střední) | úvěr(ne)) P(nezaměstnaný(ne) | úvěr(ne)) = 0.0416

Tedy naivní bayesovský klasifikátor zařadí tohoto uchazeče do třídy *úvěr(ano)*. Můžeme si všimnout, že jsme úspěšně klasifikovali neúplně popsaný případ, který by zůstal nezařazen dříve vytvořenými rozhodovacími stromy i pravidly.

Výše uvedený způsob výpočtu pravděpodobnosti $P(A_j(v_k) | C(v_t))$ má některé nevýhody:

- pokud se v trénovacích datech neobjeví pro danou třídu $C(v_t)$ hodnota v_k atributu A_j, bude odpovídající pravděpodobnost $P(A_j(v_k) | C(v_t))$ rovna 0 a tedy bude nulová i aposteriorní pravděpodobnost této třídy bez ohledu na hodnoty ostatních atributů,
- pokud je četnost $n_t(A_j(v_k))$ vzájemného výskytu $A_j(v_k)$ a $C(v_t)$ malá, je spočítaná hodnota podhodnocením skutečné pravděpodobnosti $P(A_j(v_k) | C(v_t))$.

Proto se při výpočtu $P(A_j(v_k) | C(v_t))$ provádějí různé korekce [5]:

- Laplaceova korekce

$$\frac{n_t(A_j(v_k)) + 1}{n_t + T}$$

- m-odhad

$$\frac{n_t(A_j(v_k)) + m*f_k}{n_t + m}$$

kde T je počet tříd, $f_k = n(A_j(v_k))/n$ je relativní četnost hodnoty $A_j(v_k)$ [6] a m je parametr.

Přestože předpoklad podmíněné nezávislosti bývá v reálných úlohách jen málokdy splněn [7], vykazuje naivní bayesovský klasifikátor překvapivě dobré vlastnosti ve smyslu úspěšnosti klasifikace [8]. To z něj činí velmi oblíbený nastroj. Druhým důvodem obliby bayesovského klasifikátoru je jeho snadné vytvoření. Jistou nevýhodou je tedy (z pohledu nezkušeného uživatele) pouze o něco menší srozumitelnost reprezentace znalostí pomocí pravděpodobností.

[5] Na tyto korekce jsme již narazili v souvislosti se systémy *CN4* a *KEX*.

[6] V případě nedostatku informací lze předpokládat rovnoměrné rozdělení hodnot atributu A_j.

[7] Tento málokdy splněný předpoklad je důvodem, proč je klasifikátor nazýván „naivní".

[8] Domingos a Pazzani to vysvětlují tím, že nehledáme přesné hodnoty aposteriorních pravděpodobností ale jen hodnotu největší [Domingos, Pazzani, 1996]. Navíc pro klasifikaci používáme mnohem více atributů než je potřeba a pracujeme tedy se značnou redundancí.

5.6.3 Bayesovské sítě

Pravděpodobnostně korektní způsob, jak se v plné šíři vypořádat se skutečností, že evidence nejsou navzájem nezávislé, je uvažovat pravděpodobnosti výskytu všech možných kombinací (jednočlenných, dvojčlenných, …) hodnot všech evidencí. Tento postup, ač teoreticky zcela jasný, bývá v praktických úlohách nerealizovatelný. Pro n binárních atributů bychom potřebovali znát 2^n hodnot pravděpodobností. Naštěstí existuje jiná, pravděpodobnostně korektní cesta, známá jako bayesovské sítě.

5.6.3.1 Reprezentace znalostí

Bayesovské sítě (též nazývané pravděpodobnostní sítě) umožňují reprezentovat znalosti[9] o částečně nezávislých evidencích a tyto znalosti použít při usuzování. Základním pojmem, se kterým operují bayesovské sítě je podmíněná nezávislost. Veličiny A a B jsou podmíněně nezávislé při dané veličině C jestliže

$$P(A,B \mid C) = P(A \mid C)\, P(B \mid C).$$

Tomu ekvivalentní jsou vztahy ([Jiroušek, 1995])

$$P(A \mid B,C) = P(A \mid C)$$

$$P(B \mid A,C) = P(B \mid C).$$

Podmíněná nezávislost A a B při daném C se obvykle značí

$$A \perp\!\!\!\perp B \mid C$$

Bayesovská síť je acyklický orientovaný graf zachycující pomocí hran pravděpodobnostní závislosti mezi náhodnými veličinami. Ke každému uzlu u (náhodné veličině) je přiřazena pravděpodobnostní distribuce tvaru $P(u \mid rodiče(u))$, kde $rodiče(u)$ jsou uzly, ze kterých vycházejí hrany do uzlu u. Uspořádejme (a očíslujme) všechny uzly sítě tak, že rodiče jsou před svými dětmi (mají nižší pořadové číslo). Potom pro každý uzel u_i platí, že je podmíněně nezávislý na všech uzlech s nižším pořadovým číslem s výjimkou svých rodičů podmíněno svými rodiči[10]:

$$u_i \perp\!\!\!\perp \{u_k\}_{k=1,\ldots,i-1} \setminus \text{rodiče}(u_i) \mid \text{rodiče}(u_i).$$

