13. Procjena parametara

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

Jan Šnajder, predavanja, v2.3

Drugu polovicu predmeta započet ćemo razmatranjem probabilističkih (odnosno vjerojatnosnih) modela. Općenito govoreći, to su modeli koji se oslanjaju na **teoriju vjerojatnosti** kako bi modelirali distribuciju primjera i njihovih oznaka, i na temelju toga radili klasifikaciju ili regresiju.

Osnovni mehanizam kod probabilističkih modela jest **procjena parametara (engl. parameter estimation)**: kako, na temelju označenog skupa podataka, odrediti koji su parametri distribucije podataka. Kod probabilističkih modela, parametri koje tako procijenimo su onda ujedno i parametri našeg modela. Naš označeni skup podataka je ustvari **uzorak**, a procjenom parametara na temelju uzorka zapravo se bavi **statistika**, pa ćemo danas dakle malo pozabaviti statistikom. Danas ćemo objasniti samo osnovnu ideju procjene parametara, a onda ćemo idući put pričati o konkretnim metodama procjene parametara.

Također, u predavanjima koje slijede vratit ćemo se na neke stvari koje smo već bili naučili u prvom dijelu predmeta i pogledati ih ponovo, ali ovog puta gledat ćemo na njih kao na problem procjene parametara. To će dovesti do nekih zanimljivih spoznaja.

1 Motivacija

Prednost probabilističkih modela je trojaka. Prvo, ti su modeli temeljeni na **teoriji vjerojatnosti**, koja je vrlo razrađena, pa dakle imamo dobru teorijsku podlogu koju dobro razumijemo. Druga prednost probabilističkih modela jest što **modeliraju vjerojatnosti**, dakle imat ćemo informaciju koliko je klasifikacija pouzdana. Treća prednost je da su probabilistički modeli idealni u situacijama kada imamo nekakvo **apriorno znanje** o problemu, i želimo to znanje ugraditi u naš model, kako bi model profitirao od tog znanja i radio još točniju klasifikaciju/regresiju. To znanje može biti jednostavno, ali i vrlo složeno, što može dovesti do modela složenih struktura. Nadalje, probabilistički modeli će raditi dosta dobro i onda kada nemamo puno podataka, ako u model uspijemo ugraditi pretpostavke o podatcima koje su točne.

Postoje **parametarski i neparametarski** probablistički modeli. Mi ćemo se u ovom predmetu fokusirati na parametarske modele, koji se ionako češće koriste. Prisjetimo se: parametarski modeli su modeli koji pretpostavljaju da se podatci pokoravaju nekoj vjerojatnosnoj distribuciji, koju moramo unaprijed odabrati.

Za motivaciju, spomenimo jedan konkretan – i najjednostavniji – primjer probabilističkog klasifikatora: **Bayesov klasifikator**. Bayesov klasifikator modelira vjerojatnost oznake za zadani primjer, $P(y|\mathbf{x})$, i to čini na temelju dviju pretpostavljenih vjerojatnosti: **vjerojatnosti primjera za zadanu oznaku** $P(\mathbf{x}|y)$ i **apriorne vjerojatnosti oznaka** $P(y)$. Točnije, vjerojatnost oznake za dani primjer bit će proporcionalna umnošku ovih dviju vjerojatnosti:

$$P(y|\mathbf{x}) \propto P(\mathbf{x}|y)P(y)$$

Dakle, da bismo izračunali vjerojatnost neke oznake y za zadani primjer $\mathbf{x}$, moramo izračunati dvije vjerojatnosti: vjerojatnost da slučajna varijabla $\mathbf{x}$ poprimi neku konkretnu vrijednost ako znamo da je slučajna varijabla y poprimila neku konkretnu vrijednost (uvjetna vjerojatnost) i

vjerojatnost da slučajna varijabla y poprimi neku konkretnu vrijednost. Pitanje je kako ćemo izračunati te vjerojatnosti. To ćemo napraviti tako da ćemo – ovisno o vrsti podataka – odabrati dvije konkretne teorijske distribucije – npr., Gaussovu distribuciju za $P(y|\mathbf{x})$ i Bernoullijevu distribuciju za $P(y)$ – te ćemo pretpostaviti da se podatci pokoravaju tim distribucijama. Onda ćemo **procijeniti parametre** tih distribucija na temelju podataka. Kada to imamo, možemo raditi predikciju za nove primjere, koja se onda svodi na izračun vjerojatnosti $P(y|\mathbf{x})$.

