# 15. Bayesov klasifikator

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, predavanja, v3.1

Prošli puta pričali smo o statističkoj procjeni parametara, što je ključni mehanizam učenja probabilističkih modela. Prisjetili smo se osnovnih vjerojatnosnih distribucija, koje se koriste u strojnom učenju te smo zatim pričali o **procjeniteljima**. Objasnili smo procjenitelj **najveće izglednosti (MLE)**, a zatim smo razmotrili procjenitelj **maksimum aposteriori (MAP)**, koji nam omogućava da u procjenu ugradimo naše apriorno znanje o parametrima.

Danas ćemo razmotriti najjednostavniji probabilistički model – **Bayesov klasifikator**. Prvo ćemo pričati o Bayesovom klasifikatoru za kontinuirane značajke, odnosno **Gaussovom Bayesovom klasifikatoru**. U idućem predavanju pogledat ćemo onda Bayesov klasifikator za diskretne značajke.

Bayesov klasifikator ujedno će biti naš prvi **generativni model** – svi modeli koje smo razmatrali do sada bili su diskriminativni. Reći ćemo nešto više o toj podjeli na generativne i diskriminativne modele.

## 1 Pravila vjerojatnosti

Prije nego što zaronimo u Bayesov klasifikator, prisjetit ćemo se nekih jednostavnih ideja iz teorije vjerojatnosti, a koje se sveprisutne kod probabilističkih modela. Cijela algebra teorije vjerojatnosti svodi se na dva jednostavna pravila: pravilo zbroja i pravilo umnoška. **Pravilo zbroja** je:

$$P(x) = \sum_y P(x, y)$$

Vjerojatnost $P(x, y)$ je vjerojatnost zajedničke realizacije $x$ i $y$. U strojnom učenju, to je vjerojatnost da primjer $\mathbf{x}$ ima oznaku $y$. Tu vjerojatnost zovemo **zajednička (združena) vjerojatnost** (engl. *joint probability*). Pravilo zbroja nam govori da iz zajedničke vjerojatnosti možemo dobiti vjerojatnost pojedinačnih varijabli ili podskupa varijabli. Ta vjerojatnost se onda zove **marginalna vjerojatnost**. Zove se marginalna jer se dobiva **marginalizacijom** – izračunom vjerojatnosti podskupa varijabli. Marginalizacija, naravno, vrijedi i kada imamo više varijabli. Tada možemo marginalizirati po, npr., samo jednoj varijabli:

$$P(x, y) = \sum_z P(x, y, z)$$

ili možemo marginalizirati po više varijabli:

$$P(x) = \sum_y \sum_z P(x, y, z)$$

Drugo pravilo teorije vjerojatnosti je **pravilo umnoška**. Najprije, prisjetimo se da je **uvjetna vjerojatnost** definirana kao:

$$P(y|x) = \frac{P(x, y)}{P(x)} \qquad \text{ili} \qquad P(x|y) = \frac{P(x, y)}{P(y)}$$

To je vjerojatnost vrijednosti $y$, ako znamo da je ostvarena vrijednost $x$ (ili obrnuto, za drugu formulu). Pravilo umnoška samo je drugi pogled na definiciju uvjetne vjerojatnosti:

$$P(x, y) = P(y|x)P(x) = P(x|y)P(y)$$

Ovo su, dakle, ta dva osnovna pravila. Oba pravila vrijede i ako vjerojatnost $P$ zamijenimo s gustoćom vjerojatnosti $p$. Sve što ćemo mi raditi u nastavku svodit će se na pametnu primjenu ova dva pravila. Zapravo, odmah ćemo upotrijebiti ova dva pravila kako bismo izveli osnovno pravilo koje nam treba za Bayesov klasifikator, a to je **Bayesovo pravilo**:

$$P(x, y) = P(y|x)P(x) = P(x|y)P(y) \quad /P(x)$$
$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Snaga bayesovog pravila leži u tome što nam omogućava da obrnemo smjer zaključivanja: iz vjerojatnosti $P(x|y)$ možemo zaključiti o vjerojatnosti $P(y|x)$. U logici bi se to zvalo **abduktivno zaključivanje**. U klasičnoj logici abduktivno zaključivanje je "persona non grata", zato jer je to pravilo evidentno neispravno (zaključak nije logička posljedica premisa), no u strojnom učenju abdukcija je dragi gost koji nosi lijepe darove (mogućnost zaključivanja od podataka prema klasi, tj. od posljedice prema uzroku).

Sada kada znamo osnovna dva pravila vjerojatnosti (pravilo zbroja i pravilo umnoška) te iz njih izvedeno Bayesovo pravilo, spremni smo za Bayesov klasifikator.

## 2 Bayesov klasifikator

**Bayesov klasifikator** izravno primijenjuje Bayesov teorem kako bi izračunao vjerojatnost oznake $y$ za zadani ulazni primjer $\mathbf{x}$:

$$P(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)P(y)}{p(\mathbf{x})}$$

Primijetite da se ovdje pojavljuju **vjerojatnosti** (veliko "$P$") i **gustoće vjerojatnosti** (malo "$p$"), ovisno o tome je li varijabla diskretna ili kontinuirana. Konkretno, varijabla $y$ je diskretna oznaka klase, dok je vektor primjera $\mathbf{x}$ općenito vektor značajki, diskretnih ili kontinuiranih, pa pišemo gustoću vjerojatnosti jer je to općenitiji slučaj. To što množimo gustoću vjerojatnosti $p$ s vjerojatnošću $P$ nije nikakav problem (rezultat će biti gustoća vjerojatnosti).

Vjerojatnost $P(y|\mathbf{x})$ nazivamo **aposteriorna vjerojatnost oznake** (engl. *posterior*): to je vjerojatnost oznake $y$ za zadani primjer $\mathbf{x}$. I to je ono što nas zapravo zanima: koliko je vjerojatno da primjer $\mathbf{x}$ pripada klasi $y$. Gustoću vjerojatnosti $p(\mathbf{x}|y)$ nazivamo **izglednost klase** (engl. *class likelihood*). To je gustoća vjerojatnosti primjera uz zadanu klasu, odnosno [1] distribucija primjera unutar dotične klase. Intuitivno, izglednost klase nam govori kolika je vjerojatnost primjera za zadanu klasu ("ako gledam samo primjere iz klase $y$, koliko je vjerojatno da vidim $\mathbf{x}$"). Primijetite da je to točno obrnuto od aposteriorne vjerojatnosti, koja nam govori koja je vjerojatnost klase za zadani primjer. Vjerojatnost $P(y)$ naziva se **apriorna vjerojatnost klase** (engl. *class prior*). To je vjerojatnost klase neovisno o primjerima. Npr., ako radimo klasifikaciju treba li klijentu banke odobriti kredit, i ako je situacija takva da se većina (npr. 80%) kredita odobrava (a vjerojatno jest jer inače ne bi bili tu gdje jesmo), onda je $P(y = 1) = 0.8$. Konačno, gustoća vjerojatnosti $p(\mathbf{x})$ je gustoća vjerojatnosti primjera neovisno o klasi ("koliko je vjerojatno da vidim $\mathbf{x}$ iz bilo koje klase").

Možda ste se pitali – ili, ako niste, možda da se sada pitate – zašto uopće rastavljamo zajedničku gustoću vjerojatnosti $p(\mathbf{x}, y)$ na izglednost $p(\mathbf{x}|y)$ i apriornu vjerojatnost $P(y)$? Naime, matematički gledano, očito je to opet jednako zajedničkoj vjerojatnosti, pa ispada da zapravo ništa nismo napravili. No, problem je u tome što je zajednička vjerojatnost (odnosno [2]

zajednička gutoća vjerojatnosti) općenito može biti vrlo složena distribucija. Npr., $\mathbf{x}$ može biti vektor kontinuiranih vrijednosti, dok je $y$ diskretna varijabla (npr. klasifikacija učenika po uspjehu u srednjoj školi). Takve složene distribucije ne možemo modelirati nekom standardnom teorijskom distribucijom. Međutim, ako zajedničku distribuciju rastavimo, dobivamo dvije jednostavnije distribucije, koje možemo mnogo jednostavnije modelirati, koristeći standardne distribucije. Štoviše, gustoću $p(\mathbf{x}|y)$ možemo modelirati *zasebno* za svaku pojedinu klasu $y$, što vrlo pojednostavljuje stvari. Rastavljanje neke složene distribucije na umnožak jednostavnijih distribucija naziva se **faktorizacija**.