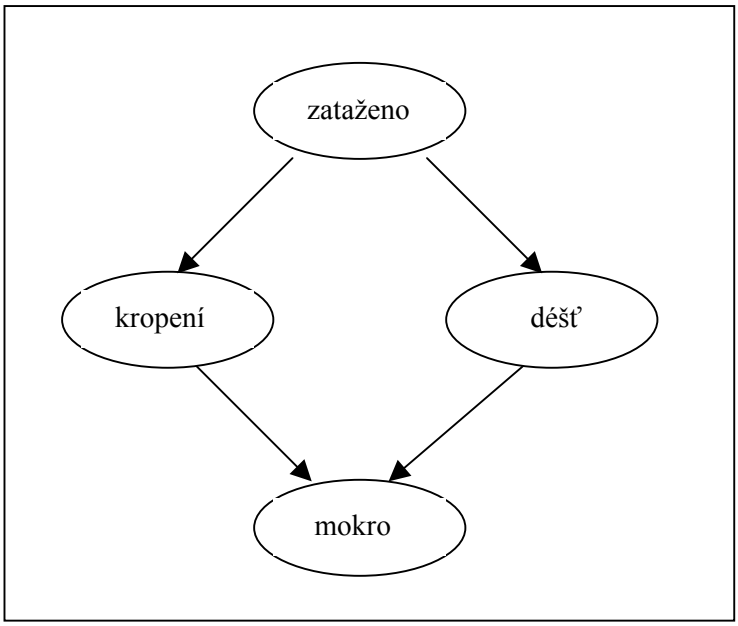
To umožňuje spočítat sdruženou pravděpodobnostní distribuci celé sítě jako [11]

$$P(u_1,\ldots,u_n) = \prod_{i=1}^{n} P(u_i \mid \text{rodiče}(u_i))$$

[9] Bayesovské sítě stojí tak říkajíc na půl cesty mezi pravidly a neuronovými sítěmi.

[10] Jiná formulace říká, že při daných rodičích, dětech a dětech rodičů je uzel podmíněně nezávislý na všech ostatních uzlech sítě.

[11] Uvedenému vztahu se říká faktorizace sdružené distribuce (Lauritzen) nebo chain rule (Jensen). Místo 2^n hodnot nyní vystačíme s $n \times 2^k$ hodnotami, kde k je max. počet „vstupů“ do uzlu v síti.



Obr. 1 Bayesovská síť

Má-li tedy bayesovská síť podobu uvedenou na Obr. 1[12], bude mít sdružená distribuce tvar

$$P(Z,K,D,M) = P(Z)\,P(K|Z)\,P(D|Z)\,P(M|K,D)$$

K jejímu výpočtu tedy potřebujeme znát čtyři dílčí podmíněné pravděpodobnostní distribuce. V případě diskrétních veličin vystačíme s tabulkami podmíněných pravděpodobností, tak jak je uvedeno v Tab. 2.

Z	**P(K=0)**	**P(K=1)**
0	0.5	0.5
1	0.9	0.1

P(K|Z)

P(Z=0)	**P(Z=1)**
0.5	0.5

P(Z)

Z	**P(D=0)**	**P(D=1)**
0	0.8	0.2
1	0.2	0.8

P(D|Z)

K	**D**	**P(M=0)**	**P(M=1)**
0	**0**	1.0	0.0
1	**0**	0.1	0.9
0	**1**	0.1	0.9
1	**1**	0.01	0.99

P(M|K,D)

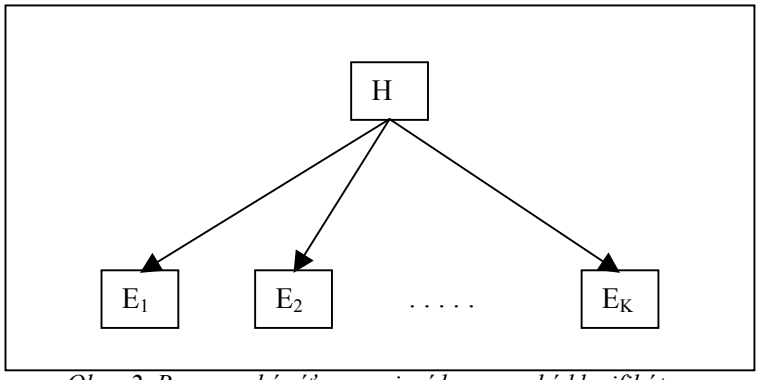
Tab. 2 Podmíněné pravděpodobnosti uzlů

Pomocí bayesovské sítě můžeme reprezentovat i naivní bayesovský klasifikátor. Vzhledem k předpokládané nezávislosti vstupních atributů bude odpovídající síť obsahovat jeden (cílový) uzel

[12] Příklad je převzat z [Murphy,2000]. K hlubšímu studiu jsou vhodné knihy [Jensen, 1996] a [Lauritzen, 1996].

