Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu
Strojno učenje, ak. god. 2011/12. – **bilješka za predavanja** (v2.1)

Bilješka 3
Bayesov klasifikator

Bayesov klasifikator je vjerojatno najpoznatiji klasifikacijski model i tipičan predstavnik parametarskih generativnih modela. Klasifikacija primjera ostvaruje se pomoću Bayesovog pravila, koji nam za svaku klasu daje vjerojatnost da primjer pripada toj klasi.

1 Bayesovo pravilo

Kod Bayesovog klasifikatora, klasifikacija primjera $\mathbf{x}$ temelji se na izračunu **aposteriorne vjerojatnosti** $P(Y = \mathcal{C}_j|X = \mathbf{x})$, tj. vjerojatnosti da primjer $\mathbf{x}$ pripada klasi $\mathcal{C}_j$. Tu vjerojatnost izračunavamo posredno, na temelju zajedničke gustoće $p(\mathbf{x}, \mathcal{C}_j)$, primjenom Bayesovog pravila:

$$P(\mathcal{C}_j|\mathbf{x}) = \frac{p(\mathbf{x}, \mathcal{C}_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)}{\sum_{k=1}^{K} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)} \,. \tag{1}$$

Marginalnu vjerojatnost $P(\mathcal{C}_j)$ nazivamo **apriorna vjerojatnost klase** (engl. *class prior*), a uvjetnu gustoću $p(\mathbf{x}|\mathcal{C}_j)$ nazivamo **klasom uvjetovana gustoća** (engl. *class conditional density*) ili, općenitije, **izglednost klase** (engl. *class likelihood*). Ako je $\mathbf{x}$ diskretna varijabla, umjesto gustoća $p(\mathbf{x}|\mathcal{C}_j)$ i $p(\mathbf{x})$ koristimo odgovarajuće vjerojatnosti.

Svi izrazi koji se pojavljuju u (1) mogu se dobiti marginalizacijom odnosno normalizacijom zajedničke gustoće $p(\mathbf{x}, \mathcal{C}_j)$, koja uvijek sadržava potpunu informaciju o podatcima. Međutim, faktorizacija zajedničke gustoće u izglednost i apriornu vjerojatnost bitno pojednostavljuje modeliranje jer omogućava da zasebno, pomoću odabrane parametarske razdiobe, modeliramo izglednost svake klase (tj. razdiobu primjera unutar svake klase). Na primjer, za kontinuiranu varijablu $\mathbf{x}$ izglednost se uobičajeno modelira Gaussovom razdiobom.

Optimalna klasifikacijska odluka jest ona koja maksimizira aposteriornu vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$. Drugim riječima, primjer $\mathbf{x}$ treba klasificirati u onu klasu $\mathcal{C}_j$ za koju je $P(\mathcal{C}_j|x)$ najveći. Takvu hipotezu nazivamo **maksimalna aposteriorna hipoteza** (MAP):

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathcal{C}_k} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) \,. \tag{2}$$

Primijetite da je dovoljno maksimizirati samo brojnik izraza (1), budući da je marginalna gustoća $p(\mathbf{x})$ za sve klase $\mathcal{C}_j$ jednaka i služi samo normalizaciji na jedinični interval. Ako želimo vjerojatnosnu interpretaciju hipoteze, možemo definirati zasebnu hipotezu h_j za svaku klasu $\mathcal{C}_j$:

$$h_j(\mathbf{x}) = P(\mathcal{C}_j|\mathbf{x}) \,.$$

Zanemarimo li nazivnik $p(\mathbf{x})$, hipotezu možemo još jednostavnije definirati kao

$$h_j(\mathbf{x}) = p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)$$

Ovime gubimo vjerojatnosnu interpretaciju jer vrijednost nije normalizirana na jedinični interval, no vrijednost $h_j(\mathbf{x})$ može se tumačiti kao pouzdanosti klasifikacije primjera $\mathbf{x}$ u klasu $\mathcal{C}_j$. U oba ova slučaja govorimo o **klasifikatoru s ocjenom pouzdanosti** (engl. *confidence-rated classifier*).

Primjer 1 (Bayesova klasifikacija) Razmotrimo problem klasifikacije u tri klase: $\mathcal{C}_1, \mathcal{C}_2$ i $\mathcal{C}_3$. Neka su apriorne vjerojatnosti tih klasa $P(\mathcal{C}_1) = P(\mathcal{C}_2) = 0.3$ i $P(\mathcal{C}_3) = 0.4$. Poznato je da za primjer $\mathbf{x}$ izglednosti iznose $p(\mathbf{x}|\mathcal{C}_1) = 0.9$ i $p(\mathbf{x}|\mathcal{C}_2) = p(\mathbf{x}|\mathcal{C}_3) = 0.4$. Izračunajmo maksimalnu aposteriornu hipotezu za svaku od klasa. U brojniku izraza (1) imamo

$$\begin{aligned} p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) &= 0.9 \times 0.3 = 0.27 \\ p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) &= 0.4 \times 0.3 = 0.12 \\ p(\mathbf{x}|\mathcal{C}_3)P(\mathcal{C}_3) &= 0.4 \times 0.4 = 0.16 \end{aligned}$$

Apriorna gustoća vjerojatnosti primjera $\mathbf{x}$ jest $p(\mathbf{x}) = \sum_{k=1}^{3} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) = 0.55$. Za aposteriorne vjerojatnosti $P(\mathcal{C}_2|\mathbf{x})$ dobivamo

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{0.27}{0.55} = 0.49 \qquad P(\mathcal{C}_2|\mathbf{x}) = \frac{0.12}{0.55} = 0.22 \qquad P(\mathcal{C}_3|\mathbf{x}) = \frac{0.16}{0.55} = 0.29$$

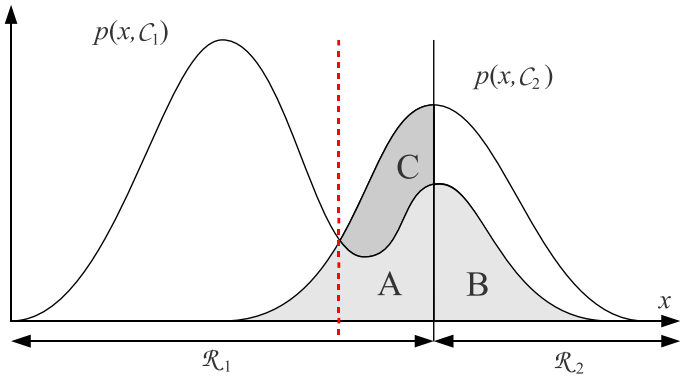
U ovom slučaju aposteriorna vjerojatnost najveća je za klasu $\mathcal{C}_1$, pa je maksimalna aposteriorna hipoteza $h(\mathbf{x}) = \mathcal{C}_1$.

1.1 Minimizacija pogreške klasifikacije

Premda je intuitivno jasno da je optimalna klasifikacijska odluka ona koja maksimizira aposteriornu vjerojatnost $P(\mathcal{C}_j|\mathbf{x})$, uvjerimo se da je to doista tako. Pretpostavimo da je naš cilj minimizirati broj pogrešaka klasifikacije. Ograničimo se na slučaj dvije klase. Pogreška klasifikacije nastupa ako se primjer $\mathbf{x} \in \mathcal{C}_2$ klasificira u klasu $\mathcal{C}_1$, ili obrnuto. Označimo sa $\mathcal{R}_1 \subseteq \mathcal{X}$ primjere koje naš klasifikator klasificira kao $\mathcal{C}_1$, tj. $\mathcal{R}_1 = \{\mathbf{x} \in \mathcal{X} \mid h_1(\mathbf{x}) = 1\}$, a sa $\mathcal{R}_2 = \mathcal{X} \backslash \mathcal{R}_1$ primjere koje naš klasifikator klasificira kao $\mathcal{C}_2$. Vjerojatnost pogreške je

$$P(\text{pogreška}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) = \int_{\mathbf{x}\in\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathbf{x}\in\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}\,.$$

Postavlja se pitanje: kako odabrati područja $\mathcal{R}_1$ i $\mathcal{R}_2$, odnosno gdje postaviti granicu između klasa, a da vjerojatnost pogrešne klasifikacije bude najmanja? Očito je da je granicu potrebno postaviti tako da se svaki primjer $\mathbf{x}$ klasificira u onu klasu za koju je vrijednost ispod integrala manja, jer time za taj primjer smanjujemo vjerojatnost pogreške. To pak znači da, ako je $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$, tada primjer $\mathbf{x}$ trebamo klasificirati u klasu $\mathcal{C}_1$ (tj. $\mathcal{R}_1$ treba namjestiti tako da je $\mathbf{x} \in \mathcal{R}_1$), jer je tada vjerojatnost pogreške manja i jednaka je $p(\mathbf{x}, \mathcal{C}_2)$. Budući da vrijedi $p(\mathbf{x}, \mathcal{C}_k) = P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$, a da je faktor $p(\mathbf{x})$ zajednički obama pribrojnicima, to slijedi da je vjerojatnost pogreške minimalna ako se svaki primjer $\mathbf{x}$ klasificira u klasu za koju je aposteriorna vjerojatnost $P(\mathcal{C}_k|\mathbf{x})$ najveća. Opisana situacija prikazana je slikom 1 za jednodimenzijski prostor primjera $\mathcal{X} = \mathbb{R}$.



Slika 1: Ilustracija zajedničkih gustoća vjerojatnosti $p(x, \mathcal{C}_1)$ i $p(x, \mathcal{C}_2)$ za prostor primjera $\mathcal{X} = \mathbb{R}$. Granice između područja odluke $\mathcal{R}_1$ i $\mathcal{R}_2$ predstavljena je okomitom linijom. Vjerojatnost pogreške klasifikacije jednaka je površini A+B+C. Površina A+B=konst., dok je površinu C moguće ukloniti ako se kao granica između klasa izabere iscrtkana linija, tj. vrijednost x za koju $p(x, \mathcal{C}_1) = p(x, \mathcal{C}_2)$. To je istovjetno odabiru klase za koju je aposteriorna vjerojatnost $P(\mathcal{C}_k|\mathbf{x})$ maksimalna.

1.2 *Minimizacija rizika

U gornjim razmatranjima pretpostavili smo da je naš cilj minimizirati broj pogrešnih klasifikacija. Situacija je drugačija kada gubitci nisu jednoliki, odnosno kada matrica gubitka (L_{kj}) nije nula-jedan. Tada je potrebno minimizirati očekivanu vrijednost funkcije gubitka L. Očekivana vrijednost funkcije gubitka L naziva se **funkcija rizika** i definira kao

$$\mathbb{E}[L] = \sum_{k=1}^{K} \sum_{j=1}^{K} \int_{\mathbf{x} \in \mathcal{R}_j} L_{kj}\, p(\mathbf{x}, \mathcal{C}_k)\, \mathrm{d}\mathbf{x} \tag{3}$$

gdje je L_{kj} gubitak uslijed pogrešne klasifikacije primjera iz klase $\mathcal{C}_k$ u klasu $\mathcal{C}_j$. Integracija se provodi po regiji $\mathcal{R}_j \subseteq \mathcal{X}$, koja obuhvaća sve primjere koje klasifikator klasificira kao $\mathcal{C}_j$, tj. $\mathcal{R}_j = \{\mathbf{x} \in \mathcal{X} \mid h_j(\mathbf{x}) = 1\}$.