Jedini preostali problem je $p(\mathbf{x})$ u nazivniku Bayesovog pravila. To je gustoća vjerojatnosti primjera $\mathbf{x}$, neovisno o klasi. Zašto nam je ona problematična? Zato jer je i ta distribucija također općenito vrlo složena. Zato ćemo i nju faktorizirati, i to ovako:

$$p(\mathbf{x}) = \sum_y p(\mathbf{x}, y) = \sum_y p(\mathbf{x}|y)P(y)$$

Dakle, faktorizaciju smo napravili tako da smo najprije primijenili pravilo zbroja (po svim mogućim oznakama klase), a zatim pravilo umnoška. Uvjerite se da je to doista jednako $p(\mathbf{x})$. Ovako modelirati $p(\mathbf{x})$ jednostavnije je iz istog razlog koji smo spomenuli gore: modeliramo zasebno izglednost klase i zasebno apriornu vjerojatnost klase, i te dvije distribucije možemo modelirati jednostavnim teorijskim distribucijama. To što je $p(\mathbf{x})$ onda složena distribucija koja se dobiva po gornjem izrazu nas više ne zanima.

Nakon svih ovih intervencija, dolazimo napokon do **modela Bayesovog klasifikatora**:

$$h_j(\mathbf{x}; \boldsymbol{\theta}) = P(y = j|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_{y'} p(\mathbf{x}|y')P(y')}$$

Ovako definirana hipoteza daje nam vjerojatnost klasifikacije u klasu $j$. Primijetite da Bayesov klasifikator bez problema može raditi s više klasa (formalno, imamo po jednu hipotezu $h_j$ za svaku od $K$ klasa).

Ako nas, međutim, ne zanimaju vjerojatnosti klasifikacije u pojedine klase, nego samo želimo odrediti oznaku primjera, onda ćemo ga klasificirati u klasu čija je vjerojatnost najveća. To je tzv. **maksimum aposteriori hipoteza (MAP)**. Model tada definiramo kao: 3

$$h(\mathbf{x}; \boldsymbol{\theta}) = \operatorname*{argmax}_y \; p(\mathbf{x}|y)P(y)$$

▸ **PRIMJER**

Radimo klasifikaciju u $K = 3$ klase. Recimo da smo naučili model (tj. procijenili parametre modela). Apriorne vjerojatnosti klasa su $P(y = 1) = P(y = 2) = 0.3$, $P(y = 3) = 0.4$. Za neki konkretan primjer $\mathbf{x}$, izglednosti klasa su $p(\mathbf{x}|y = 1) = 0.9$, $p(\mathbf{x}|y = 2) = p(\mathbf{x}|y = 3) = 0.4$. U koju klasu klasificiramo primjer $\mathbf{x}$?

$$p(x_1|y = 1)P(y = 1) = 0.9 \cdot 0.3 = 0.27$$
$$p(x_1|y = 2)P(y = 2) = 0.4 \cdot 0.3 = 0.12$$
$$p(x_1|y = 3)P(y = 3) = 0.4 \cdot 0.4 = 0.16$$
$$p(x) = \sum_{j=1}^{3} p(x|y = j)P(y = j) = 0.55$$
$$P(y = 1|x) = 0.27/0.55 = 0.49 \quad \Leftarrow \text{MAP (primjer klasificiramo u ovu klasu)}$$
$$P(y = 2|x) = 0.12/0.55 = 0.22$$
$$P(y = 3|x) = 0.16/0.55 = 0.29$$

Vektor $\boldsymbol{\theta}$ je vektor parametara apriorne distribucije i izglednosti klase. Koji su to točno parametri i koliko ih ukupno ima ovisi o tome koje smo distribucije odabrali. Što se apriorne vjerojatnosti klase tiče, ako je klasifikacija binarna ($K = 2$), koristit ćemo **Bernoullijevu distribuciju**, dok ćemo za višeklasnu ($K > 2$) klasifikaciju koristiti **kategoričku (multinulijevu) distribuciju**. Što se izglednosti klase tiče, ako su značajke diskretne, koristit ćemo **Bernoullijevu ili kategoričku distribuciju**, ovisno o tome je li značajka binarna ili viševrijednosna. Za kontinuirane značajke koristit ćemo **Gaussovu (normalnu) distribuciju**. Zapravo, budući da ćemo uvijek imati više od jedne značajke, koristit ćemo **multivarijatnu Gaussovu distribuciju**.

Bayesov klasifikator je **parametarski model**. Što to znači? Općenito, to znači da broj parametara modela ne ovisi o broju primjera. No, budući da se ovdje konkretno radi o probabilističkom modelu, to također znači da model pretpostavlja da se primjeri $\mathbf{x}$ i oznake $y$ pokoravaju nekoj teorijskoj vjerojatnosnoj distribuciji. Naravno, broj parametara tih distribucija ne ovisi o broju primjera, pa zato ni broj parametara cjelokupnog modela ne ovisi o broju primjera.

Ovime smo definirali model Bayesovog klasifikatora. Što je s druge dvije komponente algoritma strojnog učenja: funkcijom pogreške i optimizacijskim postupkom? Drugim riječima, kako ćemo trenirati Bayesov klasifikator? Prisjetimo se da kod probabilističkih modela trenirati model znači **procijeniti parametre**. Za procjenu parametara upotrijebit ćemo ono što smo usvojili prošli tjedan: MLE procjenitelje ili MAP procjenitelje. Dakle, optimizacijski postupak bit će maksimizacija izglednosti parametara (MLE) ili maksimizacija aposteriorne vjerojatnosti parametara (MAP). Sukladno tome, funkcija pogreške bit će jednostavno negativna izglednost parametara (MLE) ili negativna aposteriorna vjerojatnost parametara (MAP).

Bayesov klasifikator naš je prvi **generativni model**. Sada je dobar trenutak da kažemo nešto više o toj familiji modela.

# 3 Generativni modeli

Generativni modeli modeliraju **zajedničku vjerojatnost** (odnosno zajedničku gustoću vjerojatnosti) $p(\mathbf{x}, y)$. Kao što smo vidjeli, na temelju te vjerojatnosti može se vrlo jednostavno, primjenom Bayesovog pravila, izračunati aposteriorna vjerojatnost $P(y|\mathbf{x})$, tj. vjerojatnost da primjer $\mathbf{x}$ pripada klasi $y$.

Ovakav pristup, koji modelira zajedničku distribuciju primjera $\mathbf{x}$ i klasa $y$, nazivamo generativnim jer modelira postupak **generiranja (nastanka) podataka**. Da bismo razumijeli što pod time mislimo, pogledajmo brojnik Bayesovog klasifikatora:

$$P(\mathbf{x}, y) = p(\mathbf{x}|y)P(y)$$

Želimo li opisati način nastanka označenih primjera $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_i$, koji se ravnaju po zajedničkoj distribuciji $p(\mathbf{x}, y)$, možemo reći da primjeri nastaju **stohastičkim procesom** koji se sastoji od dva koraka: u prvome koraku odabrana je oznaka $y$ po distribuciji $P(y)$, a u drugome je koraku za odabrani $y$ odabran primjer $\mathbf{x}$ po distribuciji $p(\mathbf{x}|y)$. Takva priča, koja objašnjava stohastički postupak generiranja podataka, naziva se **generativna priča** (engl. *generative story*).

Kod Bayesovog klasifikatora generativna je priča dosta kratka i pomalo dosadna, jer se sastoji od svega dva koraka, ali općenito, kod složenijih generativnih modela, priča može biti dulja i uzbudljivija. Takvi složeniji generativni modeli su, npr., **Bayesove mreže**, **mješavina Gaussovih** distribucija (engl. *Gaussian mixture model, GMM*), **skriven Markovljev model** (engl. *Hidden Markov model, HMM*), i **latentna Dirichletova alokacija** (engl. *latent Dirichlet allocation, LDA*). Prva dva modela ćemo pogledati u narednim tjedima.