který bude rodičem všech ostatních (vstupních) uzlů. Vstupní uzly nebudou navzájem spojeny hranami (Obr. 2). Sdružená distribuce této sítě bude tedy

$$P(H,E_1,E_2,\ldots,E_K) = P(H)\,P(E_1|H)\,P(E_2|H)\,\ldots.\,P(E_k|H)$$



Obr. 2 Bayesovská síť pro naivní bayesovský klasifikátor

5.6.3.2 Inference

Bayesovská síť se používá pro pravděpodobnostní odvozování (inferenci). Známe-li strukturu sítě (Obr. 1) a podmíněné pravděpodobnostní distribuce přiřazené k jednotlivým uzlům (Tab. 2), můžeme spočítat aposteriorní pravděpodobnost libovolného uzlu.

Řekněme, že pozorujeme, že je mokro, a zajímá nás, co je příčinou. Budeme tedy provádět *diagnostickou inferenci*, někdy též nazývanou "zdola nahoru" (od projevů nějakého jevu k jeho příčinám). Podle Bayesova vzorce a za použití faktorizace sdružené distribuce spočítáme

$$P(K=1|M=1) = \frac{P(K=1,M=1)}{P(M=1)} = \frac{\sum_{z,d} P(Z=z,K=1,D=d,M=1)}{P(M=1)} =$$

$$= \frac{\sum_{z,d} (P(Z=z)\,P(K=1|Z=z)\,P(D=d|Z=z)\,P(M=1|K=1,D=d))}{P(M=1)} = \frac{0.2781}{P(M=1)}$$

$$P(D=1|M=1) = \frac{P(D=1,M=1)}{P(M=1)} = \frac{\sum_{z,k} P(Z=z,K=k,D=1,M=1)}{P(M=1)} =$$

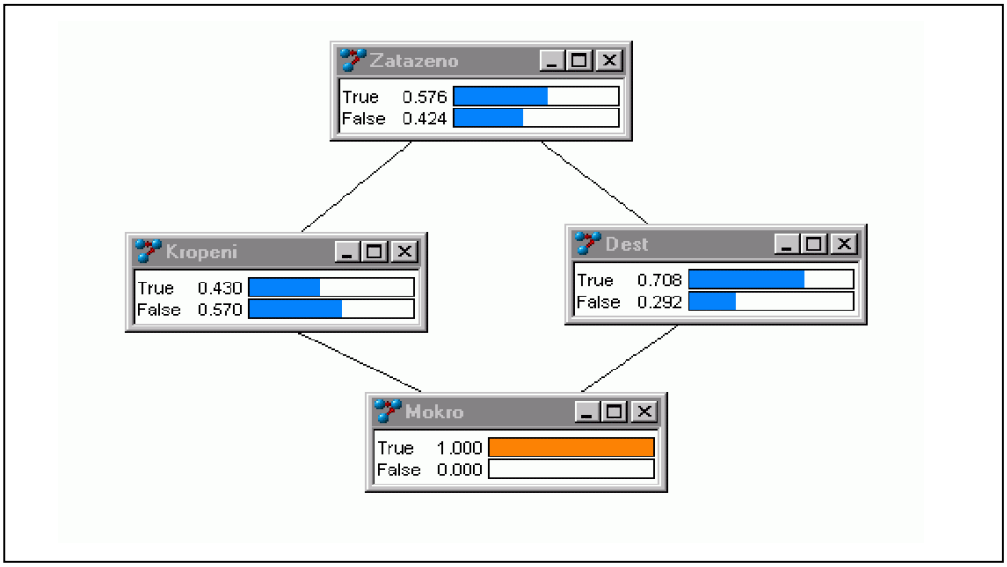
$$= \frac{\sum_{z,k} (P(Z=z)\,P(K=k|Z=z)\,P(D=1|Z=z)\,P(M=1|K=k,D=1))}{P(M=1)} = \frac{0.4581}{P(M=1)}$$

kde

$$P(M=1) = \sum_{z,k,d} (P(Z=z)\,P(K=k|Z=z)\,P(D=d|Z=z)\,P(M=1|K=k,D=d)) = 0.6471$$ [13]

Maximální aposteriorní pravděpodobnost má tedy příčina déšť; $P(D=1|M=1) > P(K=1|M=1)$. Obr. 3 ukazuje aposteriorní pravděpodobnostní distribuce všech uzlů v síti, odvozené pro zadanou pravděpodobnost $P(M=1) = 1$.

[13] Pro nalezení hypotézy s maximální aposteriorní pravděpodobností tuto hodnotu vlastně nepotřebujeme.