Možda ćete sada uočiti da procjena parametara zapravo znači da određujemo parametre modela na temelju podataka. Nije li to zapravo učenje modela? Jest! Međutim u svijetu statistike, to se zove procjena parametara. Dakle, kada statističar kaže "procjena parametara modela", to je kao da kaže "učenje/treniranje modela".

Statistika se temelji na teoriji vjerojatnosti. Hajdemo se malo prisjetiti osnova teorija vjerojatnosti. 1

2 Slučajne varijable

2.1 Slučajna varijabla i distribucija

Kod probabilističkih modela, primjere $\mathbf{x}$ i oznake y modelirat ćemo kao **slučajne varijable**. 2
Slučajna varijabla X ima unaprijed definiran skup vrijednosti $\{x_j\}$ (diskretan ili kontinuiran, konačan ili beskonačan). Ako je skup vrijednosti diskretan, onda govorimo o **diskretnoj slučajnoj varijabli**. Npr., oznaka klase y je diskretna slučajna varijabla.

Vrijednost $P(X = x)$ jest **vjerojatnost** da diskretna slučajna varijabla X poprimi vrijednost x, tj. vjerojatnost da se slučajna varijabla X realizira kao x. U nastavku ćemo umjesto $P(X = x)$ pisati kraće $P(x)$. Vrijedi $P(x_i) \geqslant 0$ i $\sum_i P(x_i) = 1$. Time je zapravo definirana **diskretna distribucija (razdioba) vjerojatnosti**. 3

Navedeno vrijedi za diskretnu slučajnu varijablu. Ako je skup mogućih vrijednosti slučajne varijable kontinuiran, npr., prosjek ocjena studenta, onda govorimo o **kontinuiranoj (neprekidnoj) slučajnoj varijabli**. Za kontinuiranu (neprekidnu) slučajnu varijablu definiramo **funkciju gustoće vjerojatnosti** (engl. *probability density function; PDF*), koju označavamo kao $p(x)$. Za 4
funkciju gustoće vjerojatnosti vrijedi $p(x) \geqslant 0$ i $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$. 5

Kažemo da funkcijom gustoće vjerojatnosti $p(x)$ definirana **kontinuirana distribucija (razdioba) vjerojatnosti**. U nastavku ćemo, kao što je tipično u literaturi iz strojnog učenja, koristiti izraz *"gustoća $p(x)$"* ili (pomalo neprecizno) *"distribucija $p(x)$"*, misleći pritom na funkciju gustoće vjerojatnosti $p(x)$.

2.2 Očekivanje, varijanca i kovarijanca

Budući da slučajne varijable mogu poprimiti različite vrijednosti, korisno je da ih pokušamo nekako okarakterizirati. U nastavku ćemo se ograničiti na **numeričke** slučajne varijable (bilo diskretne ili kontinuirane). Numeričke slučajne varijable možemo okarakterizirati preko njihovog očekivanja, varijance i kovarijance.

Krenimo od očekivanja slučajne varijable. Prosječna vrijednost diskretne slučajne varijable X čija je distribucija $P(x)$ naziva se **(matematičko) očekivanje** ili **očekivana vrijednost** varijable X i definira se kao:

$$\mathbb{E}[X] = \sum_x xP(x)$$

To je zapravo težinska suma vrijednosti varijable.

▸ **PRIMJER**

Zamislite da student rješava blic s pitanjima koja imaju 4 ponuđena odgovora, i samo jedan odgovor je točan. Točan odgovor nosi 1 bod, a netočan -0.5 bodova (predrastično, kao što ćemo vidjeti).

Prepostavimo da student pojma nema i da odgovara nasumično. To znači da su bodovi koje će dobiti na svakom pojedinačnom zadatku slučajna varijabla. Budući da student odgovara nasumično, vjerojatnost da pogodi 1 točan odgovor od 4 ponuđena je 0.25, a vjerojatnost da promaši je 0.75. Dakle, očekivanje bodova na jednom zadatku je:

$$\mathbb{E}[X] = 1 \cdot 0.25 + (-0.5) \cdot 0.75 = -0.125$$

U slučaju kontinuirane slučajne varijable X s gustoćom vjerojatnosti $p(x)$, očekivanje je

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x)\, \mathrm{d}x$$

Osim očekivanja, drugi način da se okarakterizira slučajna varijabla je **varijanca**. Varijanca slučajne varijable X iskazuje koliko vrijednosti varijable variraju oko očekivane vrijednosti:

$$\mathrm{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2\,.$$

Konačno, zanimljiva će nam biti i **kovarijanca**. Kovarijanca opisuje odnos između **dviju** slučajnih varijabli, odnosno opisuje u kojoj mjeri slučajne varijable zajednički variraju oko svojih očekivanih vrijednosti. Kovarijanca varijabli X i Y definirana je kao

$$\mathrm{Cov}(X, Y) = \sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Vrijedi $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$. Primijetite da je kovarijanca varijable sa samom sobom varijanca, tj. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ odnosno $\sigma_{X,X} = \sigma_X^2$.