Generativna priča također se može upotrijebiti za generiranje sintetičkih podataka. Jednom kada je model naučen, možemo slijediti generativnu priču kako bismo uzorkovali primjere iz zajedničke distribucije. Kod Bayesovog klasifikatora, za svaki primjer prvo bismo uzorkovali klasu $y$ prema distribuciji $P(y)$, a onda bismo uzorkovali primjer iz distribucije $P(\mathbf{x}|y)$ koja odgovara izglednosti za dotičnu klasu $y$.

## 3.1 Generativno vs. diskriminativno

Usporedimo sada generativne modele sa **diskriminativnim modelima**. Svi modeli s kojima smo se do sada bavili (linearna regresija, logistička regresija, SVM, k–NN) bili su diskriminativni, a ne generativni. U čemu je razlika?

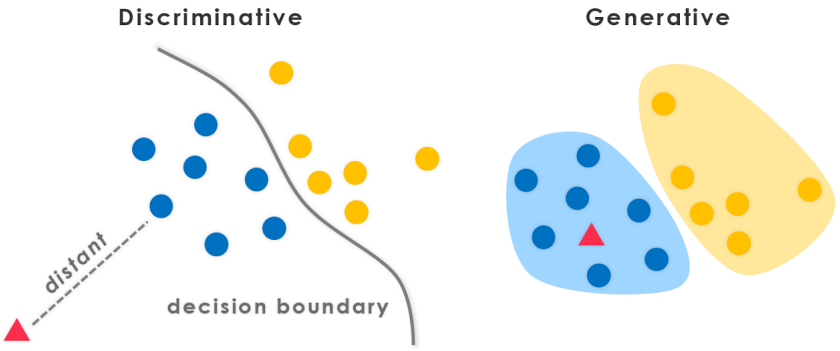
Diskriminativni modeli izravno modeliraju aposteriornu vjerojatnost $P(y|\mathbf{x})$. Prisjetimo se logističke regresije. Model je tamo bio definiran ovako: [4]

$$h(\mathbf{x}; \mathbf{w}) = P(y|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Primijetite da tu niti u jednom trenutku nismo modelirali zajedničku vjerojatnost, nego smo direktno modelirali aposteriornu vjerojatnost. To je drugačije od generativnih modela, kod kojih aposteriornu vjerojatnost modeliramo indirektno, preko zajedničke vjerojatnosti.

Logistička regresija je probabilistički diskriminativni model – modelira vjerojatnost oznake za neki primjer. Međutim, u porodicu diskriminativnih modela pripadaju i oni modeli koji uopće ne modeliraju vjerojatnost, npr. SVM. Zapravo, većina diskriminativnih modela nisu probabilistički (npr., SVM, perceptron, višeslojni perceptron, stabla odluke, k-NN). Ti modeli doduše ne modeliraju aposteriornu vjerojatnost, ali, slično kao i logistička regresija, direktno modeliraju granicu između klasa (ni u kojem trenutku SVM ne modelira zajedničku vjerojatnost primjera i oznaka).

Razliku između generativnih i diskriminativnih modela ilustrira sljedeća slika:




Prikazan je problem binarne klasifikacije (plavi vs. žuti primjeri). Na lijevoj je slici prikazan rezultat dobiven nekim diskriminativnim modelom. Sve što ovdje dobivamo je granica između klasa (decizijska granica), koja je općenito nekakva hiperpovršina u ulaznom prostoru za koju $h(\mathbf{x}) = 0.5$ (kod logističke regresije) ili $h(\mathbf{x}) = 0$ (kod SVM-a). Koncept pouzdanosti klasifikacije povezan je s konceptom udaljenosti primjera od granice. Na desnoj slici prikazan je rezultat na istom skupu primjera dobiven nekim generativnim modelom. Ovdje modeliramo vjerojatnosne distribucije primjera unutar svake od dviju klasa $p(\mathbf{x}|y)p(y)$, što, nakon množenja s apriornom vjerojatnošću klasa $P(y)$, daje distribuciju $p(\mathbf{x}, y)$, tj. zajedničku vjerojatnost (odnosno gustoću vjerojatnosti) primjera i oznaka. Kada želimo raditi klasifikaciju primjera, iz te zajedničke vjerojatnosti možemo izračunati aposteriornu vjerojatnost $p(y|\mathbf{x})$, i klasificirati primjere u klasu za koju je ta vjerojatnost najveća (MAP hipoteza). Time je onda implicitno definirana i granica između dviju klasa kao $h(\mathbf{x}) = p(y|\mathbf{x}) = 0.5$.

## 3.2 Prednosti i nedostatci

Ovdje se sad, naravno, postavlja neizbježno pitanje: jesu li bolji generativni ili diskriminativni modeli? Odgovor je, kako to često biva: ovisi. Prisjetite se **teorema o nebesplatnom ručku** (engl. *no free lunch theorem*): niti jedan algoritam nije univerzalno najbolji. U nekim situacijama generativni modeli mogu biti bolji, a u nekim diskriminativni. [5]

Da bismo znali ocijeniti koji je model bolji za koji problem, trebamo znati njihove prednosti i nedostatke. Krenimo od prednosti generativnih modela:

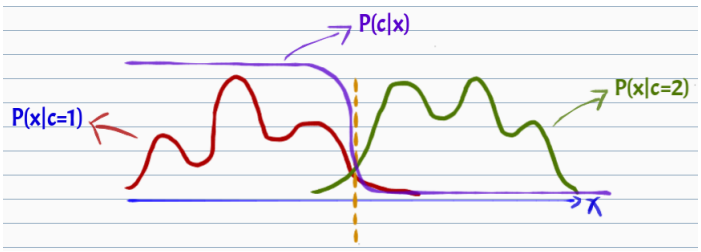
- U generativne modele relativno je lako ugraditi **pozadinsko/stručno znanje o problemu**. To znanje je u obliku apriornih distribucija (apriorna distribucija klasa, apriorna distribucija parametara) ili čak specifičnih struktura modela. Takvo znanje može se onda elegantno kombinirati sa znanjem dobivenim na temelju podataka – naznaku te ideje već smo vidjeli kod procjenitelja MAP;
- Druga prednost je **interpretabilnost rezultata i mogućnost različitih analiza** – generativni modeli nude vrlo intuitivnu interpretaciju podataka temeljenu na teoriji vjerojatnosti. Također, budući da modeliramo zajedničku distribuciju, možemo raditi različite analize ovisnosti između varijabli (tj. između značajki), a također možemo predviđati vrijednosti značajki na temelju drugih značajki, itd. Dakle, nismo ograničeni samo na izračun aposteriorne vjerojatnosti $p(y|\mathbf{x})$, nego možemo izračunati razne vjerojatnosti koje nas zanimaju. Naime, zajednička distribucija sadrži sve informacije – sve ostale vjerojatnosti (uvjetne, marginalne) mogu se iz nje izvesti i sadrže manju informaciju.

Druge (manje važne) prednosti generativnih modela su:

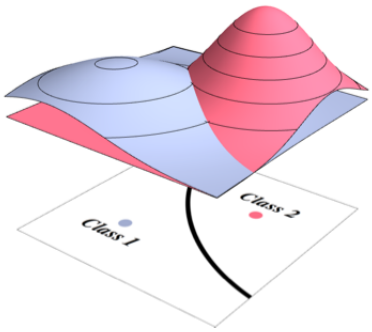
- **Ocjena pouzdanosti klasifikacije** – izlaz klasifikatora može se tumačiti kao vjerojatnost ili pouzdanost da primjer $\mathbf{x}$ pripada klasi $y$ – ovo je doduše prednost svih modela koji imaju probabilistički izlaz, a koji ne moraju nužno biti generativni, npr. logistička regresija;
- **Odbijanje klasifikacije** – ako je za neki primjer $\mathbf{x}$ izlaz klasifikatora manji od unaprijed zadanog praga, klasifikator može odbiti klasificirati primjer $\mathbf{x}$ i tako smanjiti broj pogrešnih klasifikacija. Primjeri koje klasifikator odbije klasificirati mogu se proslijediti na ručnu klasifikaciju. Ovo je također prednost koja nije ekskluzivna za generativne modele;
- Nalaženje **stršećih vrijednosti** (engl. *outliers*) – marginalizacijom vjerojatnosti $P(\mathbf{x}, y)$ možemo odrediti vjerojatnost primjera $P(\mathbf{x})$ i tako detektirati vrijednosti koje odskaču.