Primjer 2 (Očekivanje gubitka) Za neki binaran klasifikacijski problem dana je asimetrična matrica gubitka

$$L = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}.$$

Dan je (diskretan) prostor primjera $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$. Pomoću generativnog probabilističkog modela izračunate su izglednosti klasa

$$P(\mathbf{x}^{(1)}|\mathcal{C}_1) = 0.75 \qquad P(\mathbf{x}^{(2)}|\mathcal{C}_1) = 0.1 \qquad P(\mathbf{x}^{(3)}|\mathcal{C}_1) = 0.15$$
$$P(\mathbf{x}^{(1)}|\mathcal{C}_2) = 0.25 \qquad P(\mathbf{x}^{(2)}|\mathcal{C}_2) = 0.55 \qquad P(\mathbf{x}^{(3)}|\mathcal{C}_2) = 0.2$$

Apriorne vjerojatnosti klasa neka su $P(\mathcal{C}_1) = 0.8$ i $P(\mathcal{C}_2) = 0.2$. Zajedničke vjerojatnosti $P(\mathbf{x}, \mathcal{C}_j) = P(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j)$ su

$$P(\mathbf{x}^{(1)}, \mathcal{C}_1) = 0.6 \qquad P(\mathbf{x}^{(2)}, \mathcal{C}_1) = 0.08 \qquad P(\mathbf{x}^{(3)}, \mathcal{C}_1) = 0.12$$
$$P(\mathbf{x}^{(1)}, \mathcal{C}_2) = 0.05 \qquad P(\mathbf{x}^{(2)}, \mathcal{C}_2) = 0.11 \qquad P(\mathbf{x}^{(3)}, \mathcal{C}_2) = 0.04$$

iz čega za aposteriorne vjerojatnosti dobivamo

$$P(\mathcal{C}_1|\mathbf{x}^{(1)}) = 0.92 \qquad P(\mathcal{C}_1|\mathbf{x}^{(2)}) = 0.42 \qquad P(\mathcal{C}_1|\mathbf{x}^{(3)}) = 0.75$$
$$P(\mathcal{C}_2|\mathbf{x}^{(1)}) = 0.08 \qquad P(\mathcal{C}_2|\mathbf{x}^{(2)}) = 0.58 \qquad P(\mathcal{C}_2|\mathbf{x}^{(3)}) = 0.25$$

Pretpostavimo da se sva tri primjera klasificiraju tako da se smanji broj pogrešaka klasifikacije, tj. svaki primjer klasificira se u aposteriorno najvjerojatniju klasu: primjeri $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(3)}$ u klasu $\mathcal{C}_1$, a primjer $\mathbf{x}^{(2)}$ u klasu $\mathcal{C}_2$. Tada $\mathcal{R}_1 = \{\mathbf{x}^{(1)}, \mathbf{x}^{(3)}\}$ i $\mathcal{R}_2 = \{\mathbf{x}^{(2)}\}$. Očekivani gubitak je

$$\begin{aligned}
\mathbb{E}[L] &= \sum_{k=1}^{2}\sum_{j=1}^{2}\sum_{\mathbf{x}\in\mathcal{R}_j} L_{kj}P(\mathbf{x},\mathcal{C}_k) = \sum_{k=1}^{2}\Big(\sum_{\mathbf{x}\in\mathcal{R}_1} L_{k,1}P(\mathbf{x},\mathcal{C}_k) + \sum_{\mathbf{x}\in\mathcal{R}_2} L_{k,2}P(\mathbf{x},\mathcal{C}_k)\Big) \\
&= L_{21}\sum_{\mathbf{x}\in\mathcal{R}_1} P(\mathbf{x},\mathcal{C}_2) + L_{12}\sum_{\mathbf{x}\in\mathcal{R}_2} P(\mathbf{x},\mathcal{C}_1) \\
&= L_{21}P(\mathbf{x}^{(1)},\mathcal{C}_2) + L_{21}P(\mathbf{x}^{(3)},\mathcal{C}_2) + L_{12}P(\mathbf{x}^{(2)},\mathcal{C}_1) \\
&= 1\times 0.05 + 1\times 0.04 + 10\times 0.08 = 0.89
\end{aligned}$$

U ovom slučaju očekivani gubitak nije minimalan jer klasifikacija primjera $\mathbf{x}^{(2)}$ u smislu minimizacije rizika nije optimalna. Ako bismo primjer $\mathbf{x}^{(2)}$ klasificirali u klasu $\mathcal{C}_1$, očekivani gubitak bio bi optimalan i iznosio bi

$$\mathbb{E}[L] = 1\times 0.05 + 1\times 0.04 + 1\times 0.11 = 0.2\,.$$

Razlog tome jest što je, sukladno zadanoj matrici gubitka L, gubitak uslijed pogrešne klasifikacije u klasu $\mathcal{C}_2$ deset puta veći od gubitka uslijed pogrešne klasifikacije u klasu $\mathcal{C}_1$.

Želimo li dakle minimizirati očekivani gubitak $\mathbb{E}[L]$, svaki primjer $\mathbf{x}$ treba klasificirati u klasu $\mathcal{C}_j$ za koju je vrijednost $\sum_k L_{kj}\,p(\mathbf{x},\mathcal{C}_k)$ minimalna. Budući da vrijedi $p(\mathbf{x},\mathcal{C}_k) = P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$, a da je faktor $p(\mathbf{x})$ zajednički svim klasama, optimalna klasifikacijska odluka jest ona koja minimizira izraz

$$R(\mathcal{C}_j|\mathbf{x}) = \sum_{k=1}^{K} L_{kj}P(\mathcal{C}_k|\mathbf{x}) \tag{4}$$

i to je **očekivani rizik** pri klasifikaciji primjera $\mathbf{x}$ u klasu $\mathcal{C}_j$. Ako dakle želimo optimalnu klasifikaciju u smislu minimizacije rizika, funkciju hipoteze možemo definirati kao

$$h(\mathbf{x}) = \operatorname*{argmin}_{\mathcal{C}_k} R(\mathcal{C}_k|\mathbf{x})\,. \tag{5}$$

ili, ako želimo klasifikaciju s ocjenom pouzdanosti, kao

$$h_j(\mathbf{x}) = -R(\mathcal{C}_j|\mathbf{x})\,. \tag{6}$$

Primjer 3 (Minimizacija rizika) Za neki višeklasni klasifikacijski problem dana je matrica gubitka

$$L = \begin{pmatrix} 0 & 1 & 5 \\ 1 & 0 & 5 \\ 10 & 100 & 0 \end{pmatrix}.$$

Ako su za neki primjer $\mathbf{x}$ temeljem probabilističkog modela izračunate aposteriorne vjerojatnosti $P(\mathcal{C}_1|\mathbf{x}) = 0.25$, $P(\mathcal{C}_2|\mathbf{x}) = 0.6$, $P(\mathcal{C}_3|\mathbf{x}) = 0.15$, odredimo u koju klasu $\mathcal{C}_j$ treba klasificirati primjer $\mathbf{x}$, a da rizik pogrešne klasifikacija bude minimalan. Trebamo minimizirati vrijednost (4) za $j \in \{1, 2, 3\}$:

$$\begin{aligned}
(j = 1): &\quad R(\mathcal{C}_1|\mathbf{x}) = \textstyle\sum_k L_{k,1} P(\mathcal{C}_k|\mathbf{x}) = 1 \times 0.6 + 10 \times 0.15 = 2.10 \\
(j = 2): &\quad R(\mathcal{C}_2|\mathbf{x}) = \textstyle\sum_k L_{k,2} P(\mathcal{C}_k|\mathbf{x}) = 1 \times 0.25 + 100 \times 0.15 = 15.25 \\
(j = 3): &\quad R(\mathcal{C}_3|\mathbf{x}) = \textstyle\sum_k L_{k,3} P(\mathcal{C}_k|\mathbf{x}) = 5 \times 0.25 + 5 \times 0.6 = 4.25
\end{aligned}$$

Rizik je najmanji ako se primjer $\mathbf{x}$ klasificira u klasu $\mathcal{C}_1$. Ako međutim u obzir ne bismo uzimali rizik, ili ako bi gubitak bio jednolik (matrica nula-jedan), primjer $\mathbf{x}$ klasificirali bismo u klasu $\mathcal{C}_2$, budući da je aposteriorna vjerojatnost $P(\mathcal{C}_2|\mathbf{x})$ najveća.

1.3 *Kompenzacija neujednačene zastupljenosti klasa

U slučajevima kada broj pozitivnih primjera izrazito nadmašuje broj negativnih primjera (ili obrnuto), problematično je naučiti model koji dobro generalizira. Problem se može riješiti tako da se skup za učenje najprije umjetno uravnoteži, a zatim se, nakon učenja, aposteriorne vjerojatnosti $P(\mathcal{C}_j|\mathbf{x})$ skaliraju i normaliziraju uzevši u obzir stvarnu zastupljenost klasa.

Primjer 4 (Kompenzacija neujednačenosti klasa) Bayesov klasifikator koristimo za detekciju određene vrste raka na temelju medicinskih nalaza pacijenata. Neka je $\mathcal{C}$ klasa primjera u kojima je detektiran rak. Pretpostavimo da se dotična vrsta raka u slučajnom uzorku pacijenata pojavljuje u 1/1000 slučajeva, pa je najizglednija procjena $P(\mathcal{C}) = 0.001$. Ako bi se takav slučajan uzorak koristio kao skup za učenje, on bi bio izrazito neuravnotežen. U namjeri da to kompenziramo, pripremili smo umjetno uravnotežen skup za učenje u kojemu je od 1000 primjera njih 400 pozitivnih. Na tako uravnoteženom skupu vrijedi $P'(\mathcal{C}) = 0.4$.

Pretpostavimo da, nakon učenja na uravnoteženom skupu, model za neki primjer $\mathbf{x}$ izračunava aposteriornu vjerojatnost $P'(\mathcal{C}|\mathbf{x}) = 0.7$. Onda je stvarna aposteriorna vjerojatnost, ispravljena s obzirom na pravu zastupljenost klase $\mathcal{C}$, sljedeća:

$$P(\mathcal{C}|\mathbf{x}) = P'(\mathcal{C}|\mathbf{x}) \times \frac{P(\mathcal{C})}{P'(\mathcal{C})} = 0.7 \times \frac{0.001}{0.4} = 0.00175\,.$$

Uz pretpostavku simetričnog gubitka (matrice gubitka nula-jedan), primjer $\mathbf{x}$ ne bismo klasificirali u klasu $\mathcal{C}$.

2 Naivan Bayesov klasifikator

Razmotrimo sada Bayesov klasifikacijski model za slučaj diskretnih ulaznih varijabli. Pretpostavimo da raspolažemo skupom primjera za učenje $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ te da je problem višeklasan, $y^{(i)} \in \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$. Naš model je

$$h(x_1, \ldots, x_n) = \operatorname*{argmax}_j P(x_1, \ldots, x_n|\mathcal{C}_j) P(y = \mathcal{C}_j)\,. \tag{7}$$

Trebamo dakle procjeniti parametre diskretnih razdioba $P(x_1, \ldots, x_n|\mathcal{C}_j)$ i $P(\mathcal{C}_j)$. Varijabla y je diskretna multinomijalna varijabla, pa za najizgledniju procjenu apriornih vjerojatnosti $P(\mathcal{C}_j)$ imamo

$$\hat{P}(\mathcal{C}_j) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} = \frac{N_j}{N} \tag{8}$$

tj. relativan udio primjera koji su označeni klasom $\mathcal{C}_j$. Ukupan broj ovakvih parametara je $K-1$, gdje je K broj klasa (za K-tu klasu vjerojatnost je određena zbog ograničenja $\sum_j P(\mathcal{C}_j) = 1$).

Na isti način mogli bismo pokušati procjeniti vjerojatnost $P(x_1, \ldots, x_n|\mathcal{C}_j)$, tako da varijable $x_1, \ldots, x_n$ tretiramo zajednički, tj. da vektor $\mathbf{x} = (x_1, \ldots, x_n)$ tretiramo kao jednu multinomijalnu varijablu. Međutim, to rješenje u praksi ne funkcionira. Problem je što broj mogućih stanja varijable $\mathbf{x}$ raste eksponencijalno s dimenzijom n, a to znači da i broj parametara modela raste eksponencijalno. Npr., binaran vektor $\mathbf{x}$ može poprimiti 2^n različitih vrijednosti. Za svaki takav vektor trebat će procijeniti vjerojatnost pripadanja jednoj od K klasa, što znači da ukupno treba procijeniti $(2^n - 1)K$ parametara. Zbog velikog broja parametara, model će imati vrlo visoku varijancu i bit će sklon prenaučenosti. Zapravo, takav model samo pohranjuje vjerojatnosti viđenih primjera, pa savršeno klasificira primjere iz skupa zaučenje, dok svim neviđenim primjerima dodijeljuje vjerojatnost nula. Kako bi model mogao generalizirati, nužno je uvesti neke pretpostavke.

Prije no što uvedemo ikakvu pretpostavku, primijetimo da vjerojatnost $P(x_1, \ldots, x_n|\mathcal{C}_j)$ možemo faktorizirati primjenom pravila lanca. Općenito, **pravilo lanca** (engl. *chain rule*) glasi

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\cdots P(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_{k-1}). \end{aligned} \tag{9}$$

Dodavanjem uvjetne varijable $\mathcal{C}_j$, dobivamo:

$$P(x_1, \ldots, x_n|\mathcal{C}_j) = \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_{k-1}, \mathcal{C}_j). \tag{10}$$

Pretpostavka koju sada možemo uvesti jest da su varijable međusobno **uvjetno nezavisne** za zadanu klasu, odnosno da vrijedi

$$P(x_i|x_j, \mathcal{C}) = P(x_i|\mathcal{C})\,. \tag{11}$$

Višestrukom primjenom ove jednakosti izraz (10) faktorizira se na

$$P(x_1, \ldots, x_n|\mathcal{C}_j) = \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_{k-1}, \mathcal{C}_j) \overset{(11)}{=} \prod_{k=1}^{n} P(x_k|\mathcal{C}_j) \tag{12}$$

što daje model

$$h(x_1, \ldots, x_n) = \operatorname*{argmax}_j P(\mathcal{C}_j) \prod_{k=1}^{n} P(x_k|\mathcal{C}_j)\,. \tag{13}$$

Ovaj model nazivamo **naivan Bayesov klasifikator** (engl. *naïve Bayes classifer*).[1] Nazivamo ga naivnim jer pretpostavka o uvjetnoj nezavisnosti u praksi uglavnom ne vrijedi.

[1] Od 1961., kada je ovaj model predložen, javlja se i pod drugim imenima: *idiot Bayes*, *simple Bayes*, *idependent Bayes*.