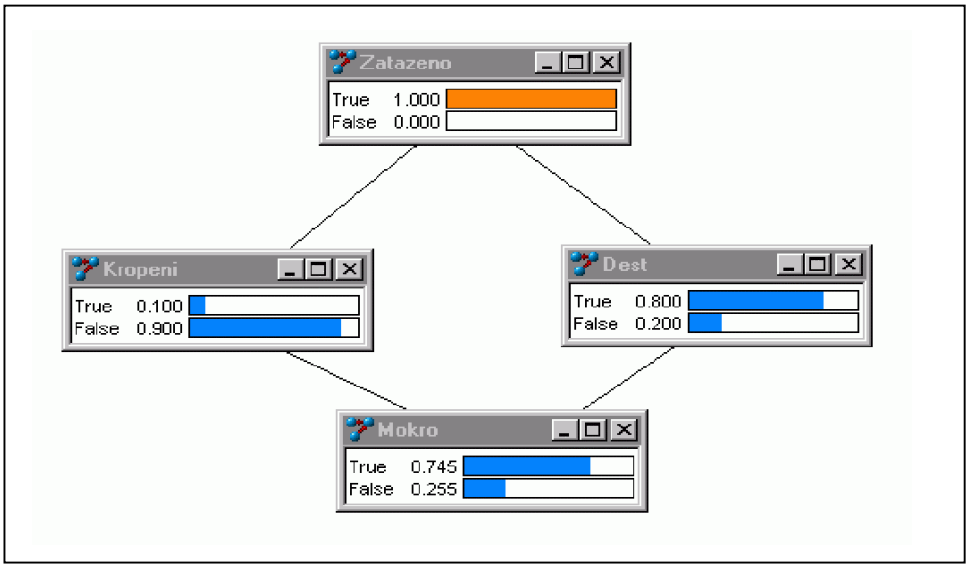
Obr. 3 Diagnostická inference (systém BKD)

Z naší sítě můžeme rovněž spočítat pravděpodobnost toho, že bude mokro, pokud vidíme, že je zataženo (Obr. 4). V tomto případě se jedná o *kauzální inferenci*, kdy postupujeme „shora dolů“ (od příčin k důsledkům).

$$P(M=1\,|\,Z=1) = \frac{P(M=1,Z=1)}{P(Z=1)} = \frac{\sum_{k,d} P(Z=1,K=k,D=d,M=1)}{P(Z=1)} =$$

$$= \frac{P(Z=1)\sum_{k,d}\left(P(K=k\,|\,Z=1)\,P(D=d\,|\,Z=1)\,P(M=1\,|\,K=k,D=d)\right)}{P(Z=1)} = 0.7452$$

Přisuzovat kauzální interpretaci bayesovské síti je však poněkud ošemetné. Zatímco kauzální vztahy (známe-li je) můžeme použít pro konstrukci bayesovské sítě, ne každá hrana v bayesovské síti musí mít kauzální interpretaci. Někdy se může jednat o pouhé korelace dvou jevů.



Obr. 4 Kauzální inference (systém BKD)

5.6.3.3 Učení

Bayesovské sítě v sobě kombinují dva typy znalostí: znalosti o struktuře vazeb mezi atributy (hrany v grafu) a znalosti o pravděpodobnostech hodnot těchto atributů (ohodnocení uzlů v grafu). Při dobývání znalostí (učení) jsou tedy v zásadě dvě možnosti:

- vyjít ze známé struktury (získané např. od experta) a z dat odvozovat pouze podmíněné pravděpodobnosti,
- z dat odvodit strukturu sítě i pravděpodobnosti.

Obě uvedené možnosti mohou být ještě komplikovány skutečností, že v datech mohou být chybějící hodnoty, resp. že některé veličiny nelze pozorovat. Při učení bayesovské sítě tedy mohou nastat následující, postupně stále složitější situace:

1. *Známá struktura, veličiny plně pozorovatelné (metoda maximálně věrohodného odhadu)*.

   V nejjednodušším případě jde o to spočítat z dat odhady podmíněných pravděpodobnostních distribucí pro jednotlivé uzly sítě. Cílem je nalézt maximálně věrohodný odhad vzhledem k trénovacím datům. Podobně jako v případě naivního bayesovského klasifikátoru můžeme tyto odhady spočítat na základě četností[14]:

$$P(M=m \mid K=k, D=d) = \frac{n(M(m) \wedge K(k) \wedge D(d))}{n(K(k) \wedge D(d))}$$

2. *Známá struktura, veličiny částečně pozorovatelné (gradientní metoda nebo EM algoritmus)*

   Situace, kdy některé veličiny (uzly sítě) nelze pozorovat (jsou takzvaně skryté nebo latentní), je analogická situaci, kdy se u neuronové sítě učí váhy skryté vrstvy. Lze tedy použít *gradientní metodu* pro určení hodnot v tabulce podmíněných pravděpodobností (hypotéza h). Místo minimalizování ztrátové funkce jako v případě neuronových sítí zde budeme maximalizovat pravděpodobnost $P(D \mid h)$, že pozorujeme daná trénovací data D pokud platí podmíněné pravděpodobnosti. Tento postup, navržený Russellem [Russell a kol, 1995], je založen na výpočtu gradientu funkce $\ln P(D \mid h)$. Nechť p_{ijk} je podmíněná pravděpodobnost toho, že veličina Y_i nabývá hodnotu $y_{i,j}$, pokud rodiče U_i této veličiny nabývají hodnotu $u_{i,k}$ Tedy např. pro první hodnotu $P(M \mid K,D)$ uvedenou v Tab. 2 je $y_{i,j}=0$ a $u_{i,k}=\{0,0\}$. Potom

$$\frac{\partial \ln P(D \mid h)}{\partial p_{ijk}} = \sum_{d \in D} \frac{P(Y_i=y_{ij}, U_i=u_{ik} \mid d)}{p_{ijk}}.$$