Naravno, generativni modeli u odnosu na diskriminativne modele imaju i neke nedostatke. Glavni nedostatci su:

- **Potreba za velikim brojem primjera** – modeliranje zajedničke vjerojatnosti $P(\mathbf{x}, y)$ iziskuje velik broj primjera, a da bi procjena parametara bila pouzdana. To je pogotovo problem kada je ulazni prostor visoke dimenzije, dakle kada vektor $\mathbf{x}$ sadrži mnogo značajki.
- **Nepotrebna složenost modeliranja** – ako je naš konačni cilj klasifikacija, onda je zapravo nepotrebno modelirati zajedničku distribuciju $P(\mathbf{x}, y)$, koja može biti vrlo složena i za čiju procjenu treba mnogo primjera. U tom slučaju dovoljno je izravno modelirati samo aposteriornu vjerojatnost $P(y|\mathbf{x})$, kao što to čine diskriminativni modeli. Ovaj nedostatak lijepo ilustrira sljedeća slika:



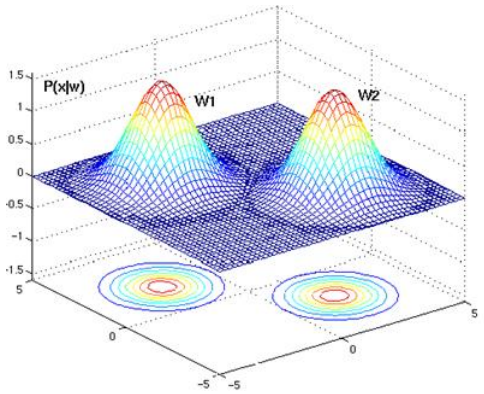

Ako je sve što nas zanima samo granica između klasa, onda je logistička regresija (sigmoida na slici) puno jefiniji model (u smislu složenosti modeliranja, tj. broja parametara) od generativnog modela koji je ovdje poprilično složen (složene višemodalne izglednosti klasa). Isto vrijedi i u ulaznome prostoru s više značajki, npr. dvije značajke:




Budući da složenost modela korelira s brojem parametara (složeniji modeli imaju više parametara), to će generativni model u pravilu imati više parametara nego diskriminativni model koji otprilike radi isti posao u smislu klasifikacije. (Vidjet ćemo primjer toga na idućem predavanju.) Vjerojatno upravo zbog te veće složenosti, u praksi se pokazuje da diskriminativni modeli – ako nas zanima samo točnosti klasifikacije – rade bolje od generativnih. Ipak, imajte na umu da to ne treba uvijek biti slučaj (nebesplatan ručak). Također, imajte na umu sve prednosti generativnih modela: nekad nam treba više od same klasifikacije – nekad je ključno da u model ugradimo pozadinsko znanje, ili želimo dubinski analizirati značajke i tumačiti rezultate. Onda su generativni modeli bolji izbor od diskriminativnih. 6

# 4 Gaussov Bayesov klasifikator

Pogledajmo sada napokon kako definirati Bayesov klasifikator, i to Bayesov klasifikator za kontinuirane značajke, tzv. **Gaussov Bayesov klasifikator**. Kod tog modela primjer $\mathbf{x}$ predstavljen je kao vektor brojeva, tj. značajke su numeričke, što znači da ćemo izglednosti klasa $P(\mathbf{x}|y)$ modelirati **Gaussovom distribucijom**. Npr., za dvije značajke (dvodimenzijski ulazni prostor) to izgleda ovako: 7




Ideja je da izglednost klase svake modeliramo kao jednu zasebnu multivarijatnu Gaussovu distribuciju. Srednja vrijednost te distribucije, dakle vektor $\boldsymbol{\mu}$, predstavlja **prototipni primjer** te klase, i taj primjer ima najveću gustoću vjerojatnosti. Primjeri koji su udaljeni od središta su manje vjerojatni da pripadaju toj klasi. Idealno, svi primjeri koji pripadaju ovoj klasi bi bili jednaki prototipnom primjeru, međutim zbog **šuma** to nije slučaj. Dakle, Gaussova distribucija ovdje modelira odstupanje od ideala uslijed šuma. 8

## 4.1 Univarijatni Gaussov Bayesov klasifikator

Jednostavnosti radi, za početak ćemo pogledati (nerealan) slučaj kada je prostor primjera jednodimenzijski, tj. kada imamo samo jednu značajku. U tom slučaju izglednost klase $P(x|y)$

modeliramo univarijatnom Gaussovom distribucijom:

$$x|y \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

Prisjetimo se Gaussove distribucije:

$$p(x|y=j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\Big\{ -\frac{(x-\mu_j)^2}{2\sigma_j^2} \Big\}$$

Model je:

$$h(x) = \underset{y}{\operatorname{argmax}}\, p(x,y) = \underset{y}{\operatorname{argmax}}\, p(x|y)P(y)$$

Ako nas zanima samo **pouzdanost** za svaku klasu $j$, a ne vjerojatnost, možemo definirati model čije je izlaz jednak zajedničkoj vjerojatnosti:

$$h_j(x) = p(x, y=j) = p(x|y=j)P(y=j)$$

Radi matematičke jednostavnosti, prelazimo u logaritamsku domenu:

$$\begin{aligned} h_j(x) &= \ln p(x|y=j) + \ln P(y=j) \\ &= -\frac{1}{2}\cancel{\ln 2\pi} - \ln \sigma_j - \frac{(x-\mu_j)^2}{2\sigma_j^2} + \ln P(y=j) \end{aligned}$$

Uklanjanje konstante (ne utječe na maksimizaciju):

$$h_j(x|\boldsymbol{\theta}_j) = -\ln \sigma_j - \frac{(x-\mu_j)^2}{2\sigma_j^2} + \ln P(y=j)$$

gdje je vektor parametara jednak

$$\boldsymbol{\theta}_j = (\mu_j, \sigma_j, \mu_j')$$

pri čemu smo sa $\mu_j'$ označili parametar distribucije $P(y=1)$, koji je jednak apriornoj vjerojatnosti klase $y=1$. Ovdje se radi o poznatim distribucijama (Gaussova i Bernoullijeve). Njihove parametre možemo procijeniti pomoću procjenitelja MLE (kojeg smo izveli prošli put):

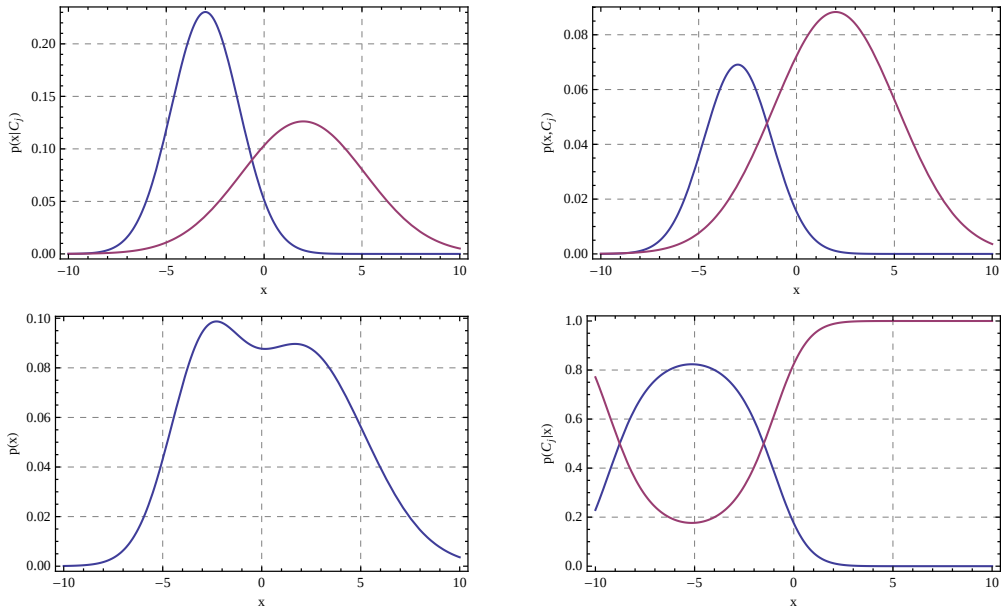
$$\begin{aligned} \hat{\mu}_j &= \frac{1}{N_j} \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = j\} x^{(i)} \\ \hat{\sigma}_j^2 &= \frac{1}{N_j} \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = j\} (x^{(i)} - \hat{\mu}_j)^2 \\ P(y=j) = \hat{\mu}_j' &= \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N} \end{aligned}$$