Primjerice, u kontekstu klasifikacije dokumenata, ova pretpostavka kaže da, ako nam je poznato da dokument pripada klasi "Sport", vjerojatnost pojavljivanja riječi "nogomet" jednaka je neovisno o tome pojavljuje li se u istome tekstu riječ "lopta". Pretpostavka je sasvim sigurno pogrešna, no ipak se pokazuje da naivan Bayesov klasifikator u praksi vrlo dobro funkcionira.

Vjerojatnosti $P(x_k|\mathcal{C}_j)$ možemo jednostavno procjeniti metodom najveće izglednosti:

$$\hat{P}(x_k|\mathcal{C}_j) = \frac{\sum_{i=1}^{N} \mathbf{1}\{x_k^{(i)} = x_k \wedge y^{(i)} = \mathcal{C}_j\}}{\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \mathcal{C}_j\}} = \frac{N_{kj}}{N_j} \tag{14}$$

ili Laplaceovim procjeniteljem:

$$\hat{P}(x_k|\mathcal{C}_j) = \frac{\sum_{i=1}^{N} \mathbf{1}\{x_k^{(i)} = x_k \wedge y^{(i)} = \mathcal{C}_j\} + \lambda}{\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} + K_k\lambda} = \frac{N_{kj} + \lambda}{N_j + K_k\lambda} \tag{15}$$

gdje je K_k broj mogućih vrijednosti značajke x_k.

Ukupan broj ovakvih vjerojatnosti koje moramo procijeniti je $\sum_{k=1}^{n}(K_k - 1)K$. Ako su značajke binarne, taj broj je nK. Broj parametara sada linearno ovisi o dimenziji n, a ne više eksponencijalno. Dakle, naivna pretpostavka u uvjetnoj nezavisnosti ulaznih varijabli omogućila nam je značajno smanjenje broja parametara.

Primjer 5 (Naivan Bayesov klasifikator) Naivan Bayesov klasifikator koristimo za klasifikaciju SF-filmova u kategoriju *Dobar film*. Koristimo sljedeće značajke:

$$\begin{aligned}
\text{Mjesto radnje} &= \{\text{svemir}, \text{Zemlja}, \text{drugdje}\} \\
\text{Glavni lik} &= \{\text{znanstvenica}, \text{dijete}, \text{kriminalac}, \text{policajac}\} \\
\text{Vrijeme radnje} &= \{\text{prošlost}, \text{budućnost}, \text{sadašnjost}\} \\
\text{Vanzemaljci} &= \{\text{da}, \text{ne}\}
\end{aligned}$$

	x_1	x_2	x_3	x_4	y
i	Mjesto radnje	Glavni lik	Vrijeme radnje	Vanzemaljci	Dobar film
1	svemir	znanstvenica	sadašnjost	da	ne
2	Zemlja	kriminalac	budućnost	ne	ne
3	drugdje	dijete	prošlost	da	ne
4	svemir	znanstvenica	sadašnjost	ne	da
5	svemir	kriminalac	prošlost	ne	ne
6	Zemlja	dijete	prošlost	da	da
7	Zemlja	policajac	budućnost	da	ne
8	svemir	policajac	budućnost	ne	da

Model je binaran ($K = 2$), a broj vrijednosti pojedinih značajki je $K_1 = 3$, $K_2 = 4$, $K_3 = 3$ i $K_4 = 2$. Ukupan broj parametara ovog modela je:

$$K - 1 + \sum_{k=1}^{n}(K_k - 1)K = 1 + 2 \times (2 + 3 + 2 + 1) = 17\,.$$

Preglednosti radi, u nastavku navodimo redundantan skup od ukupno 26 parametara. ML-procjene za apriorne vjerojatnosti klasa su:

$$P(y = \text{da}) = \tfrac{1}{8}\textstyle\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \text{da}\} = \tfrac{3}{8} \qquad P(y = \text{ne}) = \tfrac{1}{8}\textstyle\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \text{ne}\} = \tfrac{5}{8} \tag{16}$$

ML-procjene za vjerojatnosti $P(x_k|\mathcal{C}_j)$ su:

$$\begin{array}{llll}
P(x_1 = \text{svemir}|y = \text{da}) & = 2/3 & P(x_1 = \text{svemir}|y = \text{ne}) & = 2/5 \\
P(x_1 = \text{Zemlja}|y = \text{da}) & = 1/3 & P(x_1 = \text{Zemlja}|y = \text{ne}) & = 2/5 \\
P(x_1 = \text{drugdje}|y = \text{da}) & = 0 & P(x_1 = \text{drugdje}|y = \text{ne}) & = 1/5 \\
 & & & \\
P(x_2 = \text{znanstvenica}|y = \text{da}) & = 1/3 & P(x_2 = \text{znanstvenica}|y = \text{ne}) & = 1/5 \\
P(x_2 = \text{dijete}|y = \text{da}) & = 1/3 & P(x_2 = \text{dijete}|y = \text{ne}) & = 1/5 \\
P(x_2 = \text{kriminalac}|y = \text{da}) & = 0 & P(x_2 = \text{kriminalac}|y = \text{ne}) & = 2/5 \\
P(x_2 = \text{policajac}|y = \text{da}) & = 1/3 & P(x_2 = \text{policajac}|y = \text{ne}) & = 1/5 \\
 & & & \\
P(x_2 = \text{prošlost}|y = \text{da}) & = 1/3 & P(x_2 = \text{prošlost}|y = \text{ne}) & = 2/5 \\
P(x_2 = \text{sadašnjost}|y = \text{da}) & = 1/3 & P(x_2 = \text{sadašnjost}|y = \text{ne}) & = 1/5 \\
P(x_2 = \text{budućnost}|y = \text{da}) & = 1/3 & P(x_2 = \text{budućnost}|y = \text{ne}) & = 2/5 \\
 & & & \\
P(x_2 = \text{da}|y = \text{da}) & = 1/3 & P(x_2 = \text{da}|y = \text{ne}) & = 3/5 \\
P(x_2 = \text{ne}|y = \text{da}) & = 2/3 & P(x_2 = \text{ne}|y = \text{ne}) & = 2/5
\end{array}$$

Razmotrimo sada klasifikaciju novog primjera $\mathbf{x}^{(1)} = (\text{svemir}, \text{dijete}, \text{sadašnjost}, \text{da})$:

$$\begin{aligned}
P(\text{da}|\mathbf{x}^{(1)}) &\propto \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{da}) \\
&= P(\text{svemir}|\text{da})P(\text{dijete}|\text{da})P(\text{sadašnost}|\text{da})P(\text{da}|\text{da})P(\text{da}) \\
&= 2/3 \times 1/3 \times 1/3 \times 1/3 \times 3/8 = 0.009 \\
P(\text{ne}|\mathbf{x}^{(1)}) &\propto \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{ne}) \\
&= P(\text{svemir}|\text{ne})P(\text{dijete}|\text{ne})P(\text{sadašnost}|\text{ne})P(\text{da}|\text{ne})P(\text{ne}) \\
&= 2/5 \times 1/5 \times 1/5 \times 3/5 \times 5/8 = 0.006
\end{aligned}$$

MAP-hipoteza je:

$$h(\mathbf{x}^{(1)}) = \underset{\mathcal{C} \in \{\text{da},\text{ne}\}}{\operatorname{argmax}} P(\mathcal{C}|\mathbf{x}) = \text{da}$$

Što je s klasifikacijom primjera $\mathbf{x}^{(2)} = (\text{svemir}, \text{kriminalac}, \text{sadašnjost}, \text{ne})$?

$$\begin{aligned}
P(\text{da}|\mathbf{x}^{(2)}) &\propto \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{da}) \\
&= P(\text{svemir}|\text{da})P(\text{kiminalac}|\text{da})P(\text{sadašnost}|\text{da})P(\text{ne}|\text{da})P(\text{da}) \\
&= 2/3 \times 0 \times 1/3 \times 2/3 \times 3/8 = 0 \\
P(\text{ne}|\mathbf{x}^{(2)}) &\propto \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{ne}) \\
&= P(\text{svemir}|\text{ne})P(\text{kriminalac}|\text{ne})P(\text{sadašnost}|\text{ne})P(\text{ne}|\text{ne})P(\text{ne}) \\
&= 2/5 \times 2/5 \times 1/5 \times 2/5 \times 5/8 = 0.008
\end{aligned}$$

Zato što je $P(x = \text{kriminalac}|y = \text{da}) = 0$, aposteriorna vjerojatnost hipoteze $y = \text{da}$ je nula, neovisno o drugim vjerojatnostima u umnošku. Ovo je tipičan primjer prenaučenosti koji može nastupiti kod uporabe ML-procjena.

Primjer 6 (Zaglađeni naivan Bayesov klasifikator) Razmotrimo gornji skup za učenje i klasifikaciju primjer $\mathbf{x}^{(2)}$, ali uz zaglađivanje. Umjesto ML-procjena koristit ćemo Laplaceove procjene s $\lambda = 1$:

$$\hat{P}(x_k|\mathcal{C}_j) = \frac{N_{kj} + 1}{N_j + K_k}\,.$$

Tako za procjene izglednosti dobivamo:

$$\begin{array}{llll}
P(x_1 = \text{svemir}|y = \text{da}) & = \frac{2+1}{3+3} = 1/2 & P(x_1 = \text{svemir}|y = \text{ne}) & = \frac{2+1}{5+3} = 3/8 \\
P(x_1 = \text{kriminalac}|y = \text{da}) & = \frac{0+1}{3+4} = 1/7 & P(x_1 = \text{kriminalac}|y = \text{ne}) & = \frac{2+1}{5+4} = 1/3 \\
P(x_1 = \text{sadašnjost}|y = \text{da}) & = \frac{1+1}{3+3} = 1/3 & P(x_1 = \text{sadašnjost}|y = \text{ne}) & = \frac{1+1}{5+3} = 1/4 \\
P(x_1 = \text{ne}|y = \text{da}) & = \frac{2+1}{3+2} = 3/5 & P(x_1 = \text{ne}|y = \text{ne}) & = \frac{2+1}{5+2} = 3/7
\end{array}$$

Za posteriorne vjerojatnosti sada dobivamo:

$$\begin{aligned}
P(\text{da}|\mathbf{x}^{(2)}) &= \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{da}) \\
&= P(\text{svemir}|\text{da})P(\text{kiminalac}|\text{da})P(\text{sadašnost}|\text{da})P(\text{ne}|\text{da})P(\text{da}) \\
&= 1/2 \times 1/7 \times 1/3 \times 3/5 \times 3/8 = 0.0054 \\
P(\text{ne}|\mathbf{x}^{(2)}) &= \prod_{k=1}^{n} P(x_k|y = \text{da})P(y = \text{ne}) \\
&= P(\text{svemir}|\text{ne})P(\text{kriminalac}|\text{ne})P(\text{sadašnost}|\text{ne})P(\text{ne}|\text{ne})P(\text{ne}) \\
&= 3/8 \times 1/3 \times 1/4 \times 3/7 \times 5/8 = 0.0084
\end{aligned}$$

Aposteriorno najvjerojatnija klasifikacija i dalje je $y = \text{ne}$, međutim razlika između aposteriornih vjerojatnosti sada je nešto manja.

3 Polunaivan Bayesov klasifikator

Zaključili smo da procjena za $P(x_1, \ldots, x_n|\mathcal{C}_j)$ izravno za vektor $\mathbf{x} = (x_1, \ldots, x_n)$ nema smisla jer ne omogućava generalizaciju. S druge strane, pretpostavka naivnog Bayesa o uvjetnoj nezavisnosti varijabli x_i s obzirom na klasu $\mathcal{C}_j$ vrlo je radikalna i u praksi nije zadovoljena. Postavlja se pitanje: možemo li napraviti model koji bi bio između ove dvije krajnosti? Takav model neke bi varijable tretirao kao uvjetno nezavisne, dok bi druge tretirao zajednički. Npr. umjesto faktorizacije:

$$P(\mathcal{C}_j|x_1, x_2, x_3) \propto P(x_1|\mathcal{C}_j)P(x_2|\mathcal{C}_j)P(x_3|\mathcal{C}_j)P(\mathcal{C}_j)$$

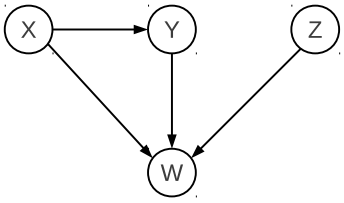
mogli bismo faktorizirati na sljedeći način:

$$P(\mathcal{C}_j|x_1, x_2, x_3) \propto P(x_1|\mathcal{C}_j)P(x_2, x_3|\mathcal{C}_j)P(\mathcal{C}_j)$$

ili, ekvivalentno:

$$P(\mathcal{C}_j|x_1, x_2, x_3) \propto P(x_1|\mathcal{C}_j)P(x_2|\mathcal{C}_j)P(x_3|x_2, \mathcal{C}_j)P(\mathcal{C}_j)$$

pogotovo ukoliko se pokaže da ne vrijedi $x_2 \perp x_3|\mathcal{C}_j$, odnosno da varijable x_2 i nisu x_3 uvjetno nezavisne za klasu $\mathcal{C}_j$.



Slika 2: Bayesova mreža za četiri varijable.