   Hodnotu p_{ijk} pak (po jedné iteraci přes celá trénovací data) upravíme podle vztahu

$$p_{ijk} = p_{ijk} + \eta \sum_{d \in D} \frac{P(Y_i=y_{ij}, U_i=u_{ik} \mid d)}{p_{ijk}}.$$

   Výsledné hodnoty p_{ijk} ještě musíme normalizovat, aby se jednalo o pravděpodobnosti.

   Pro výpočet gradientu potřebujeme znát hodnoty $P(Y_i=y_{ij}, U_i=u_{ik})$ pro každý příklad v trénovacích datech. Jsou-li tyto veličiny nepozorovatelné, můžeme potřebné hodnoty spočítat z hodnot pozorovatelných pomocí běžné inference v bayesovské síti.

[14] Lze samozřejmě použít i Laplaceovu korekci nebo m-odhad.

Jinou možností je použít *EM* algoritmus [Dempster a kol, 1997]. Základní myšlenka algoritmu vychází z následující úvahy: „Kdybychom znali hodnoty skrytých veličin, mohli bychom přímo spočítat parametry podmíněné pravděpodobnostní distribuce. Zkusme tedy vyjít z hodnot, které očekáváme na základě „nástřelu“ parametrů distribuce. Tyto hodnoty pak použijeme pro výpočet parametrů. *EM* algoritmus tedy pracuje iterativně ve dvou opakujících se krocích (Obr. 5). V kroku *expektace* (krok 1) se počítají očekávané hodnoty skryté veličiny, v kroku *maximalizace* (krok 2) se počítají parametry distribuce maximalizující dané kritérium.

EM Algoritmus

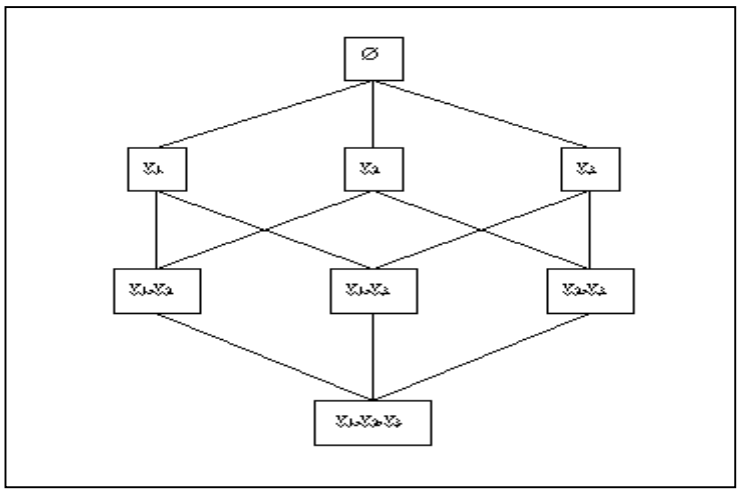
Inicializace
Náhodně zvol parametry distribuce

Hlavní cyklus
1. Na základě parametrů distribuce spočítej hodnoty skrytých veličin
2. Z takto spočítaných hodnot urči maximálně věrohodný odhad parametrů distribuce
3. Pokud došlo ke změně parametrů, vrať se k bodu 1.

Obr. 5 Expectation Maximization (EM) algoritmus

3. *Neznámá struktura, veličiny plně pozorovatelné* (*prohledávání prostoru modelů)*

V tomto případě musíme z dat odhadovat i strukturu sítě. Jedná se tedy vlastně o úlohu prohledávání prostoru hypotéz, různých topologií sítě. V případě, že známe uspořádání uzlů (od rodičů k dětem), můžeme (na základě dat) ke každému uzlu v síti určovat jeho rodiče nezávisle. Množina těchto rodičů může být 0-prvková, 1-prvková, 2-prvková , …. Můžeme tedy všechny množiny potenciálních rodičů uvažovaného uzlu uspořádat podle obecnosti a pak tyto množiny procházet shora-dolů, nebo zdola-nahoru. Vhodnější je postup shora-dolů, který preferuje jednodušší sítě. Úloha prohledávání je ale výpočetně velice náročná. Obr. 6 ukazuje možné rodiče uzlu u vytvářené z tříprvkové množiny v_i, $i=1..k$; různých množin rodičů tedy může být 2^k.