▸ **PRIMJER**

Klasificiramo u dvije klase, $y=1$ i $y=0$. Izglednosti tih klasa definirane su sljedećim Gaussovovim gustoćama vjerojatnosti:

$$\begin{aligned} p(x|y=1) &= \mathcal{N}(-3, 3) \\ p(x|y=0) &= \mathcal{N}(2, 10) \end{aligned}$$

Apriorne vjerojatnosti klasa neka su $P(y=1) = 0.3$ i $P(y=0) = 0.7$. Gustoće vjerojatnosti onda izgledaju ovako:




Krenimo od gornje lijeve slike. Ona prikazuje izglednosti za dvije klase $y = 1$ i $y = 0$, tj. gustoće vjerojatnosti $p(x|y = 1)$ (plava krivulja) i $p(x|y = 0)$ (ljubičasta krivulja). Budući da gustoća vjerojatnosti za $y = 1$ ima manju varijancu od gustoće vjerojatnosti za $y = 0$, to je ova prva "uža i viša". Na gornjoj desnoj slici prikazane su zajedničke gustoće vjerojatnosti $p(x, y = 1)$ i $p(x, y = 0)$ (to je funkcija dviju varijabli, ali ju prikazujemo odvojeno, kao dvije krivulje, za $y =$ i $y = 0$). Te se gustoće vjerojatnosti dobivaju iz $p(x|y)$ množenjem sa $P(y)$, pa je dakle razlika između ove slike i one prethodne u tome što su gustoće vjerojatnosti pomnožene s vrijednošću apriorne vjerojatnosti za $y = 1$ odnosno $y = 0$. Budući da je $P(y = 0) > P(y = 1)$, to je ljubičasta krivulja sada puno viša od plave. Na donjoj lijevoj slici prikazana je marginalna gustoća vjerojatnosti $p(x)$, koja se dobiva marginalizacijom zajedničke vjerojatnosti po oznaci $y$, tj. $p(x) = p(x, y = 0) + p(x, y = 1)$, što odgovara zbrajanju plave i ljubičaste krivulje. Konačno, na donjoj desnoj slici prikazane su aposteriorne gustoće vjerojatnosti $p(y = 1|x)$ i $p(y = 0|x)$. One su, prema Bayesovom teoremu, dobivene dijeljenjem zajedničke gustoće vjerojatnosti $p(x, y)$ (slika gore desno) s gustoćom vjerojatnosti primjera $p(x)$ (donja lijeva slika). Primijetite da je zbroj $p(y = 0|x) + p(y|x)$ nužno jednak jedinici. Npr., za primjer $x = -5$ vjerojatnost klasifikacije u klasu $y = 1$ iznosi $p(y = 1|x = 0.5) \approx 0.8$, a vjerojatnost klasifikacije u klasu $y = 0$ iznosi $p(y = 0|x = 0.5) \approx 0.2$). Također primijetite da je granica između klasa mjesto u ulaznom prostoru (ovdje jednodimenzijskom) gdje $p(y = 1|x) = p(y = 0|x) = 0.5$. To je u ovom slučaju na dva mjesta (jer izglednost za drugu klasu ima veću varijancu od izglednosti za prvu klasu, pa se Gaussove krivulje sijeku na dva mjesta; i slici gore lijevo to se ne vidi najbolje).

## 4.2 Multivarijatni Gaussov Bayesov klasifikator

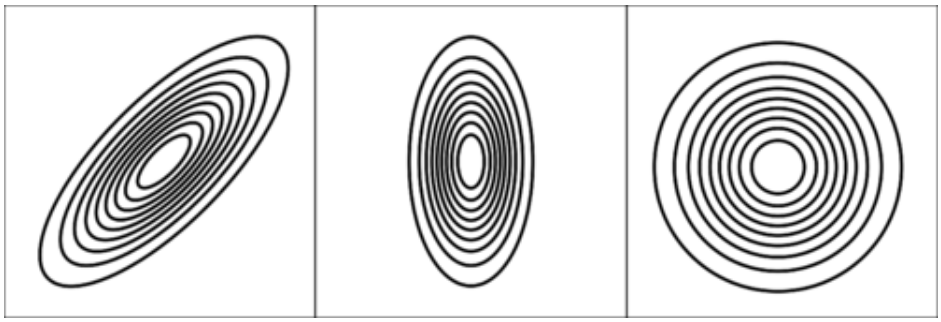
Fokusirajmo se sada na realističan slučaj: primjer $\mathbf{x}$ je vektor realnih brojeva. Tada izglednost svake klase $j$ modeliramo **multivarijatnom Gaussovom distribucijom**:

$$p(\mathbf{x}|y = j) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)\Big)$$

gdje je $\boldsymbol{\Sigma}_j$ matrica kovarijacije, a $\boldsymbol{\mu}_j$ je vektor srednjih vrijednosti za klasu $j$. Nakon logaritmiranja, model za svaku klasu $j$ pojedinačno je: 9

$$\begin{aligned} h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\ &= -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) + \ln P(y=j) \\ &\Rightarrow -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) + \ln P(y=j) \end{aligned}$$

Interpretacija za $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ analogna je kao i za jednodimenzijski slučaj: $\boldsymbol{\mu}_j$ je prototipna vrijednost primjera u klasi $j$, dok je $\boldsymbol{\Sigma}_j$ količina šuma i korelacija između izvora šuma unutar klase $j$. Ovisno o tome kako definiramo kovarijacijsku matricu, možemo na različite načine modelirati korelaciju između izvora šuma, npr. (slika koju smo već bili diskutirali prošli put):



$\boldsymbol{\Sigma}$ $\qquad$ $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ $\qquad$ $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

Koliko ovaj model ukupno ima parametara? Odgovor je: $\frac{n}{2}(n+1)K + K\cdot n + K - 1 \Rightarrow \mathcal{O}(n^2)$ (primijetite da je kovarijacijska matrica simetrična, pa je potrebno pohraniti samo dijagonalu i jedan trokut matrice). Ovdje, dakle, imamo **kvadratnu ovisnost** parametara o broju značajki (zbog kovarijacijske matrice). Procijeniti tolike parametre bi mogao biti problem, pogotovo onda kada je $n$ velik a $N$ malen.

Za procjenu parametara opet možemo koristiti MLE:

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

Što mislite, je li ovo **linearan ili nelinearan model**? Tj., je li granica između klasa linearna (hiperravnina) ili nelinearna (hiperpovršina)? I zašto je to uopće bitno? Bitno je, jer nelinearnost granice direktno upućuje na složenost (kapacitet) modela. Ako model ima linearnu granicu, onda je manje složenosti, i moguće prejednostavan za probleme koji iziskuju nelinearnu granicu između klasa. Da bismo utvrdili je li granica linearna ili ne, najbolje je da je izračunamo, odnosno izrazimo kao jednadžbu. Granica između klasa $y=1$ i $y=0$ jesu točke (odnosno hiperpovršina) za koje:

$$h_1(\mathbf{x}) = h_0(\mathbf{x})$$

tj. točke za koje vrijedi:

$$h_1(\mathbf{x}) - h_0(\mathbf{x}) = 0$$

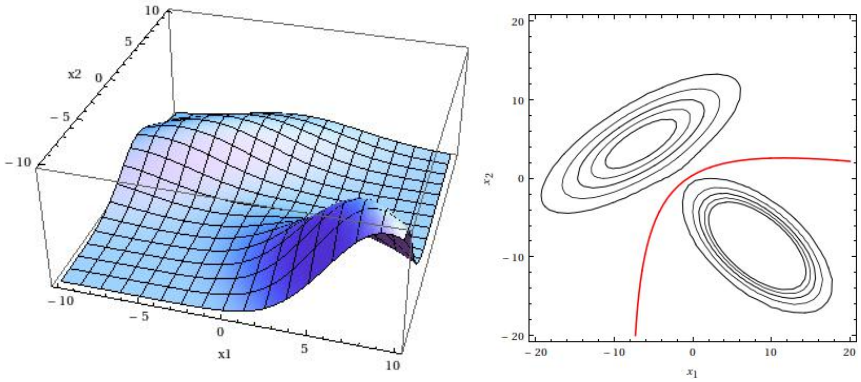
Model za klasu $j$ definirali smo kao (ekspandiramo kvadrat):

$$\begin{aligned} h(\mathbf{x})_j &= -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) + \ln P(y=j) \\ &= -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}\big(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\big) + \ln P(y=j) \end{aligned}$$