Ovakav model, kod kojega za neke varijable pretpostavljamo uvjetnu nezavisnost, dok za druge varijable to ne pretpostavljamo i modeliramo ih zajednički, nazivamo **polunaivan Bayesov klasifikator** (engl. *seminaïve Bayes classifier*). Očito je da će polunaivan model točnije modelirati podatke i davati točnije procjene. S druge strane, polunaivan model je složeniji od naivnog modela, odnosno ima više parametara i njegovo je učenje složenije. Naime, kada varijable tretiramo zajednički, broj parametara eksponencijalno ovisi o broju varijabli, a ne više linearno. Npr., za faktore $P(x_2|\mathcal{C}_j)P(x_3|\mathcal{C}_j)$ treba procijeniti $K(K_2 - 1) + K(K_3 - 1)$ parametara, dok za faktor $P(x_2, x_3|\mathcal{C}_j)$ treba procijeniti $K(K_2K_3 - 1)$ parametara. Posljedično, polunaivan model ima manju pristranost i veću varijancu, pa ga je lakše prenaučiti. Taj se problem međutim u načelu može izbjeći pravilnim odabirom modela, primjerice unakrsnom provjerom.

Prije nego što se uopće krene učiti model polunaivnog Bayesovog klasifikatora, potrebno je odrediti koje varijable treba tretirati zajednički, a koje se mogu tretirati kao uvjetno nezavisne za danu klasu. Kada je broj varijabli malen, moguće je iscrpno ispitati sve moguće modele.[2] Za realne probleme (modele s desetak ili više varijabli) tako nešto nije izvedivo, već se moramo osloniti na heurističke metode optimizacije. Nekoliko takvih postupaka razmotrit ćemo u nastavku.

3.1 Bayesova mreža

Za daljnje razmatranje bit će nam korisno Bayesov klasifikator grafički prikazati kao usmjereni graf. Općenito, modele koji koriste graf kako bi na sažet način prikazali zajedničku distribuciju nazivamo **probabilistički grafički modeli** (engl. *probabilistic graphical models*). Posebice, grafički model kod kojega je graf aciklički i usmjeren nazivamo **Bayesova mreža** (engl. *Bayesian networks*).[3] Čvorovi Bayesove mreže predstavljaju slučajne varijable, a lukovi predstavljaju zavisnosti između varijabli. Ako varijabla X zavisi o varijabli o Y, tada crtamo lûk od čvora Y do čvora X.

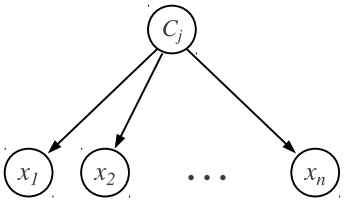
Na slici 2 prikazana je jednostavna Bayesova mreža za četiri varijable. Ova mreža predstavlja grafički zapis zajedničke distribucije $P(X, Y, Z, W)$, faktorizirane kao:

$$P(X, Y, Z, W) = P(X)P(Y|X)P(Z)P(W|X, Y, Z)\,.$$

Nepostojanje određenih lukova upućuje na (uvjetnu) nezavisnost između varijabli. Tako su varijable X i Z nezavisne od svih drugih varijabli, varijabla Y ovisi samo o varijabli X (odnosno varijabla Y uvjetno je nezavisna od drugih varijabli, ako je dana varijabla X), dok varijabla W ovisi o sve tri varijable.

[2] Broj mogućih združivanja jednak je broju particija nad n varijabli i dan je Bellovim brojem B_n. Primjerice $B_3 = 5$, $B_5 = 52$, $B_{10} = 115975$.

[3] Također: *mreža vjerovanja* (engl. *belief network*) i *usmjeren aciklički grafički model*.



Slika 3: Naivan Bayesov klasifikator prikazan kao Bayesova mreža.

Općenito, Bayesova mreža sažeto zapisuje funkciju gustoće vjerojatnosti $p(\mathbf{x})$ kao:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\text{pa}(x_i))$$

gdje je $\text{pa}(x_i)$ skup čvorova roditelja čvora x_i.

Na slici 3 prikazana je Bayesova mreža za naivan Bayesov klasifikator. Između varijabli x_i ne postoje izravne zavisnosti, no postoje zavisnosti između klase $\mathcal{C}_j$ i svake od varijabli x_i. Modeli polunaivnog Bayesovog klasifikatora razlikovat će se od ovog modela po tome što će neke od varijabli x_i biti združene u zajednički čvor, ili po tome što će između nekih od varijabli x_i postojati lukovi koji modeliraju zavisnost.

3.2 Algoritam FSSJ

Algoritam **unaprijednog slijednog odabira i združivanja** (engl. *forward sequential selection and joining*, FSSJ) (Pazzani, 1997) izgrađuje model polunaivnog Bayesovog klasifikatora na način da minimizira njegovu pogrešku. Algoritam kreće s nepovezanom Bayesovom mrežom. U svakoj idućoj iteraciji algoritam nastoji neku od varijabli x_i ili povezati s čvorem $\mathcal{C}_j$, ili je dodati u već postojeći čvor, ovisno o tome što daje manju pogrešku generalizacije. Riječ je zapravo o pohlepnom pretraživanju prostora stanja metodom usponom na vrh, pri čemu se kao kriterij optimizacije koristi pogreška generalizacije.

Algoritam FSSJ

1. Inicijaliziraj skup varijabli x_i koje se koriste u modelu na prazan skup. Graf je početno nepovezan, pa

   $$P(x_1, \ldots, x_n, \mathcal{C}_j) = P(x_1) \cdots P(x_n) P(\mathcal{C}_j)$$

   odnosno

   $$P(\mathcal{C}_j|x_1, \ldots, x_n) = P(\mathcal{C}_j)\,.$$

   Klasificiraj sve primjere iz skupa za provjeru u klasu $\mathcal{C}^*$ koja je najzastupljenija u skupu za učenje, $\mathcal{C}^* = \text{argmax}_j\, P(\mathcal{C}_j)$.

2. Za varijablu x_i koja još nije uključena u model, razmotri sljedeće operacije:

   (a) Uključivanje varijable x_i u model tako da se ona doda kao uvjetno nezavisna u odnosu na ostale varijable za danu klasu $\mathcal{C}_j$, tj. dodavanje lûka $(\mathcal{C}_j, x_i)$.

   (b) Uključivanje varijable u model tako da se ona doda u zajednički čvor (tzv. superčvor) s nekom već uključenom varijablom.

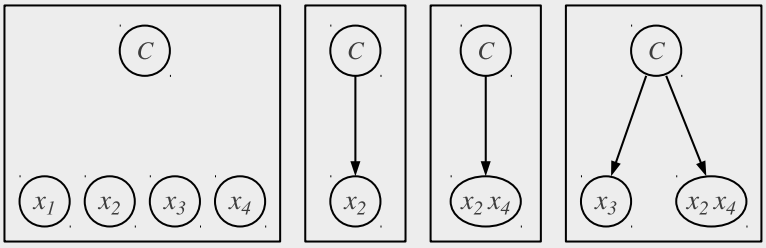
3. Izaberi varijablu i operaciju koja minimizira pogrešku generalizacije.

4. Ponavljaj od koraka 2 sve dok nema novih poboljšanja pogreške generalizacije.

Opisani algoritam je pohlepan jer nikada ne preispituje svoje odluke. Zbog toga nema garancije da je pronađeni model doista optimalan. (Teoretski je moguće da model koji izgradi algoritam FSSJ bude lošiji od modela naivnog Bayesovog klasifikatora.)

Treba naglasiti da je bitno da se optimizacija provodi unakrsnom provjerom, odnosno temeljem pogreške generalizacije (mjerene na skupu za provjeru), a ne temeljem empirijske pogreške (mjerene na skupu za učenje). Naime, minimalnu empirijsku pogrešku ostvario bi potpuno nefaktoriziran model (model u kojem je svih n varijabli združeno), ali, kao što smo već smo naglasili, takav bi model imao veliku pogrešku generalizacije.

Primjer 7 (Algoritam FSSJ) Razmotrimo izgradnju polunaivnog Bayesovog klasifikatora za četiri ulazne varijable: x_1, x_2, x_3, x_4. Jedan mogući slijed koraka izgradnje klasifikatora je ovaj:



Modelu odgovara aposteriorna razdioba

$$P(\mathcal{C}|x_1, x_2, x_3, x_4) \propto P(x_3|\mathcal{C})P(x_2, x_4|\mathcal{C})P(\mathcal{C})\,.$$

Označimo sa $E(x_i, [x_k, x_l])$ pogrešku generalizacije modela koji varijable x_k i x_l tretira združeno i uvjetno nezavisno od varijable x_i. Na temelju slijeda koraka algoritma FSSJ, možemo zaključiti da za vrijedi:

$$\begin{aligned}
E(x_2) &\leqslant E(x_i),\ i = 1, 3, 4\\
E([x_2, x_4]) &\leqslant E(x_2, x_i),\ i = 1, 3, 4\\
E([x_2, x_4]) &\leqslant E([x_2, x_i]),\ i = 1, 3\\
E([x_2, x_4], x_3) &\leqslant E([x_2, x_4], x_1)\\
E([x_2, x_4], x_3) &\leqslant E([x_2, x_4, x_i]),\ i = 1, 3\\
E([x_2, x_4], x_3) &\leqslant E(x_1, [x_2, x_4], x_3)\\
E([x_2, x_4], x_3) &\leqslant E([x_1, x_2, x_4], x_3)\\
E([x_2, x_4], x_3) &\leqslant E([x_2, x_4], [x_1, x_3])
\end{aligned}$$

Primijetite da algoritam u konačan model nije uključio varijablu x_1. Algoritam FSSJ dakle ujedno provodi i **odabir značajki** (engl. *feature selection*).

Također primijetite da je algoritam FSSJ pohlepan. Npr., u ovom primjeru algoritam je provjerio i odbacio samo 15 modela od ukupno 52 moguća.[4] Možda se globalni minimum doseže sa $E([x_1, x_4], x_2, x_3)$, no taj model algoritam FSSJ nije provjeravao.

3.3 Klasifikator TAN

TAN (engl. *tree augmented naive Bayes*) (Friedman et al., 1997) je polunaivan Bayesov klasifikator čija se izgradnja temelji na mjeri uzajamne informacije među varijablama. **Uzajamna informacija** (engl. *mutual information*) između slučajnih varijabli X i Y definirana je kao:

$$I(X,Y) = \sum_{j=1}^{K_Y} \sum_{i=1}^{K_X} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \,. \tag{17}$$

U okviru teorije informacije, $I(X,Y)$ se tumači kao količina informacije koju varijable X i Y međusobno dijele, odnosno u kojoj mjeri poznavanje vrijednosti jedne od varijabli smanjuje nesigurnost o vrijednosti druge varijable. Mjera uzajamne informacije kvantificira stupanj stohastičke zavisnosti među varijablama i to je veća što je omjer između $P(X,Y)$ i $P(X)P(Y)$ manji; varijable X i Y su nezavisne akko $I(X,Y) = 0$. **Uvjetna uzajamna informacija** (engl. *conditional mutual information*) definirana je kao

$$I(X,Y|Z) = \sum_{k=1}^{K_Z} P(z_k) I(X,Y|z_k) = \sum_{k=1}^{K_Z} \sum_{j=1}^{K_Y} \sum_{i=1}^{K_X} P(x_i, y_j, z_k) \log \frac{P(x_i, y_j|z_k)}{P(x_i|z_k)P(y_j|z_k)} \,. \tag{18}$$

Klasifikator TAN zasniva se na ideji da je varijable za koje je mjera uvjetne uzajamne informacije $I(x_i, x_j|\mathcal{C}_j)$ najveća potrebno modelirati kao zavisne. Počevši od nepovezane Bayesove mreže, algoritam nastoji povezati parove čvorova s najvećom mjerom uvjetne uzajamne informacije, osim ako bi takvo povezivanje narušilo svojstvo acikličnosti grafa.

Algoritam izgradnje klasifikatora TAN

1. Izračunaj $I(x_i, x_j|\mathcal{C})$ za $i < j$, $i = 1, \ldots, n$ i sortiraj silazno.
2. Izgradi nepovezanu Bayesovu mrežu s čvorovima $x_1, \ldots, x_n$.
3. Razmotri par (x_i, x_j) s najvećom vrijednošću $I(x_i, x_j|\mathcal{C})$ i dodaj (neusmjereni) brid (x_i, x_j) ako time ne nastaje ciklus. Inače preskoči taj par i razmatraj idući par u listi.
4. Ponavljaj korak 3 dok ne izgradiš $n - 1$ bridova.
5. Pretvori neusmjeren graf u usmjeren graf tako da nasumično odabereš jedan čvor kao korijen.
6. Dodaj čvor $\mathcal{C}$ i poveži ga lukovima sa svim ulaznim varijablama.