▸ **PRIMJER**

Recimo da $X = \{1, 2\}$ i $Y = \{1, 2\}$. Neka su vjerojatnosti $P(X, Y)$ sljedeće:

$$P(1,1) = 0.5, P(1,2) = 0.1, P(2,1) = P(2,2) = 0.2$$

Marginalne vjerojatnosti onda su: 6

$$P(X = 1) = 0.6, P(X = 2) = 0.4, P(Y = 1) = 0.7, P(Y = 2) = 0.3$$

Očekivanja varijabli onda su:

$$\mathbb{E}[X] = 1 \cdot 0.6 + 2 \cdot 0.4 = 1.4$$
$$\mathbb{E}[Y] = 1 \cdot 0.7 + 2 \cdot 0.3 = 1.3$$

Varijance varijabli onda su:

$$\mathrm{Var}(X) = 0.6 \cdot (1 - 1.4)^2 + 0.4 \cdot (2 - 1.4)^2 = 0.24$$
$$\mathrm{Var}(Y) = 0.7 \cdot (1 - 1.3)^2 + 0.3 \cdot (2 - 1.3)^2 = 0.21$$

Korelacija između slučajnih varijabli X i Y onda je:

$$\begin{aligned}
\mathrm{Cov}(X, Y) = \sigma_{X,Y} &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \sum_{x \in \{1,2\}} \sum_{y \in \{1,2\}} P(x, y)(x - \mathbb{E}[X])(y - \mathbb{E}[Y]) \\
&= 0.5 \cdot (1 - 1.4) \cdot (1 - 1.3) + 0.1 \cdot (1 - 1.4) \cdot (2 - 1.3) + \ldots \\
&= 0.08
\end{aligned}$$

Za slučajne varijable X i Y za koje vrijedi $\mathrm{Var}(X) \neq 0$ i $\mathrm{Var}(Y) \neq 0$ definiran je **Pearsonov koeficijent korelacije**:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

Pearsonov koeficijent upućuje na to koliko su varijable X i Y međusobno **linearno zavisne**. Za savršenu pozitivnu linearnu ovisnost vrijedi $\rho_{X,Y} = 1$, dok za savršenu negativnu linearnu ovisnost vrijedi $\rho_{X,Y} = -1$.

Npr, visina u centimetrima i visina u metrima će biti savršeno linearno korelirane, starosna dob i broj godina radnog staža će biti pozitivno korelirane, premda ne savršeno, dok su, npr., starost automobila i njegova cijena negativno korelirane. Pritom primijetite da vrijednost koeficijenta korelacije ne ovisi o skalama varijabli X i Y (koeficijent korelacije je bezdimenzijska mjera).

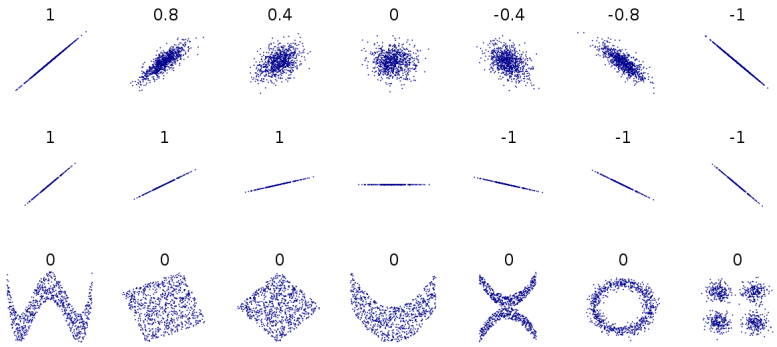
▸ **PRIMJER**

Za raniji primjer, koeficijent korelacije je:

$$\rho_{X,Y} = \frac{0.08}{\sqrt{0.24}\sqrt{0.21}} = \frac{0.08}{0.23} = 0.35$$

što je umjerena pozitivna korelacija.