Granica između dva modela, za klase $y=1$ i $y=0$, bila bi onda:

$$\begin{aligned} h_{10}(\mathbf{x}) &= h_1(\mathbf{x}) - h_0(\mathbf{x}) \\ &= -\frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{1}{2}\big(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\big) + \ln P(y=1) \\ &\quad - \Big(-\frac{1}{2}\ln|\boldsymbol{\Sigma}_0| - \frac{1}{2}\big(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\big) + \ln P(y=0)\Big) \\ &= \ldots\mathbf{x}^{\mathrm{T}}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\mathbf{x}\ldots \end{aligned}$$

Vidimo da će granica biti **nelinearna** (točnije: kvadratna, tj. opisana parabolom) jer postoji član koji kvadratno ovisi o $\mathbf{x}$. To izgleda ovako: 10




Lijeva slika prikazuje gustoće dviju izglednosti klasa. Desna slika prikazuje to isto, ali pomoću izokonture gustoće vjerojatnosti, i prikazuje odgovarajuću granicu između dviju klasa (crvena parabola). To su točke za koje $h_1(\mathbf{x}) = h_0(\mathbf{x})$.

Ovo je lijepo, međutim problem bi moglo biti to što kvadratni model ima puno parametara: $\mathcal{O}(n^2)$. To bi onda značilo da se lako može dogoditi da model bude prenaučen. Zato ćemo sada razmotriti neka pojednostavljenja modela, koja će dovesti do **linearnosti**, a time i do manjeg broja parametara. Kako možemo pojednostaviti model? Moramo uvesti dodatne **induktivne pretpostavke**. Moguća su razna pojednostavljnja ovog modela, a mi ćemo pokazati tri inkrementalna pojednostavljenja. Dakle, krenuvši od ovog modela, uvodit ćemo dodatne pretpostavke i izvesti tri sve jednostavnija modela.

# 5 Varijante Gaussovog Bayesovog klasifikatora

Prvo pojednostavljenje koje ćemo napraviti jest da pretpostavimo da sve klase imaju identičnu kovarijacijsku matricu. Takvu kovarijacijsku matricu zovemo **dijeljena (vezana) kovarijacijska matrica** (engl. *shared (tied) covariance matrix*). Kod procjene, tu dijeljenu matricu računamo kao težinski prosjek procjena kovarijacijskih matrica za pojedinačne klase: 11

$$\hat{\boldsymbol{\Sigma}} = \sum_j \hat{\mu}_j \hat{\boldsymbol{\Sigma}}_j$$

gdje je $\hat{\mu}_j$ apriorna vjerojatnost klase $j$ (odnosno parametar multinulijeve distribucije za varijablu $y$). Ideja je da klase čija je apriorna vjerojatnost manja manje doprinose dijeljenoj kovarijacijskoj matrici.

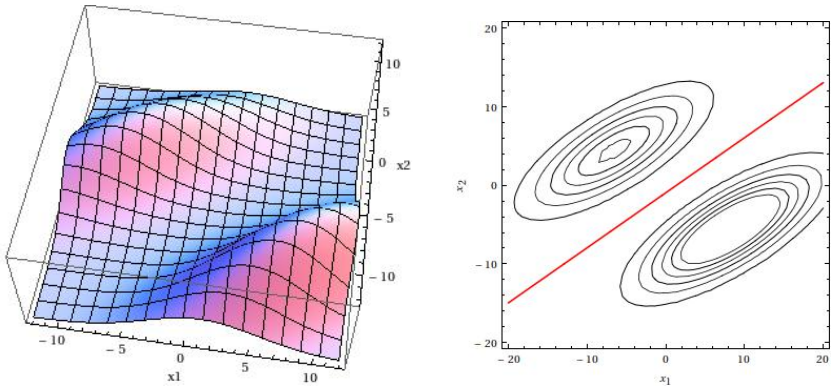
Uz dijeljenu kovarijacijsku matricu, model za klasu $j$ izgleda ovako:

$$\begin{aligned} h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\ &= \cancel{-\frac{n}{2}\ln 2\pi} - \cancel{\frac{1}{2}\ln|\boldsymbol{\Sigma}|} - \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j) + \ln P(y=j) \end{aligned}$$

Prva dva člana smo izbacili jer su konstante, tj. isti su za sve klase, pa nikako ne utječu na klasifikacijsku odluku. Granica između dviju klasa sada je:

$$\begin{aligned} h_{10}(\mathbf{x}) &= h_1(\mathbf{x}) - h_0(\mathbf{x}) \\ &= \cancel{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \ln P(y=1) \\ &\quad \cancel{\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \ln P(y=0) \\ &= \mathbf{x}^{\mathrm{T}}\underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}_{\mathbf{w}} \underbrace{-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + \ln\frac{P(y=1)}{P(y=0)}}_{w_0} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \end{aligned}$$

ovdje smo samo vektorom $\mathbf{w}$ zamijenili izraz $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, koji je $n$-dimenzijski vektor parametara, te smo skalarom $w_0$ zamijenili preostali dio izraza, koji je doista skalar. Iz toga je postalo jasno da se granica između klasa može opisati **linearnim modelom** $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$. Dakle, zaključujemo da, ako je kovarijacijska matrica dijeljena između klasa, onda je granica između klasa linearna. Pogledajmo kako bi to izgledalo u dvodimenzijskome ulaznom prostoru: 12




Izokonture izglednosti klasa su elipse s istim nagibom. Zbog toga je granica između klasa linearna.

Koliko ovaj model ima parametara? Odogovor je: $\frac{n}{2}(n+1) + nK + K - 1$. Imamo, dakle, $K$ puta manje parametara nego kod modela s nedijeljenom kovarijacijskom matricom. Nažalost, to je i dalje $\mathcal{O}(n^2)$. Je li nam se to isplatilo? Ovisi. Nekad će se možda isplatiti: ako su izvori šumova u svim klasama slični, a imamo puno klasa, onda je ovakav model možda bolji (bolje generalizira).

## 5.1 Drugo pojednostavljenje

Sada ćemo napraviti dodatno pojednostavljenje: osim što ćemo pretpostaviti da je kovarijacijska matrica dijeljena, pretpostavit ćemo i da je **dijagonalna**:

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$$

To znači da su kovarijacije između značajki jednake nuli, odnosno da nema linearne zavisnosti između značajki. (Pazite to ne znači nužno da nema nikakvih zavisnosti! Sjetite se: ako je

kovarijacija jednaka nuli, to znači da nema linearne zavisnosti, ali može biti da ima nelinearne zavisnosti.) Ako je kovarijacijska matrica dijagonalna, onda je lako izračunati njezinu determinantu i inverz:

$$|\boldsymbol{\Sigma}| = \prod_i \sigma_i^2$$
$$\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_i^2)$$

Uvrstimo to u izraz za izglednost klase:

$$\begin{aligned} p(\mathbf{x}|y=j) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\Big) \\ &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^n \sigma_i} \exp\Big(-\frac{1}{2}\sum_{i=1}^n \Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \prod_{i=1}^n \exp\Big(-\frac{1}{2}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big\{-\frac{1}{2}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big\} \\ &= \prod_{i=1}^n \mathcal{N}(\mu_{ij}, \sigma_i^2) \end{aligned}$$