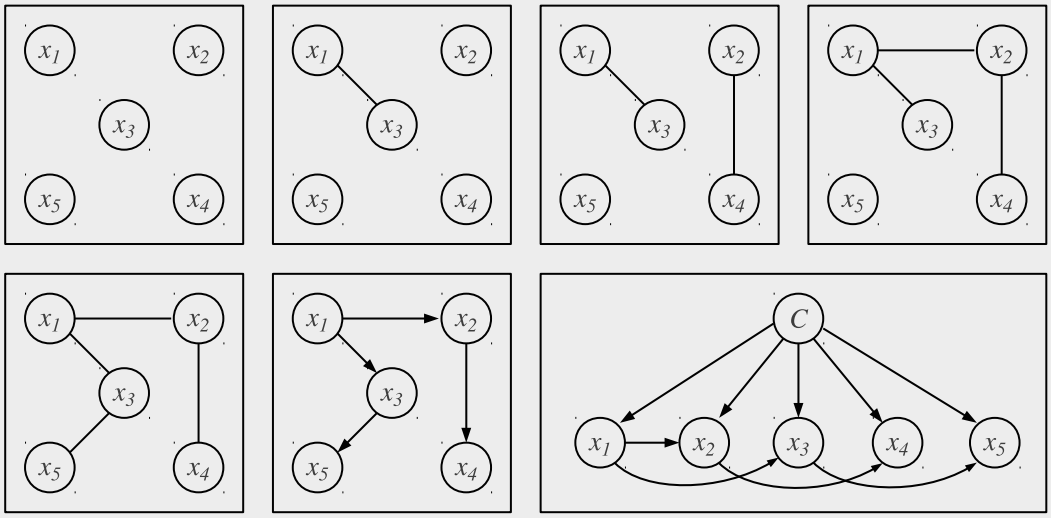
Prva četiri koraka moguće je formulirati i na ovaj način: izgradi neusmjeren potpuno povezan težinski graf, kod kojega je težina brida (x_i, x_j) jednaka $I(x_i, x_j|\mathcal{C}_j)$, a zatim izračunaj razapinjući put maksimalne težine.

[4] Budući da je moguće ne uključiti neku od varijabli, ukupan broj modela jednak je Bellovom broju B_{n+1} (jedna dodatna particija je za sve neuključene čvorove).

Primjer 8 (Klasifikator TAN) Izgradimo klasifikator TAN za ulazne varijable $x_1, \ldots, x_5$. Pretpostavimo da na skupu za učenje vrijedi:

$$I(x_1, x_3|\mathcal{C}) > I(x_2, x_4|\mathcal{C}) > I(x_1, x_2|\mathcal{C}) > I(x_3, x_4|\mathcal{C}) > I(x_1, x_4|\mathcal{C}) > I(x_3, x_5|\mathcal{C}) > I(x_1, x_5|\mathcal{C}) > I(x_2, x_3|\mathcal{C}) > I(x_2, x_5|\mathcal{C}) > I(x_4, x_5|\mathcal{C}) .$$

Iz ovih je odnosa vidljivo da su, uz danu klasu $\mathcal{C}_j$, varijable x_1 i x_3 na skupu za učenje najmanje, a varijable x_4 i x_5 najviše uvjetno nezavisne. Koraci izgradnje klasifikatora TAN su:



Kao korijenski čvor izabran je čvor x_1. Dobiveni model odgovara aposteriornoj razdiobi

$$P(\mathcal{C}|x_1, x_2, x_3, x_4, x_5) \propto P(x_1|\mathcal{C})P(x_2|x_1, \mathcal{C})P(x_3|x_1, \mathcal{C})P(x_4|x_2, \mathcal{C})P(x_5|x_3, \mathcal{C})P(\mathcal{C}) .$$

Opisani postupak zaslužuje nekoliko komentara. Nasumičan odabir korijenskog čvora na prvi se pogled može činiti kao nepoželjna nedeterminističnost. Međutim, odabir korijenskog čvora je nebitan jer nema utjecaja na konačan oblik zajedničke distribucije – zajednička distribucija može se faktorizirati na više načina budući da vrijedi $P(X|Y)P(Y) = P(Y|X)P(X)$. Treba međutim naglasiti da se vrijednosti $I(x_i, x_j|\mathcal{C}_j)$ računaju na temelju skupa za učenje $\mathcal{D}$ (npr. temeljem ML-procjena), pa se dakle radi o procjeni uvjetne uzajamne informacije, a ne o pravoj vrijednosti. Kao i uvijek kad je riječ o valjanim procjeniteljima, procjena će biti to točnija što je uzorak (odnosno broj primjera) veći.

Koja je razlika između modela izgrađenog algoritmom FSSJ, kod kojega se čvorovi združuju, i modela TAN, kod kojega se čvorovi povezuju? Razlika je u složenosti modela, odnosno u broju parametara. Klasifikator TAN može modelirati zavisnosti između pojedinačnih parova varijabli za koje zavisnost ne vrijedi tranzitivno, dok klasifikator izgrađen algoritmom FSSJ to ne može. Zbog toga će broj parametara modela TAN biti manji (v. primjer 9).

Primjer 9 (Broj parametara Bayesovog klasifikatora) Želimo izgraditi klasifikator s pet ulaznih varijabli, $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$. Pretpostavimo da sve varijable imaju tri

moguće vrijednosti, a da je klasifikator binaran. Nadalje pretpostavimo da su sve varijable uvjetno nezavisne uz danu klasu $\mathcal{C}_j$, osim parova varijabli x_1 i x_2 te varijabli x_2 i x_3, koje uvjetno zavisne uz danu klasu $\mathcal{C}_j$.

Izračunajmo ukupan broj parametara različitih varijanti diskretnih Bayesovih klasifikatora. Broj parametara ovisi o načinu faktorizacije zajedničke distribucije $P(\mathbf{x}, \mathcal{C}_j)$. Kod naivnog Bayesovog klasifikatora imamo

$$P(\mathbf{x}, \mathcal{C}_j) = P(\mathcal{C}_j) \prod_{i=1}^{5} P(x_i|\mathcal{C}_j)$$

pa je broj parametara jednak $(2-1)+5\times 2\times (3-1) = 21$. Naivan Bayesov klasifikator ima najmanje parametara, ali u potpunosti zanemaruje zavisnost koja postoji između varijabli.

Polunaivan Bayesov klasifikator izgrađen algoritmom FSSJ može uzeti u obzir navedene zavisnosti između varijabli jedino tako da u jedan superčvor združi varijable x_1, x_2 i x_3:

$$P(\mathbf{x}, \mathcal{C}_j) = P(x_1, x_2, x_3|\mathcal{C}_j)P(x_4|\mathcal{C}_j)P(x_5|\mathcal{C}_j)P(\mathcal{C}_j)$$

što daje ukupno $2 \times (3^3 - 1) + 2 \times 2 \times (3 - 1) + (2 - 1) = 61$ parametar.

Klasifikator TAN zavisnost može modelirati s manje parametara, budući da može zasebno modelirati zavisnosti između parova varijabli x_1 i x_2 te x_2 i x_3:

$$P(\mathbf{x}, \mathcal{C}_j) = P(x_1|x_2, \mathcal{C}_j)P(x_2|x_3, \mathcal{C}_j)P(x_4|\mathcal{C}_j)P(x_5|\mathcal{C}_j)P(\mathcal{C}_j)$$

što daje ukupno $2 \times 2 \times 3 \times (3 - 1) + 2 \times 2 \times (3 - 1) + (2 - 1) = 33$ parametara.

Općenito, za faktor $P(X_1, \ldots, X_i|X_{i+1}, \ldots, X_n)$, kod kojeg svaka varijabla ima K stanja, treba procijeniti $\mathcal{O}(K^n)$ parametara. Ako taj broj želimo smanjiti, faktori trebaju biti što kraći (bolje je imati veći broj kraćih faktora nego manji broj duljih faktora).

3.4 Klasifikator k-DB

Kod klasifikatora TAN svaki čvor može imati samo jednog roditelja, odnosno svaka varijabla može, pored zavisnosti o klasi, biti uvjetno zavisna samo o jednoj ulaznoj varijabli. Bayesov model koji omogućava veći broj ovisnosti naziva se **klasifikator k-DB** (engl. *k-Limited Dependence Bayesian Classifiers*) (Sahami, 1996). Klasifikator k-DB može modelirati zavisnost varijable o najviše k drugih varijabli (ne računajući varijablu $\mathcal{C}_j$). Odabir čvorova između kojih se modeliraju zavisnosti provodi se, kao i kod klasifikatora TAN, temeljem mjere uvjetne uzajamne informacije $I(x_i, x_j|\mathcal{C}_j)$. Dodatno, varijable se u model uključuju redom prema relevantnosti, izračunate pomoću mjere uzajamne informacije $I(x_i, \mathcal{C}_j)$ između ulazne varijable i varijable klase. Model k-DB zavisnosti među varijablama može modelirati točnije nego model TAN, uz cijenu većeg broja parametara i veće sklonosti prenaučenosti.

Algoritam izgradnje klasifikatora k-DB

1. Izračunaj $I(x_i, \mathcal{C}_j)$ i $I(x_i, x_j|\mathcal{C}_j)$ za svaki par varijabli. Sortiraj varijable silazno po $I(x_i, \mathcal{C}_j)$.

2. Za varijablu x_i koja je prva u listi:

(a) Dodaj varijablu x_i u model i izbaci je iz liste.

(b) Postavi čvor $\mathcal{C}_j$ za roditelja čvora x_i.

(c) Od varijabli koje su već uključene u model, njih k (ili manje, ako ih nema toliko) koje imaju najveću vrijednost $I(x_i, x_j|\mathcal{C}_j)$ postavi kao čvorove roditelje od x_i.

3. Ponavljaj prethodni korak dok lista nije prazna.

Budući da su svi uvijek usmjereni prema čvoru koji se u tekućoj iteraciji dodaje, a nikad od njega, konačan graf neće sadržavati usmjerene cikluse.

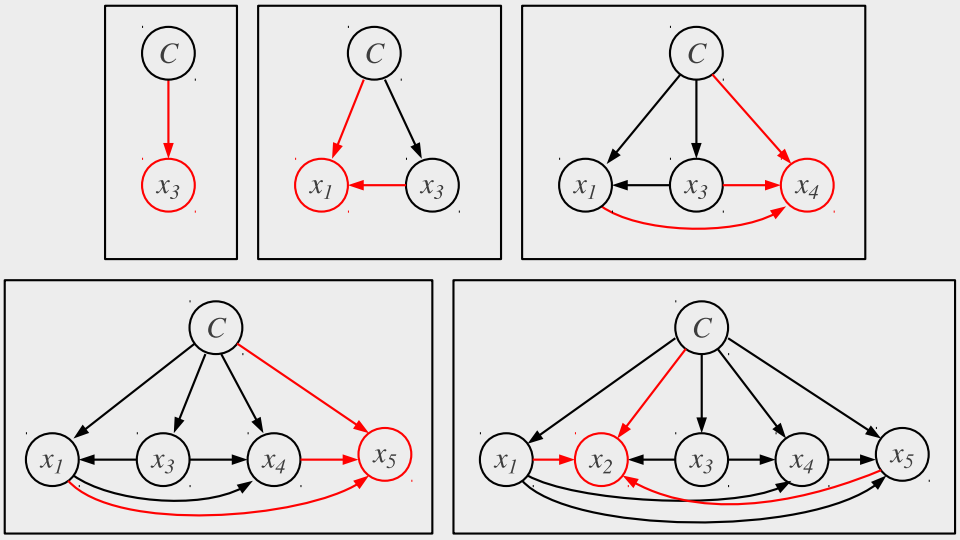
Primjer 10 (Klasifikator 2-DB) Izgradimo klasifikator 2-DB nad ulaznim varijablama $x_1, \ldots, x_5$. Neka su uzajamne informacije, procijenjene na skupu za učenje $\mathcal{D}$, takve da vrijedi:

$$I(x_3, \mathcal{C}) > I(x_1, \mathcal{C}) > I(x_4, \mathcal{C}) > I(x_5, \mathcal{C}) > I(x_2, \mathcal{C})$$

tj. varijabla $\mathcal{C}_j$ najviše ovisi o varijabli x_3, a najmanje o varijabli x_2. Ovim će redoslijedom varijable biti uključivane u model. Neka su uvjetne uzajamne informacije takve da vrijedi:

$$I(x_3, x_4|\mathcal{C}) > I(x_2, x_5|\mathcal{C}) > I(x_1, x_3|\mathcal{C}) > I(x_1, x_2|\mathcal{C}) > I(x_2, x_4|\mathcal{C}) > I(x_2, x_3|\mathcal{C}) > I(x_1, x_4|\mathcal{C}) > I(x_4, x_5|\mathcal{C}) > I(x_1, x_5|\mathcal{C}) > I(x_3, x_5|\mathcal{C})$$

tj. uz danu klasu $\mathcal{C}$, varijable x_3 i x_4 su najviše, a varijable x_3 i x_5 najmanje zavisne. Koraci izgradnje klasifikatora k-DB su (lukovi i čvorovi koji se dodaju obojani su crveno):



Dobiveni model odgovara aposteriornoj razdiobi

$$P(\mathcal{C}|x_1, x_2, x_3, x_4, x_5) \propto P(x_1|x_3, \mathcal{C})P(x_2|x_1, x_5, \mathcal{C})P(x_3|\mathcal{C})P(x_4|x_1, x_3, \mathcal{C}) P(x_5|x_1, x_4, \mathcal{C})P(\mathcal{C})\,.$$

3.5 Bayesova mreža kao klasifikator

Modeli polunaivnog Bayesovog klasifikatora koje smo ovdje opisali temelje se na ideji proširenja naivnog modela tako da se u obzir uzmu neke od zavisnosti među varijablama. Problemu klasifikacije moguće je pristupiti i obrnuto, na način da se najprije izgradi potpuna Bayesova mreža koja modelira neki problem, a zatim se lokalni dijelovi te mreže koriste za klasifikaciju. Može se pokazati da vrijednost svake varijable u Bayesovoj mreži ovisi samo o roditeljskim čvorovima, njihovoj djeci, i roditeljima čvorova djece. Taj se skup čvorova naziva **Markovljev omotač** (engl. *Markov blanket*). Klasifikacija se svodi na zaključivanje nad čvorovima koji čine Markovljev omotač, za što se može upotrebiti bilo koji od algoritama zaključivanja nad Bayesovom mrežom. Za izgradnju mreže možemo upotrijebiti neki od algoritama za učenje strukture mreže (npr. algoritmi PC ili K2).