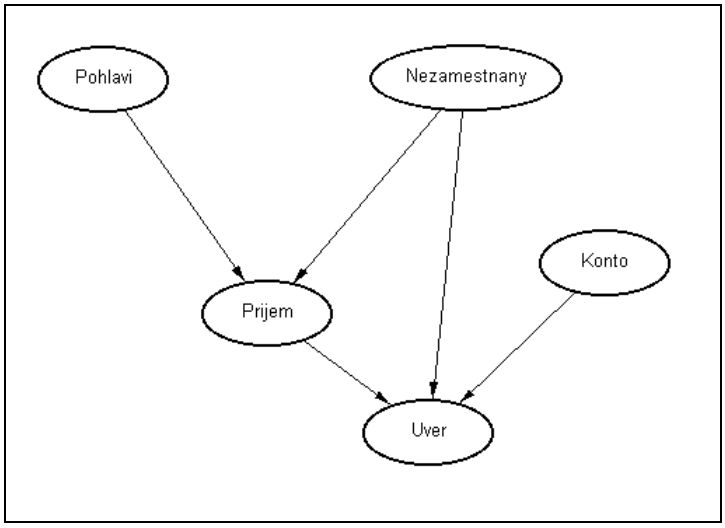


Obr. 6 Potenciální rodiče

4. *Neznámá struktura, veličiny částečně pozorovatelné (EM + prohledávání prostoru modelů)*

V nejsložitějším případě musíme odhadovat strukturu sítě z neúplných dat. Jednou možností je použít tzv. strukturální *EM* algoritmus popsaný v [Friedman,1998]. Pro každý model spočítáme optimální parametry pomocí algoritmu *EM* a pak vybereme nejlepší model.

Na příkladu si ukážeme tu nejjednodušší variantu; výpočet podmíněných pravděpodobností v případě známé struktury sítě. Opět použijeme data uvedená v Tab. 1. Řekněme, že expert zadal strukturu sítě v podobě uvedené na Obr. 7.



Obr. 7 Bayesovská síť pro příklad o úvěrech

Sdružená pravděpodobnostní distribuce pro tuto síť je tedy

P(příjem,konto,pohlaví,nezaměstnaný,uvěr) =

P(pohlaví) P(nezam) P(příjem|pohlaví,nezam) P(konto) P(úvěr|příjem,nezam,konto).

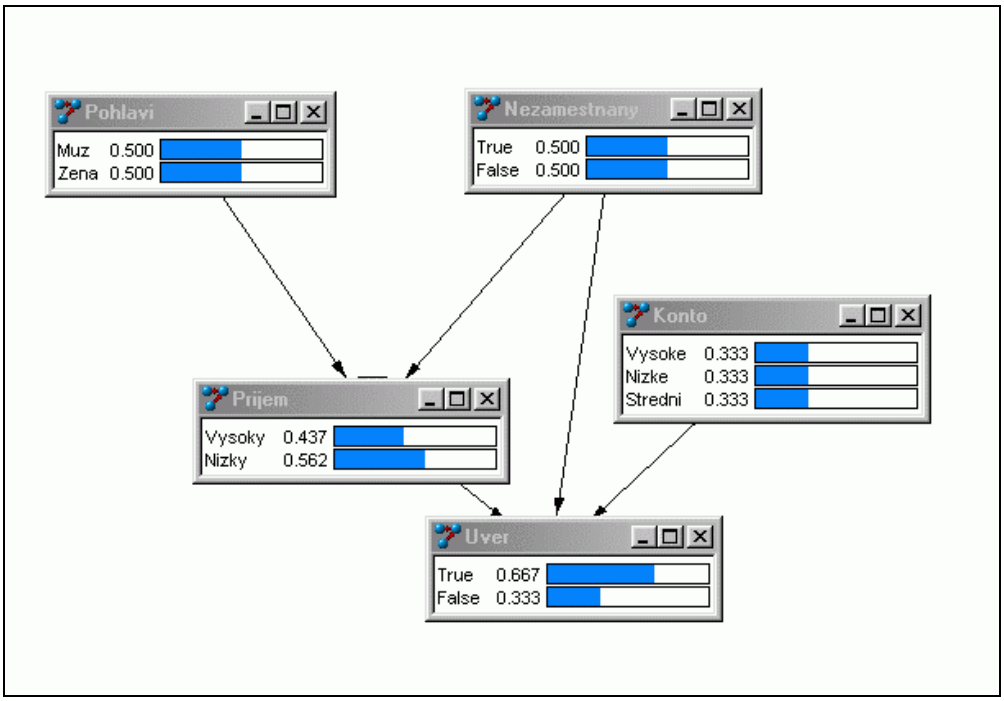
Z trénovacích dat musíme odhadnout všechny dílčí pravděpodobnosti uvedené faktorizace. Tedy např.:

$$P(\text{pohlaví}=\text{žena}) = \frac{n(\text{pohlaví}(\text{žena}))}{n} = 6/12 = 1/2$$