Što smo dobili? Dobili smo umnožak univarijatnih Gaussovih distribucija, po jednu za svaku značajku:

$$p(\mathbf{x}|y) = \prod_{i=1}^n p(x_i|y)$$

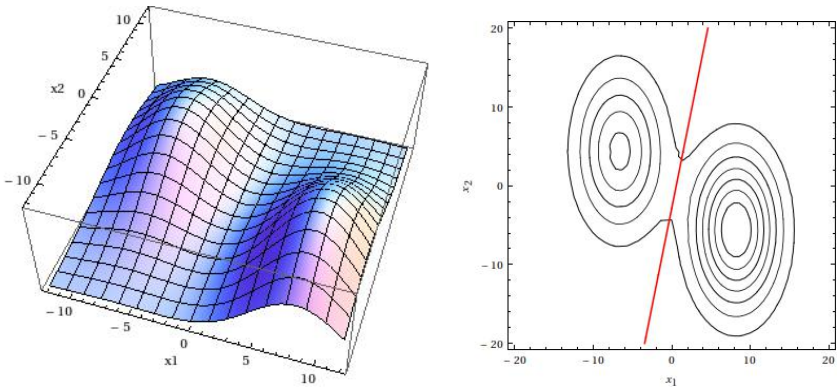
Ovo zapravo znači da su značajke $x_i$ međusobno **uvjetno nezavisne** uz zadanu klasu $y$. Bayesov klasifikator s takvom pretpostavkom naziva se **naivan Bayesov klasifikator**. O tome ćemo više drugi put. Izveli smo, dakle, naivan Bayesov klasifikator za kontinuirane značajke. Model tog klasifikatora je:

13

$$\begin{aligned} h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\ &= \ln \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\Big(-\frac{1}{2}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big) + \ln P(y=j) \\ &= \sum_{i=1}^n \ln \cancel{\frac{1}{\sqrt{2\pi}\sigma_i}} + \sum_{i=1}^n \Big(-\frac{1}{2}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big) + \ln P(y=j) \\ &\Rightarrow -\frac{1}{2}\sum_{i=1}^n \Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2 + \ln P(y=j) \end{aligned}$$

Primijetimo da će predikcija modela za klasu $j$ biti proporcionalna normiranoj euklidskoj udaljenosti između primjera $\mathbf{x}$ i vektora srednjih vrijednosti $\boldsymbol{\mu}_j$.

Kako ovdje izgleda granica između klasa? Granica je i dalje linearna, čim je matrica kovarijacije dijeljena. To se malo slabije vidi iz gornjeg izraza za model, jer se u izrazu pojavljuje $x_i$, koji kvadriramo, međutim ti kvadratni članovi se poništavaju jer su identični za svaki par klasa (za isti ulaz). U dvodimenzijskome ulaznom prostoru to izgleda ovako (izokonture su elipse poravnate s osima):




Ovaj model ima $n + n \cdot K + K - 1$ parametara, što je $\mathcal{O}(n)$. Drugim riječima, napokon smo dobili linearnu ovisnost broja parametara o broju značajki.

## 5.2 Treće pojednostavljenje

Konačno, treće pojednostavljenje jest pretpostaviti dijeljenu i **izotropnu kovarijacijsku matricu**:

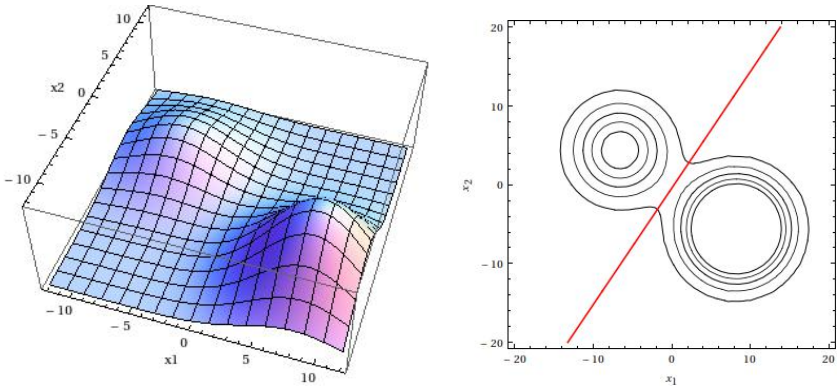
$$\mathbf{\Sigma} = \sigma^2 \mathbf{I}$$

Dakle, pretpostavljamo da je kovarijacijska matrica za sve klase identična, da je dijagonalna (značajke nisu linearno zavisne) i da su varijance za svaku značajku identične, što znači da pretpostavljamo da je raspon svih značajki isti. Ima li to smisla? Da, ako smo skalirali značajke izvan modela (kao dio predobrade podataka). Inače baš i ne.

Model je sada:

$$h_j(\mathbf{x}) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu_{ij})^2 + \ln P(y = j)$$

U dvodimenzijskome ulaznom prostoru to izgleda ovako (izokonture su kružnice):




Koliko ovaj model ima parametara? Odgovor je: $1 + Kn + K - 1$, dakle broj parametara linearan je u broju značajki. (Usput, primijetite da ne možemo odbaciti parametar $\sigma^2$, premda je identičan za svaku klasu, jer apriorne vjerojatnosti $P(y = j)$ nisu identične za svaku klasu.)

Sada se postavlja pitanje: od svih ovih modela koje smo izveli, koji bismo odabrali? Model s punom kovarijacijskom matricom ili jedno od tri pojednostavljenja? Ili neka druga pojednostavljenja, npr.:

| Pretpostavka na kov. matricu | Kov. matrica | Broj parametara | |
|---|---|---|---|
| Različite, hiperelipsoidi | $\mathbf{\Sigma}_j$ | $Kn(n+1)/2 + Kn$ | $\mathcal{O}(n^2)$ |
| Dijeljena, hiperelipsoidi | $\mathbf{\Sigma}$ | $n(n+1)/2 + Kn$ | $\mathcal{O}(n^2)$ |
| Različite, poravnati hiperelipsoidi | $\mathbf{\Sigma}_j = \text{diag}(\sigma_{i,j}^2)$ | $2Kn$ | $\mathcal{O}(n)$ |
| Dijeljena, poravnati hiperelipsoidi | $\mathbf{\Sigma} = \text{diag}(\sigma_i^2)$ | $n + Kn$ | $\mathcal{O}(n)$ |
| Različite, hipersfere | $\mathbf{\Sigma}_j = \sigma_j^2\mathbf{I}$ | $K + Kn$ | $\mathcal{O}(n)$ |
| Dijeljena, hipersfere | $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ | $1 + Kn$ | $\mathcal{O}(n)$ |

Kao i uvijek: najjednostavniji način jest napraviti **unakrsnu provjeru**: ispitati svaku varijantu modela na izdvojenom skupu označenih primjera te odabrati onu varijantu koja ima najmanju ispitnu pogrešku. Ta varijanta očekivano najbolje generalizira na neviđene primjere.

## Sažetak

- Osnovna pravila teorije vjerojatnosti su **pravilo umnoška** i **pravilo zbroja**, pomoću kojih smo izveli **Bayesovo pravilo**
- Bayesov klasifikator pomoću Bayesovog pravila izračunava **aposteriornu vjerojatnost** oznake primjera
- Bayesov klasifikator je **generativan** i **parametarski** model
- Generativni modeli opisuju nastanak podataka i omogućuju uključivanje **pozadinskog znanja** i **interpretabilnost**
- **Gaussov Bayesov klasifikator** izglednost klase modelira pomoću multivarijatnog Gaussa i taj model koristimo za kontinuirane značajke
- Uvođenjem različitih pretpostavki na **kovarijacijsku matricu** dobivamo **jednostavnije** varijante Gaussovog Bayesovog klasifikatora

## Bilješke

1 Kao što smo vidjeli prošli put, ako je $p(x|y)$ općenito neka vjerojatnost od $x$ za fiksirani $y$, onda je to ujedno i izglednost od $y$ za fiksirani $x$. Zato gustoću vjerojatnosti $p(\mathbf{x}|y)$ zovemo **izglednost klase** (ili oznake) $y$. Naravno, to je ujedno i vjerojatnost od $\mathbf{x}$ za danu klasu (oznaku) $y$.