4 Bayesov klasifikator za kontinuirane varijable

Ako su varijable kontinuirane, izglednosti klasa tipično se modeliraju Gaussovom razdiobom. Osnovni razlog za to je analitička jednostavnost. Pored toga, mnogi prirodni fenomeni doista se pokoravaju normalnoj razdiobi u smislu da se primjere jedne klase može tretirati kao vrijednosti koje blago odstupaju od neke srednje vrijednosti. Treba naravno napomenuti da kontinuirane varijable ne moraju uvijek biti normalno distribuirane i da postoje statistički testovi kojima se može utvrditi je li to doista slučaj (npr. Kolmogorov-Smirnovov test ili Shaipro-Wilkov test). Unatoč tome, u većini slučajeva normalna je razdioba dovoljno dobra aproksimacija stvarnih podataka, pod uvjetom da primjeri iz pojedine klase oblikuju jednu grupu (u protivnom treba koristiti model Gaussove mješavine).

Razmotrimo prvo jednodimenzijski (univarijatan) Bayesov klasifikator, a zatim višedimenzijski (multivarijatan) Bayesov klasifikator za kontinuirane varijable. Jednodimenzijski klasifikator nije u praksi zanimljiv, ali razmatramo ga radi boljeg razumijevanja višedimenzijskog klasifikatora.

4.1 Jednodimenzijski Bayesov klasifikator

Kod jednodimenzijskog (univarijatnog) Bayesovog klasifikatora, izglednost svake klase $\mathcal{C}_j$ modeliramo jednodimenzijskom Gaussovom gustoćom $\mathcal{N}(\mu_j, \sigma_j^2)$:

$$p(x|\mathcal{C}_j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{(x-\mu_j)^2}{2\sigma_j^2} \right\}. \tag{19}$$

Model možemo definirati kao

$$h_j(x) = p(x|\mathcal{C}_j)P(\mathcal{C}_j)$$

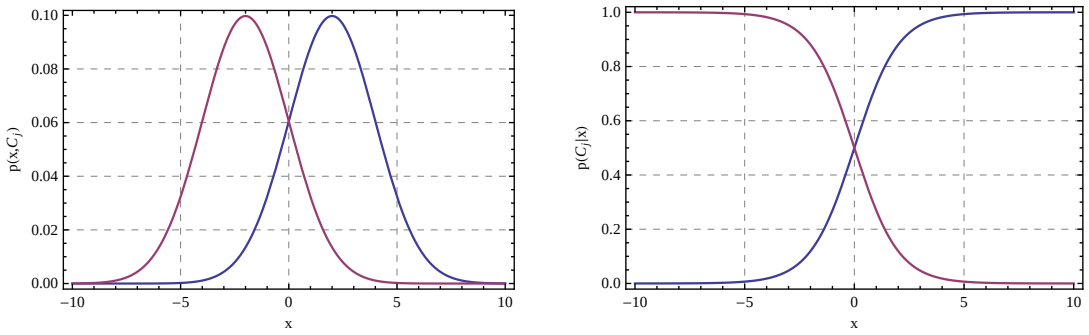
što je, budući da nas zanima samo maksimizacija, istovjetno s

$$h_j(x) = \ln p(x|\mathcal{C}_j) + \ln P(\mathcal{C}_j)\,. \tag{20}$$

Uvrštenjem (19) u (20) dobivamo

$$h_j(x|\boldsymbol{\theta}_j) = -\frac{1}{2}\ln 2\pi - \ln\sigma_j - \frac{(x-\mu_j)^2}{2\sigma_j^2} + \ln P(\mathcal{C}_j) \tag{21}$$

čime je definiran parametarski model koji se sastoji od po tri parametra za svaku klasu, $\boldsymbol{\theta}_j = (\mu_j, \sigma_j, P(\mathcal{C}_j))$.



(a) zajednička gustoća $p(x, \mathcal{C}_j)$ (b) aposteriorne vjerojatnosti $p(\mathcal{C}_j|x)$

Slika 4: Dvoklasni Gaussov model za klase jednakih varijanci i jednakih apriornih vjerojatnosti: $p(x|\mathcal{C}_1) \sim \mathcal{N}(-2, 4)$ i $p(x|\mathcal{C}_2) \sim \mathcal{N}(2, 4)$.

Raspolažemo skupom primjera za učenje $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$, gdje $y^{(i)} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$. Na temelju primjera iz skupa $\mathcal{D}$, parametre μ_j i σ_j^2 možemo procijeniti metodom najveće izglednosti, zasebno za svaku klasu $\mathcal{C}_j$. Tako dobivamo:

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} x^{(i)} \qquad \hat{\sigma}_j^2 = \frac{1}{N_j} \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = \mathcal{C}_j\} (x^{(i)} - \hat{\mu}_j)^2 \tag{22}$$

dok za procjenu apriornih vjerojatnosti $P(\mathcal{C}_j)$ koristimo (8). Uvrštavanjem ovih procjena u (21) dobivamo model

$$h_j(x) = -\ln \hat{\sigma}_j - \frac{(x - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2} + \ln \hat{P}(\mathcal{C}_j) \tag{23}$$

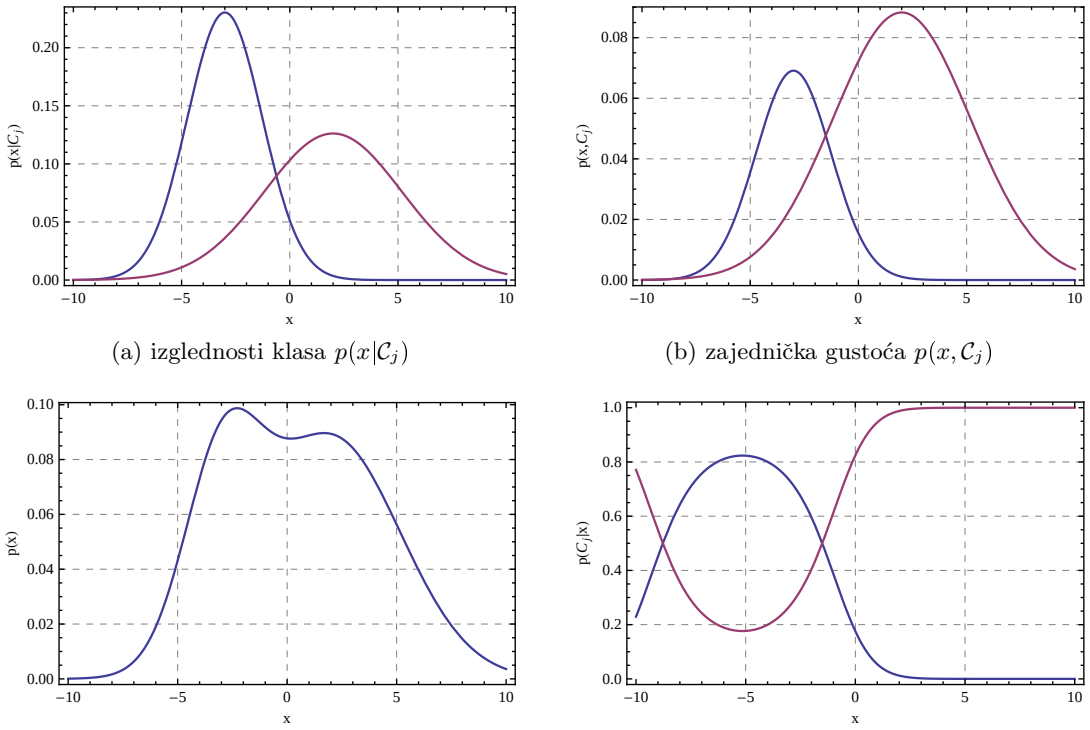
pri čemu smo prvi pribrojnik iz (21) zanemarili, budući da je za sve funkcije h_j identičan.

Na slici 4 prikazan je model za slučaj dviju klasa jednakih varijanci, $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$, i jednakih apriornih vjerojatnosti, $\hat{P}(\mathcal{C}_1) = \hat{P}(\mathcal{C}_2)$. U ovom slučaju granica između područja nalazi se točno na polovici između srednjih vrijednosti $\hat{\mu}_1$ i $\hat{\mu}_2$, budući da tada, prema (21), vrijedi:

$$\begin{aligned} h_1(x) &= h_2(x) \\ (x - \hat{\mu}_1)^2 &= (x - \hat{\mu}_2)^2 \\ x &= \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2) \end{aligned}$$

Na slici 5 prikazan je model za općenitiji slučaj dviju klasa s različitim srednjim vrijednostima, varijancama i apriornim vjerojatnostima. Zato što su varijance različite, funkcije aposteriorne vjerojatnosti presijecaju u dvije točke, pa postoje dvije granice između klasa. Također, zato što su apriorne vjerojatnosti klasa različite, granica područja primiče se klasi s manjom apriornom vjerojatnošću.

Treba naglasiti da smo, definiravši hipotezu kao (4.1) odnosno (20), izgubili vjerojatnosnu interpretaciju hipoteze. To znači da vrijednost hipoteze $h_j(x)$ ne možemo tumačiti kao vjerojatnost da primjer x pripada klasi $\mathcal{C}_j$. To je vidljivo iz usporedbe slika 5d i 5b. Primjerice, vjerojatnost da primjer $x = 5$ pripada klasi $\mathcal{C}_2$ približno je jednaka vjerojatnosti da primjer $x = 10$ pripada istoj toj klasi. Međutim, nenormalizirana vrijednost, odnosno vrijednost zajedničke gustoće $p(x, \mathcal{C}_1)$, mnogo je manja za $x = 10$ nego za $x = 5$. Vidimo dakle da, ako hipoteza nije normalizirana, ne možemo uspoređivati rezultate klasifikacija



(a) izglednosti klasa $p(x|\mathcal{C}_j)$

(b) zajednička gustoća $p(x, \mathcal{C}_j)$

(c) gustoća primjera $p(x)$

(d) aposteriorne vjerojatnosti $P(\mathcal{C}_j|x)$

Slika 5: Dvoklasni Gaussov model. Klase su modelirane izglednostima $p(x|\mathcal{C}_1) \sim \mathcal{N}(-3, 3)$ i $p(x|\mathcal{C}_2) \sim \mathcal{N}(2, 10)$. Apriorna vjerojatnosti klasa su $P(\mathcal{C}_1) = 0.3$ i $P(\mathcal{C}_2) = 0.7$.

različitih primjera. Međutim, možemo uspoređivati pouzdanosti klasifikacije pojedinačnog primjera u različite klase. Tako možemo zaključiti da je klasifikacija primjera $x = 5$ u klasu $\mathcal{C}_2$ znatno pouzdanija nego klasifikacija istog tog primjera u klasu $\mathcal{C}_1$.