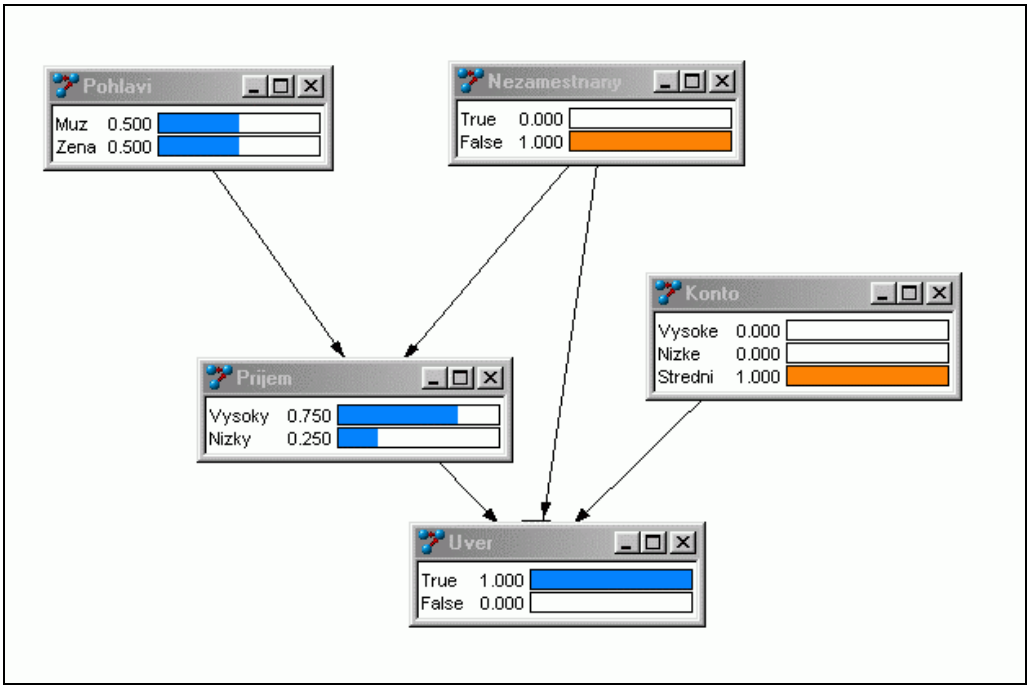
nebo

$$P(\text{příjem}=\text{vysoký}\,|\,\text{pohlaví}=\text{žena},\text{nezam}=\text{ano}) = \frac{n(\text{příjem}(\text{vysoký}) \wedge \text{pohlaví}(\text{žena}) \wedge \text{nezam}(\text{ano}))}{n(\text{pohlaví}(\text{žena}) \wedge \text{nezam}(\text{ano}))} = 1/4.$$

Bayesovskou síť pak můžeme použít pro klasifikaci nových klientů. Obr. 8 ukazuje pravděpodobnosti hodnot jednotlivých atributů tak jak byly spočítány z trénovacích dat. Obr. 9 pak ukazuje jak se tyto pravděpodobnosti změní, jestliže víme, že nový klient není nezaměstnaný a má středně vysoké konto.



Obr. 8 Pravděpodobnosti odvozené z trénovacích dat (systém BKD)



Obr. 9 Pravděpodobnosti odvozené pro nového klienta (systém BKD)

5.6.4 Systémy a aplikace

"Klasický" naivní bayesovský klasifikátor je implementován např. v systému *Bayda* z univerzity v Helsinkách (http://www.cs.Helsinki.FI/research/cosco/), nebo v systému *RoC* z Open University ve Velké Britanii (http://kmi.open.ac.uk/projects/bkd/). Oba tyto systémy jsou volně dostupné přes Internet.

Zajímavým systémem vycházejícím z naivního bayesovského klasifikátoru je systém *AutoClass* [Cheeseman, Stultz, 1996]. *AutoClass* (http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/) umožňuje odvodit maximální aposteriorní pravděpodobnosti zařazení do tříd jak pro příklady klasifikované (učení s učitelem) tak i neklasifikované (učení bez učitele). Při učení bez učitele se používá *EM* algoritmus, skrytou veličinou je v tomto případě informace o zařazení do třídy. Algoritmus *EM* umožní nalézt pro uvažovaný počet tříd nejlepší rozdělení příkladů do těchto tříd. Třídy (shluky) s malou pravděpodobností se mohou zanedbat; průchod přes *EM* se pak zopakuje pro menší počet tříd. Tento postup umožňuje současně stanovit optimální počet tříd i zařazení příkladů k těmto třídám.