2 Naravno, reći da dekompozicija nekog $X$ na dijelove $X_i$ nema smisla jer je kompozicija dijelova $X_i$ jednaka $X$ sama po sebi nema smisla. Dekompozicija je jedan od osnovnih alata znanosti, posebno računarstva. V. https://en.wikipedia.org/wiki/Decomposition_(computer_science).

3 Ovdje postoji opasnost da **maximum aposteriori hipotezu (MAP)** pobrkamo s **procjeniteljem maksimum aposteriori (MAP)**. Zabuna je očekivana, jer se radi o potpuno istom mehanizmu: maksimiziramo aposteriornu vjerojatnost. Međutim, hipoteza MAP i procjenitelj MAP su ipak dvije različite stvari jer pripadaju različitim razinama. Procjenitelj MAP je način procjene parametara modela (npr. Bayesovog klasifikatora). Jednom kada smo procijenili parametre (tj. naučili model), možemo raditi klasifikaciju. Način na koji onda odlučujemo u koju klasu klasificirati primjer određuje nam hipoteza MAP. Dakle, prvo za treniranje modela koristimo procjenitelj MAP, a onda pomoću naučenog modela radimo klasifikaciju primjera pomoću hipoteze MAP. Također, ništa nas ne sprječava da parametre procijenimo nekom drugom metodom (npr. MLE), a onda za predikciju koristimo MAP hipotezu. Vrijedi i obrnuto: možemo koristiti procjenitelj MAP, ali za predikciju ne koristiti MAP hipotezu nego neko drugo pravilo odlučivanja. Međutim, tipično ipak koristimo MAP hipotezu, jer je ona optimalna u smislu **bayesovske teorije odlučivanja** (naime, hipoteza MAP minimizira vjerojatnost pogreške; v. poglavlje 4.1.1 u skripti).

4 **Generativni modeli** ponekad se u literaturi nazivaju **modeli zajedničke vjerojatnosti** (engl. *joint*

*models*), dok se probabilistički diskriminativni modeli nazivaju i **uvjetni modeli** (engl. *conditional models*), budući da modeliraju uvjetnu (aposteriornu) vjerojatnost.

5 **Teorem o nebesplatnom ručku** (engl. *no free lunch theorem, NFL*) su zapravo dva teorema: jedan za strojno učenje, a drugi za postupke optimizacije. Oba je predložio američki matematičar David Wolpert (drugi teroem u suautorstvu sa Williamom Macreadyjem) 1996. odnosno 1997. godine. NFL za strojno učenje opisan je u Wolpert (1996). Neformalni opis možete naći na `http://www.no-free-lunch.org/`. Lijep opis i dokaz možete naći u Shalev-Shwartz and Ben-David (2014) (poglavlje 5.1). Argument NFL-a u osnovi je povezan s Humovim problemom indukcije; `https://plato.stanford.edu/entries/induction-problem/`.

6 Rivalstvo između generativnih i diskriminativnih modela otvoreno je pitanje. Pogledajte, na primjer, (Ng and Jordan, 2001) i (Xue and Titterington, 2008).

7 Ovaj dio uglavnom prati poglavlja 5.4–5.6 iz (Alpaydin, 2020).

8 Primijetite da izglednost svake klase modeliramo jednom multivarijatnom Gaussovom gustoćom vjerojatnosti. To znači da su gustoće vjerojatnosti za svaku klasu unimodalnu. To zapravo znači da pretpostavljamo da je klasa homogena, u smislu da ima jedan prototipni primjer oko kojega se svi drugi primjeri rasipaju zbog šuma. Složeniji model bi bio onaj koji omogućio modeliranje nehomogenih klasa kao kombinaciju (superpoziciju) više multivarijatnih Gaussovih gustoća vjerojatnosti. Nehomogene klase u stvarnosti nisu rijetkost. Npr., klasifikacija rukom pisanih znamenki ima nehomogenu klasu za znamenku 7, jer se ona može pisati na dva načina (s horizontalnom crticom ili bez nje, uglavnom ovisno o tome jeste li Europljanin ili Amerikanac). Bayesov klasifikator podrazumijeva da su klase homogene. Međutim, postoje modeli koji mogu modelirati nehomogene klase koje se sastoje od kombinacije Gaussovih gustoća vjerojatnosti. To su **modeli Gaussove mješavine** (engl. *mixture of Gaussians, MoG*). O tim modelima pričat ćemo u kontekstu grupiranja podataka.

9 Prisjetimo se napomene u vezi kovarijacijske matrice $\boldsymbol{\Sigma}$: ta matrica je uvijek **pozitivno semidefinitna**, što znači da $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x} \geqslant 0$, a isto tako $\Delta^2 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} \geqslant 0$, tj. za Mahalanobisovu udaljenost vrijedi $\Delta \geqslant 0$ Ali, da bi PDF bila dobro definirana, $\boldsymbol{\Sigma}$ mora biti **pozitivno definitna**: $\Delta^2 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} > 0$ za ne-nul vektor $\mathbf{x}$. Ako je $\boldsymbol{\Sigma}$ pozitivno definitna, onda je nesingularna: $|\boldsymbol{\Sigma}| > 0$ i postoji $\boldsymbol{\Sigma}^{-1}$. Ako $\boldsymbol{\Sigma}$ nije pozitivno definitna, uzroci su: $\mathrm{Var}(x_i) = 0$ (beskorisna značajka) ili $\mathrm{Cov}(x_i, x_j) = 1$ (multikolinearnost – redundantan par značajki). Također, imajte na umu da matrica može biti skoro linearna (što će biti indicirano njezinim visokim **kondicijskim brojem**.

10 Model **Gaussovog Bayesovog klasifikatora** s kvadratnom granicom istovjetan je modelu **kvadratne diskriminantne analize** (engl. *quadratic discriminant analysis, QDA*). Kvadratna diskriminantna analiza je poopćenje **linearne diskriminantne analize** (engl. *linear discriminant analysis, LDA*), kod koje je granice između klasa linearna. Linearnu diskriminantnu analizu predložio je 1936. godine predložio Ronald Fisher, jedan od najvećih statističara. U nastavku ćemo uvesti pojednostavljenja Gaussovog Bayesovog klasifikatora, koja će nas dovesti do modela istovjetnog linearnoj diskriminativnoj analizi.

11 U statistici, pretpostavka da je kovarijacijska matrica identična za sve klase naziva se **homoskedastičnost** (homogenost varijance). Izračun kovarijacijske matrice kao težinske kombinacije kovarijacijski matrica pojedinačnih klasa naziva se **udružena varijanca** (engl. *pooled variance*).

12 Gaussov Bayesov klasifikator s dijeljenom kovarijacijskom matricom, kod kojega je granica između klasa linearna, istovjetan je ranije spomenutoj **linearnoj diskriminativnoj analizi (LDA)**. Linearnu diskriminativnu analizu nemojte brkati s također ranije spomenutom **Latentnom Dirichleovom alokacijom (LDA)**. Kratice su iste, ali metode su posve različite. Jedina zajedničkost im je da to što su obje metode generativne.

13 Preciznije, ovo znači da značajke **nisu linearno zavisne** za zadanu klasu $y$. Međutim, kao što smo već napomenuli, između značajki može postojati nelinearna zavisnost, premda nema linearne zavisnosti. Primijetite da i u tom slučaju – kada kovarijacija jest jednaka nuli ali varijable su ipak nelinearno zavisne – dobivamo naivan Bayesov model. To je zato što u multivarijatnoj Gaussovoj distribuciji kroz kovarijacijsku možemo modelirati samo linearnu zavisnost između varijabli – i to samo između parova varijabli. Ako su varijable nelinearno zavisne, to ne možemo modelirati. Mogli bismo reći da

je kod kontinuiranog Bayesa dio "tereta naivnosti" preuzela i sama Gaussova distribucija, jer njenim odabirom implicitno pretpostavljamo da značajke mogu biti samo linearno zavisne.

## Literatura


E. Alpaydin. *Introduction to machine learning*. MIT press, 2020.

A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841–848, 2001.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

J.-H. Xue and D. M. Titterington. Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural processing letters*, 28(3):169, 2008.