Primjer 11 (Jednodimenzijski Bayesov klasifikator) Želimo naučiti univarijatni Bayesov klasifikator za klasifikaciju primjera u tri klase, $\mathcal{C}_1$, $\mathcal{C}_2$ i $\mathcal{C}_3$. Prostor primjera je $\mathcal{X} = \mathbb{R}$, a skup primjera za učenje je

$$\mathcal{D} = \big\{(-1.52, \mathcal{C}_2), (1.28, \mathcal{C}_1), (0.56, \mathcal{C}_1), (4.15, \mathcal{C}_3), (3.36, \mathcal{C}_1), (-7.59, \mathcal{C}_2), (-1.14, \mathcal{C}_1), (4.05, \mathcal{C}_3), (4.09, \mathcal{C}_1), (5.24, \mathcal{C}_3), (3.72, \mathcal{C}_3), (-1.07, \mathcal{C}_1), (0.09, \mathcal{C}_2), (3.68, \mathcal{C}_3), (2.49, \mathcal{C}_3)\big\}.$$

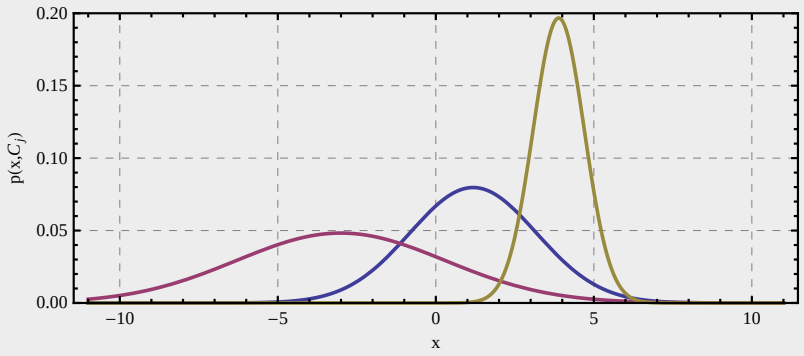
Primjeri poprimaju kontinuirane vrijednosti, pa izglednosti klasa modeliramo Gaussovim gustoćama. Procjenu parametara gustoća $p(x|\mathcal{C}_j)$, $j = 1, 2, 3$, računamo prema (22):

$$\hat{\mu}_1 = 1.18 \qquad \hat{\sigma}_1^2 = 4.01$$
$$\hat{\mu}_2 = -3.01 \qquad \hat{\sigma}_2^2 = 10.94$$
$$\hat{\mu}_3 = 3.89 \qquad \hat{\sigma}_3^2 = 0.66$$

(Primijetite da su procjene $\hat{\sigma}_j^2$ pristrane i da ih se može ispraviti tako da ih se pomnoži s $N/(N-1)$.) Procjenu apriornih vjerojatnosti klasa računamo prema (8) i dobivamo

$$\hat{P}(\mathcal{C}_1) = 0.4 \qquad \hat{P}(\mathcal{C}_2) = 0.2 \qquad \hat{P}(\mathcal{C}_3) = 0.4$$

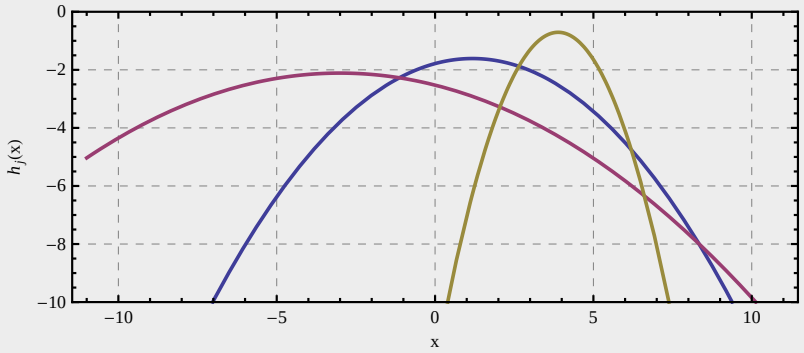
Zajedničke gustoće $p(x, \mathcal{C}_j) = p(\mathcal{C}_j|x)P(\mathcal{C}_j)$ izgledaju ovako



Uvrštavanjem u (23) za klase $\mathcal{C}_1$, $\mathcal{C}_2$ i $\mathcal{C}_3$ dobivamo sljedeće hipoteze:

$$\begin{aligned} h_1(x) &= -0.12x^2 + 0.29x - 1.79 \\ h_2(x) &= -0.05x^2 - 0.28x - 2.53 \\ h_3(x) &= -0.76x^2 + 5.91x - 12.19 \end{aligned}$$

koje izgledaju ovako:



Granice između klasa mogu se odrediti rješavanjem (ne)jednadžbi $h_j(x) = h_k(x) \geqslant h_l(x)$ $(j \neq k,\, j \neq l,\, k \neq l)$.

4.2 Višedimenzijski Bayesov klasifikator

Kod višedimenzijskog slučaja, $\mathcal{X} = \mathbb{R}^n$, izglednosti klasa modeliramo multivarijatnom Gaussovom gustoćom $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$:

$$p(\mathbf{x}|\mathcal{C}_j) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}. \tag{24}$$

Vektor $\boldsymbol{\mu}_j$ opisuje prototipnu vrijednost primjera u klasi $\mathcal{C}_j$, dok matrica kovarijacije opisuje količinu šuma u svakoj varijabli te korelaciju između izvora šuma. Klasama odgovaraju hiperelipsoidi čije je usmjerenje određeno faktorima korelacije. Uvrštavanjem u model

$$h_j(\mathbf{x}) = \ln p(\mathbf{x}|\mathcal{C}_j) + \ln P(\mathcal{C}_j) \tag{25}$$

dobivamo

$$h_j(\mathbf{x}) = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln P(\mathcal{C}_j)\,. \tag{26}$$

Ovaj model ukupno ima Kn parametara za srednje vrijednosti te $\frac{Kn}{2}(n+1)$ parametara za matrice kovarijacije (prisjetite se da je kovarijacijska matrica simetrična pa je dovoljno pohraniti polovicu od n^2 parametara).

Raspišimo (26) kako bismo utvrdili je li model linearan ili nelinearan. Primijetite da je prvi pribrojnik jednak za sve klase, pa ga možemo zanemariti. Sređivanjem izraza dobivamo

$$h_j(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j) + \ln P(\mathcal{C}_j)\,. \tag{27}$$

Ovdje smo iskoristili jednakost $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j = \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}$, koja vrijedi jer je matrica $\boldsymbol{\Sigma}$ odnosno $\boldsymbol{\Sigma}^{-1}$ simetrična. Naime

$$\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = (\mathbf{x}|\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) = ((\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}\mathbf{x}|\boldsymbol{\mu}) = (\boldsymbol{\Sigma}^{-1}\mathbf{x}|\boldsymbol{\mu}) = (\boldsymbol{\mu}|\boldsymbol{\Sigma}^{-1}\mathbf{x}) = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}$$

pri čemu smo iskoristili svojstvo skalarnog produkta $(\mathbf{x}|\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ i svojstvo $(\mathbf{A}\mathbf{x}|\mathbf{y}) = (\mathbf{x}|\mathbf{A}\mathbf{y})$ koje vrijedi za simetričnu matricu $\mathbf{A}$. Izraz $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}$ je kvadratna forma, pa funkcija $h_j(\mathbf{x})$ kvadratno ovisi o $\mathbf{x}$. Posljedično, granica između dviju klasa, $\mathcal{C}_j$ i $\mathcal{C}_k$, koju možemo izvesti rješavanjem jednadžbe $h_j(\mathbf{x}) = h_k(\mathbf{x})$, bit će paraboloid (slika 6a). Model je dakle nelinearan.

Za dani skup za učenje $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ parametre modela procjenjujemo zasebno za svaku klasu $\mathcal{C}_j$. Ako koristimo metodu najveće izglednosti, procjene su:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_j &= \frac{1}{N_j}\sum_{i=1}^{N}\mathbf{1}\{y^{(i)} = \mathcal{C}_j\}\mathbf{x}^{(i)} \\
\hat{\boldsymbol{\Sigma}}_j &= \frac{1}{N_j}\sum_{i=1}^{N}\mathbf{1}\{y^{(i)} = \mathcal{C}_j\}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_j)^{\mathrm{T}} \\
\hat{P}(\mathcal{C}_j) &= \frac{N_j}{N}
\end{aligned}$$

Opisani model može vrlo točno modelirati podatke, ali je njegov nedostatak velik broj parametara. Ako je skup za učenje malen, vrlo je teško načiniti pouzdanu procjenu tih parametara. Pored toga, velik broj parametara često predstavlja i praktičan problem u smislu računalnih resursa. Kako je broj parametara $\mathcal{O}(n^2)$, problem je osobito izražen kod velikih dimenzija. Zbog toga ćemo u nastavku razmotriti niz mogućih pojednostavljenja ovog modela.

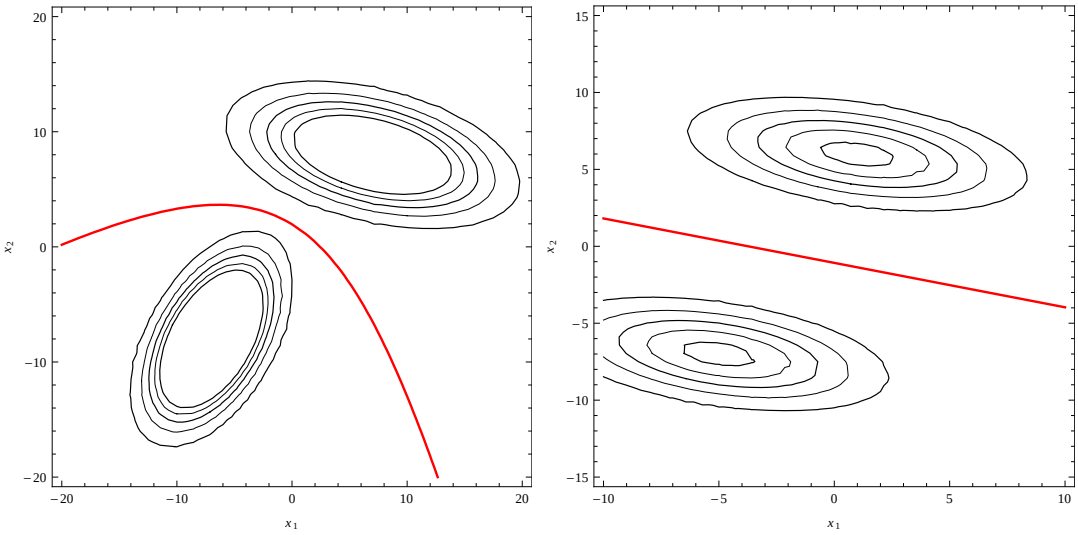
1. pojednostavljenje: dijeljenja kovarijacijska matrica

Model se može pojednostaviti ako se pretpostavi da je kovarijacijska matrica jednaka za svaku klasu, odnosno da je dijeljena između klasa. Takvu matricu možemo procijeniti kao kao težinski zbroj pojedinačnih kovarijacijskih matrica:

$$\hat{\boldsymbol{\Sigma}} = \sum_j \hat{P}(\mathcal{C}_j)\hat{\boldsymbol{\Sigma}}_j\,. \tag{28}$$

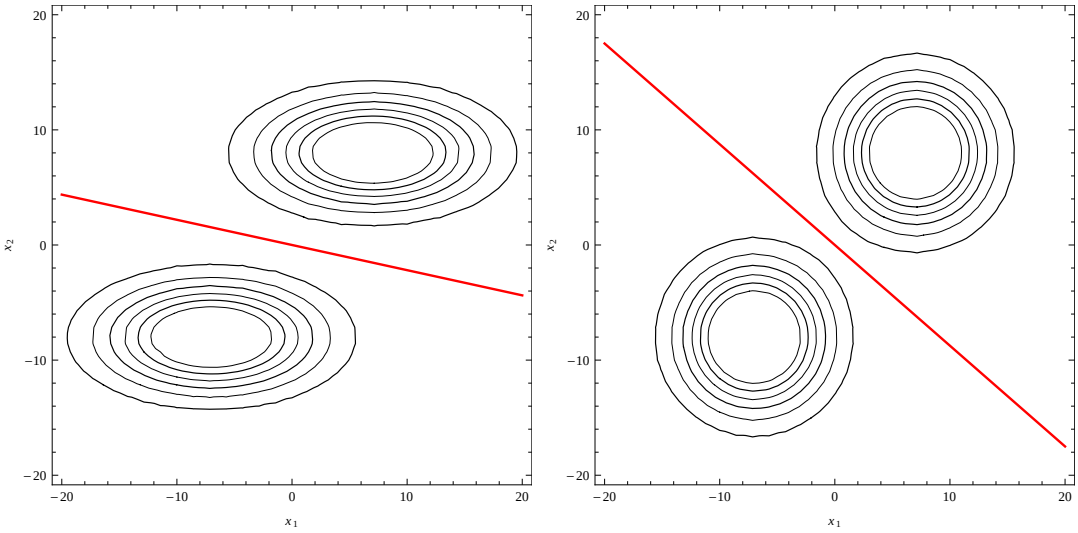
U tom slučaju kovarijacijska matrica $\boldsymbol{\Sigma}$ identična je za sve klase, pa za fiksirani $\mathbf{x}$ vrijedi

$$\underbrace{\frac{1}{2}\ln|\boldsymbol{\Sigma}|}_{=\text{konst.}} - \frac{1}{2}(\underbrace{\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}}_{=\text{konst.}} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j) + \ln P(\mathcal{C}_j)$$



(a) Različite kov. matrice

(b) Dijeljena kov. matrica



(c) Dijeljena dijagonalna kov. matrica

(d) Dijeljena izotropna kov. matrica

Slika 6: Granica između dviju klasa za bivarijatni Gaussov model.

pa model (27) degenerira u

$$h_j(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln P(\mathcal{C}_j)\,. \tag{29}$$

Primijetite da je iščeznuo član koji kvadratno ovisi o $\mathbf{x}$, pa su granice između klasa sada linearne (slika 6b). U slučaju da klase nisu linearno razdvojive, ovaj će model manje točno modelirati podatke. Model ima ukupno Kn parametara za srednje vrijednosti te $\frac{n}{2}(n+1)$ parametara za dijeljenu kovarijacijsku matricu. Broj parametara još uvijek kvadratno ovisi o n, što može biti problematično kod visokih dimenzija.