Jiným příkladem systému používajícím bayesovský přístup ke shlukování příkladů je *COBWEB*. Tento systém inkrementálním způsobem vytváří hierarchie konceptů metodou shora-dolů. Nový vstupní příklad se zařadí buď k existujícímu konceptu (uzlu hierarchie), nebo se pro něj vytvoří nový uzel (koncept). V případě, že příklad bude zařazen do existujícího uzlu se zvažuje ještě možnost tento uzel rozdělit respektive sloučit s jiným uzlem. Rozhodnutí, která operace se pro daný příklad provede (vytvoření nového uzlu, prosté přidání do existujícího uzlu, přidání a rozdělení, přidání a spojení) závisí na hodnotě

$$\frac{1}{T}\sum_t P(C(v_t))\left(\sum_j\sum_k P(A_j(v_k)\,|\,C(v_t))^2 - \sum_j\sum_k P(A_j(v_k))^2\right),$$

která porovnává podmíněnou pravděpodobnost, že atribut A_j má hodnotu v_k, pokud příklad patří ke konceptu $C(v_t)$ s apriorní pravděpodobností, že atribut A_j má hodnotu v_k. Pro zařazení příkladu se zvolí ta operace, která maximalizuje uvedenou hodnotu. Původní podoba algoritmu [Fisher, 1987] pracovala pouze s kategoriálními daty. Další verze umožňovaly pracovat i s numerickými daty ([Gennari a kol., 1989]). Systém lze získat na http://or.eng.tau.ac.il:7777/topics/ecobweb.html.

Na poli bayesovských sítí existuje celá řada různých implementací ať už komerčních nebo volně dostupných. K prvním systémům implementujícím bayesovské sítě patří *Hugin*, původně vyvinutý na univerzitě v Aalborgu v Dánsku skupinou kolem F. Jensena. Z univerzitního týmu vznikla soukromá firma Hugin (http://www.hugin.com), která tento systém nabízí v současnosti. Jiným známým příkladem komerčního systému je systém *Netica* nabízený kanadskou firmou Norsys (http://www.norsys.com). Z volně šířených implementací zmiňme systémy *GeNIe* z univerzity v Pittsburgu (http://www2.sis.pitt.edu/~genie/), *BN Power Con*structor z univerzity v Albertě (http://www.cs.ualberta.ca/~jcheng/bnsoft.htm) a *Bayesian Knowledge Discoverer* z Open University (http://kmi.open.ac.uk/projects/bkd/). Posledně jmenovaný systém byl použit pro vytvoření obrázků Obr. 3 , Obr. 4, Obr. 8 a Obr. 9.

Vývojem aplikací se mimo jiné zabývá firma Microsoft. V jejím výzkumném centru pracuje na bayesovských sítích skupina vedená D. Heckermanem. Jejich sítě můžeme nalézt v různých diagnostických modulech v operačním systému Windows. První z nich byl modul detekce chyb připojené tiskárny ve Windows95.

Literatura:

[Cheeseman, Stultz, 1996] Cheeseman, P. – Stultz,J.: Bayesian classification (AutoClass): Theory and results. In: (Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, eds.) Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, ISBN 0-262-56097-6.

[Dempster a kol,1997] Dempster,A.P. – Laird,N.M. – Rubin,D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B, 39(1), 1997, 1-38.

[Domingos, Pazzani, 1996] Domingos,P. – Pazzani,M.: Beyond independence: conditions for the optimality of the simple bayesian classifier. In: (Saitta ed.) Proc. 13th Int. Conference on Machine Learning ICML'96, 1996, 105-110.

[Duda, Hart, 1973] Duda,R.O. – Hart,P.E.: Pattern classification and scene analysis. Wiley, 1973.

[Fisher, 1987] Fisher,D.H.: Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning*, 2, 1987, 139-172.

[Friedman,1998] Friedman,N.: The bayesian structural EM algorithm. In Proc. UAI'98, 1998.

[Gennari a kol., 1989] Gennari, J.H. – Langley,P. – Fisher,D.H.: Models of incremental concept formation. *Artificial Intelligence*, 40, 1989, 11-61.

[Heckerman, 1995] Heckermann,D.: A tutorial on learning with Bayesian networks. Microsoft Research tech. Report MSR-TR-95-06.

[Jensen, 1996] Jensen, F.: An introduction to bayesian networks. UCL Press, 1996.

[Jiroušek, 1995] Jiroušek,R.: Metody reprezentace a zpracování znalostí v umělé inteligemci. Skripta VŠE, 1995.

[Kohavi a kol.,1997] Kohavi,R. – Becker,B. – Sommerfeld,D.: Improving simple Bayes. In: (Sommeren, Widmer, eds.) Poster Papers of the 9th European Conf. on Machine Learning ECML'97. 1997.

[Lauritzen, 1996] Lauritzen,S.: Graphical models. Oxford, 1996.

[Lauritzen Spiegelhalter, 1988] Lauritzen,S. – Spiegelhalter,D.: Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society,* Series B, 50, 1988, 157-224.

[Mitchell, 1997] Mitchell,T.: Machine learning. McGraw-Hill. 1997. ISBN 0-07-042807-7

[Murphy, 2000] Murphy,K.: A brief introduction to graphical models and bayesian networks. http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html

[Rissanen, 1978] Rissanen,J.: Modeling by shortest data description. *Automatica*, Vol. 14, 1978, 465-471.

[Russell a kol, 1995] Russell,S. – Binder,J. – Koller,D. – Kanazawa,K.: Local learning in probalistic networks with hidden variables. In: Proc. 14th Int. Joint Conference on Artificial Intelligence IJCAI'95, Morgan Kaufmann, 1995.