2. pojednostavljenje: dijagonalna kovarijacijska matrica

Daljnje pojednostavljenje modela moguće je uz pretpostavku da varijable nisu korelirane odnosno da su nezavisne.[5] U tom slučaju koristimo dijagonalnu kovarijacijsku matricu, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$. Matrica $\boldsymbol{\Sigma}^{-1}$ onda je također dijagonalna te vrijedi $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_i^2)$ i $|\boldsymbol{\Sigma}| = \prod_i \sigma_i$. Multivarijatna Gaussova gustoća (24) degenerira u produkt univarijatnih Gaussovih razdiobi:

$$\begin{aligned} p(\mathbf{x}|\mathcal{C}_j) &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^N \sigma_i} \exp\left\{ -\frac{1}{2}\sum_{i=1}^N \left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2 \right\} \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{1}{2}\left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2 \right\}. \end{aligned} \tag{30}$$

Model koji smo dobili jest **naivan Bayesov klasifikator** za kontinuirane varijable. (Primijetite da za $n = 1$ dobivamo (19)). Uvrštavanjem (30) u (29) i zanemarivanjem pribrojnika koji su za sve klase jednaki dobivamo

$$h_j(\mathbf{x}) = -\frac{1}{2}\sum_{i=1}^N \left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2 + \ln P(\mathcal{C}_j)\,. \tag{31}$$

Izraz $((x_i - \mu_{ij})/\sigma_i)^2$ odgovara udaljenosti između x_i i μ_{ij} izraženoj u jedinicama standardne devijacije. Geometrijski gledano, prvi pribrojnik u (31) jednak je kvadratu **normalizirane euklidske udaljenosti** između vrijednosti x_i i vrijednosti μ_{ij} (srednje vrijednosti dimenzije i za klasu $\mathcal{C}_j$). Normirana euklidska udaljenost neosjetljiva je na razlike u varijanci između pojedinih dimenzija. Varijance općenito mogu biti različite, pa klasama odgovaraju hiperelipsoidi, koji su međutim poravnati s osima jer su varijable nekorelirane (slika 6c). Broj parametara modela jest Kn za srednje vrijednosti klasa te n za varijance, što predstavlja smanjenje sa $\mathcal{O}(n^2)$ na $\mathcal{O}(n)$.

3. pojednostavljenje: izotropna kovarijacijska matrica

Moguća su daljnja pojednostavljenja modela. Možemo pretpostaviti da su varijance iste za svaku dimenziju, $\sigma_i = \sigma$, odnosno da je matrica kovarijacije izotropna, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

$$h_j(\mathbf{x}) = -\frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^N (x_i - \mu_{ij})^2}_{||\mathbf{x}-\boldsymbol{\mu}||^2} + \ln P(\mathcal{C}_j)\,. \tag{32}$$

[5] Nekoreliranost ne implicira nezavisnost, ali budući da u modelu ni na koji drugi način ne modeliramo zavisnost, ukidanje koreliranosti znači zapravo da pretpostavljamo nezavisnost.

Tablica 1: Složenost modela u ovisnosti o pretpostavkama o izglednostima klasa.

Pretpostavka	Kovarijacijska matrica	Broj parametara	
Različite, hiperelipsoidi	$\mathbf{\Sigma}_j$	$Kn(n+1)/2$	$\mathcal{O}(Kn^2)$
Dijeljena, hiperelipsoidi	$\mathbf{\Sigma}$	$n(n+1)/2$	$\mathcal{O}(n^2)$
Različite, poravnati hiperelipsoidi	$\mathbf{\Sigma}_j = \mathrm{diag}(\sigma_{i,j}^2)$	Kn	$\mathcal{O}(Kn)$
Dijeljena, poravnati hiperelipsoidi	$\mathbf{\Sigma} = \mathrm{diag}(\sigma_i^2)$	n	$\mathcal{O}(n)$
Različite, hipersfere	$\mathbf{\Sigma}_j = \sigma_j^2\mathbf{I}$	K	$\mathcal{O}(K)$
Dijeljena, hipersfere	$\mathbf{\Sigma} = \sigma^2\mathbf{I}$	1	$\mathcal{O}(1)$

Nomalizirana euklidska udaljenost svodi se na običnu euklidsku udaljenost, a klase su hipersfere sa središtima u $\boldsymbol{\mu}_j$ (slika 6d).

4. pojednostavljenje: jednake apriorne vjerojatnosti klasa

Konačno, model možemo pojednostaviti tako da pretpostavimo da su apriorne vjerojatnosti klasa jednake. Tada za model dobivamo

$$h_j(\mathbf{x}) = -||\mathbf{x} - \boldsymbol{\mu}_j||^2$$

gdje $||\cdot||$ označava normu vektora.[6] Model primjer $\mathbf{x}$ klasificira u klasu $\mathcal{C}_j$ s najbližom srednjom vrijednošću $\boldsymbol{\mu}_j$. Kao i ranije, izraz možemo raspisati kako bismo utvrdili kakav je oblik granice između klasa:

$$h_j(\mathbf{x}) = -||\mathbf{x} - \boldsymbol{\mu}_j||^2 = -(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_j) = -(\mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j)\,.$$

Prvi pribrojnik zajednički je svim klasama, pa ga možemo zanemariti. Granica između klasa dakle je linearna

$$h_j(\mathbf{x}) = \mathbf{w}_j^{\mathrm{T}}\mathbf{x} + w_{j0}$$

s težinama $\mathbf{w}_j = \boldsymbol{\mu}_j$ i slobodnim članom $w_{j0} = -\frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j = -\frac{1}{2}||\boldsymbol{\mu}_j||^2$.

Opisali smo niz Bayseovih modela temeljenih na Gaussovoj gustoći. Pregled modela dan je u tablici 1. Kod najsloženijeg modela koristili smo zasebnu kovarijacijsku matricu za svaku klasu. Prijelaz s nelinarnog modela na linearan ostvarili smo uporabom dijeljene kovarijacijske matrice. Daljnja pojednostavljenja ostvarili smo uvođenjem dodatnih pretpostavki o izglednostima klasa, i to uporabom dijagonalne ili izotropne kovarijacijske matrice. Moguće su i druge kombinacije, npr. korištenje dijagonalnih, ali nedijeljenih kovarijacijskim matrica. Uvođenje dodatnih pretpostavki uvijek dovodi do smanjenja broja parametara, što za posljedicu ima povećanje pristranosti i smanjenje varijance modela. Za odabir optimalnog modela može se koristiti npr. unakrsna provjera.

Primjer 12 (Bayesov višedimenzijski klasifikator s kontinuiranim ulazima) Želimo izgraditi Bayesov klasifikator za klasifikaciju primjera iz prostora $\mathbb{R}^3$ u klase $\mathcal{C}_1$ i $\mathcal{C}_2$. Metodom najveće izglednosti dobili smo sljedeće procjene parametara na skupu za

[6] *Normu* ili duljinu vektora $\mathbf{x}$ definiramo kao $||\mathbf{x}|| = \sqrt{(\mathbf{x}|\mathbf{x})} = \sqrt{x^{\mathrm{T}}x}$.

učenje:

$$\hat{\boldsymbol{\mu}}_1 = (1, -5, 0) \qquad \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1 & -0.75 & 0.2 \\ -0.75 & 6.25 & 1.5 \\ 0.2 & 1.5 & 4 \end{pmatrix} \qquad \hat{P}(\mathcal{C}_1) = 0.4$$

$$\hat{\boldsymbol{\mu}}_2 = (7, 2, 3) \qquad \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 4 & 0.7 & 2 \\ 0.7 & 12.25 & -3.5 \\ 2 & -3.5 & 16 \end{pmatrix} \qquad \hat{P}(\mathcal{C}_2) = 0.6$$

Razmotrimo četiri modela, od nasloženijeg do najjednostavnijeg. Najsloženiji model je onaj definiran izrazom (27). Taj model ukupno ima 19 parametara (12 za kovarijacijsku matricu, 6 za vektore srednjih vrijednosti i 1 za apriornu vjerojatnost jedne od klasa). Uvrštavanjem gornjih procjena u (27) i izračunom determinanti i inverza kovarijacijskih matrica, dobivamo konkretne izraze za hipoteze dviju klasa:

$$\begin{aligned} h_1(\mathbf{x}) = &-0.57x_1^2 - 0.10x_2^2 - 0.14x_3^2 - 0.17x_1x_2 + 0.12x_1x_3 + 0.08x_2x_3 \\ &+ 0.32x_1 - 0.83x_2 + 0.30x_3 - 4.70 \\ h_2(\mathbf{x}) = &-0.14x_1^2 - 0.04x_2^2 - 0.04x_3^2 + 0.03x_1x_2 + 0.04x_1x_3 - 0.02x_2x_3 \\ &+ 1.70x_1 + 0.06x_2 - 0.02x_3 - 9.90 \end{aligned}$$

Hipoteze kvadratno ovise o ulaznim varijablama. Granica između dviju klasa je krivulja za koju $h_1(\mathbf{x}) = h_2(\mathbf{x})$, koja će također biti kvadratna funkcija od $\mathbf{x}$. Jednostavniji model je model s dijeljenom kovarijacijskom matricom. Prema (28), dijeljena matrica jednaka je

$$\hat{\boldsymbol{\Sigma}} = 0.4\hat{\boldsymbol{\Sigma}}_1 + 0.6\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 2.8 & 0.12 & 1.28 \\ 0.12 & 9.85 & -1.5 \\ 1.28 & -1.5 & 11.2 \end{pmatrix}.$$

Ukupan broj parametara modela je 13. Uvrštavanjem u (29) dobivamo sljedeće hipoteze:

$$\begin{aligned} h_1(\mathbf{x}) &= 0.44x_1 - 0.53x_2 - 0.12x_3 - 2.50 \\ h_2(\mathbf{x}) &= 2.50x_1 + 0.17x_2 + 0.01x_3 - 9.40 \end{aligned}$$

Zbog dijeljenja kovarijacijske matrice iščezli su kvadratni članovi, pa su dobivene hipoteze linearne. Još jednostavniji model dobit ćemo ako uvedemo naivnu pretpostavku o nekoreliranosti varijabli, odnosno ako koristimo dijagonalnu dijeljenu kovarijacijsku matricu:

$$\hat{\boldsymbol{\Sigma}}_{diag} = \begin{pmatrix} 2.8 & 0 & 0 \\ 0 & 9.85 & 0 \\ 0 & 0 & 11.2 \end{pmatrix}.$$

Ovaj model ima 10 parametara. Uvrštavanjem u (31) dobivamo hipoteze:

$$\begin{aligned} h_1(\mathbf{x}) &= 0.36x_1 - 0.51x_2 - 2.4 \\ h_1(\mathbf{x}) &= 2.50x_1 + 0.20x_2 + 0.27x_3 - 9.9 \end{aligned}$$

Konačno, model možemo pojednostaviti korištenjem izotropne kovarijacijske matrice, odnosno uprosječivanjem varijanci svih varijabli:

$$\hat{\boldsymbol{\Sigma}}_{iso} = \begin{pmatrix} 7.88 & 0 & 0 \\ 0 & 7.88 & 0 \\ 0 & 0 & 7.88 \end{pmatrix}.$$

Ovaj model ima 8 parametara. Uvrštavanjem u (32) dobivamo hipoteze:

$$h_1(\mathbf{x}) = 0.13x_1 - 0.63x_2 - 2.6$$
$$h_2(\mathbf{x}) = 0.89x_1 + 0.25x_2 + 0.38x_3 - 4.4$$

Opisani modeli definiraju različite granice između dviju klasa, pa će neke primjere različito klasificirati. Na primjer, posljednji će model primjer $\mathbf{x} = (1, 3.5, 2)$ klasificirati u klasu $\mathcal{C}_2$, dok će ga složeniji modeli klasificirati u klasu $\mathcal{C}_1$.

Moguće je također izgraditi model koji djelomično kombinira više pretpostavki. Naime, kovarijacijsku matricu klase $\mathcal{C}_j$ možemo napisati kao težinsku kombinaciju triju osnovnih slučajeva:

$$\mathbf{\Sigma}'_j = \alpha\sigma^2\mathbf{I} + \beta\mathbf{\Sigma} + (1 - \alpha - \beta)\mathbf{\Sigma}_j, \qquad \alpha \geqslant 0,\ \beta \geqslant 0,\ 0 \leqslant \alpha + \beta \leqslant 1\,.$$

Za $\alpha = \beta = 0$ dobivamo najopćenitiji slučaj nelinearnog klasifikatora. Za $\alpha = 0$ i $\beta = 1$ kovarijacijska je matrica dijeljena, dok je za $\alpha = 1$ i $\beta = 0$ kovarijacijska matrica dijagonalna; u oba je slučaja klasifikator linearan. Između ovih krajnosti nalazi se čitav niz hibridnih modela. Optimalne vrijednosti hiperparametara α i β treba naravno odrediti unakrsnom provjerom.