Važno je ovdje primijetiti da Pearsonov koeficijent korelacije mjeri isključivo linearnu zavisnost dviju varijable. Varijable mogu biti nelinearno zavisne, a imati nizak Pearsonov koeficijent korelacije. To ilustriraju sljedeći primjeri: [7]



Slika prikazuje grafikone raspršenja (engl. *scatter plots*) za dvije varijable i vrijednosti Pearsonovog koeficijenta korelacije. U srednjem retku varijable su savršeno pozitivno ili negativno korelirane. Vidimo da iznos korelacije ne ovisi o nagibu pravca: nije bitno koliko se varijabla Y poveća ili smanji s povećanjem ili smanjenjem varijable X, već da li se to uvijek konzistentno događa. Slučaj u sredini (kada je varijabla Y konstanta) je slučaj za koji korelacija nije definirana. U prvome retku imamo slučajeve kada je korelacija manja od 1 (umjerena ili slaba) ili kada je jednaka nuli. Korelacija je jednaka nuli ako varijable nisu zavisne, što na grafikonu raspršenja vidimo kao "oblak". No, korelacija može biti jednaka nuli i kada varijable jesu zavisne, ali ta zavisnost nije linearna. To je prikazano u trećem retku. Za sve te slučajeve korelacija je jednaka nuli, premda je očito da između varijabli postoji neka zavisnost, koja se manifestira kao neki uzorak u grafikonu raspšenja. Specifično, u posljednjem (skroz desnom) slučaju u podatcima postoje grupe (klasteri) varijabli, međutim korelacija će tu i dalje biti nula jer je prosječni odmak točke od pravca (rezidual) jednak nuli. [8]

2.3 Višedimenzijska slučajna varijabla

U strojnom učenju, primjeri uobičajeno imaju više značajki, $\mathbf{x} = (x_1, \ldots, x_n)$. U teoriji vjerojatnosti takvu višedimenzijsku varijablu modeliramo **slučajnim vektorom**, $(X_1, \ldots, X_n)$. Često će nas zanimati i jesu li i kako značajke međusobno korelirane: npr., sjetimo se modela linearne regresije i problema multikolinearnosti, gdje želimo izbaciti značajke koje su korelirane. Jedan način da dobijemo uvid u korelacije između značajki jest da izračunamo korelacije između svih parova značajki. To nam daje matricu koju zovemo **matrica kovarijacije (kovarijacijska matrica)**, koju označavamo sa $\mathbf{\Sigma}$. Elementi te matrice su kovarijacije između svih parova varijabli:

$$\mathbf{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j) = \sigma_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

Kovarijacijska matrica je **kvadratna simetrična matrica** (jer $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$), koja na glavnoj dijagonali ima varijance varijabli (jer $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$), a izvan dijagonale kovarijance svih parova varijabli:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{pmatrix}$$

U matričnom računu, kovarijacijska je matrica definirana kao:

$$\mathbf{\Sigma} = \mathbb{E}\Big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}}\Big]$$

Primijetite da ovdje računamo vektorski produkt (vanjski produkt) vektora $(\mathbf{X} - \mathbb{E}[\mathbf{X}])$ sa samim sobom, a ne skalarni produkt, pa je rezultat matrica dimenzija $(n \times 1) \times (1 \times n) = n \times n$, a ne skalar.

Općenito, kovarijacijska matrica će imati ne-nul vrijednosti izvan dijagonale. Međutim, posebno će nam biti zanimljiva dva specifična slučaja:

(1) Ako su varijable $X_1, \ldots, X_n$ međusobno linearno nezavisne, onda $\mathrm{Cov}(X_i, X_j) = 0$ za $i \neq j$ i kovarijacijska je matrica **dijagonalna matrica**, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i^2)$;

(2) Ako nezavisne varijable $X_1, \ldots, X_n$ imaju jednaku varijancu, onda $\sigma_i^2 = \sigma^2$, pa kovarijacijska matrica degenerira u $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, gdje je $\mathbf{I}$ jedinična matrica. Takav slučaj nazivamo **izotropnom kovarijancom**.

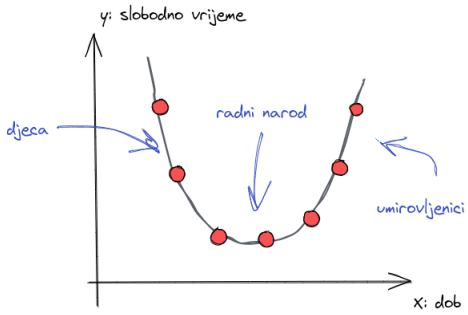
Kasnije će nam trebati i inverz kovarijacijske matrice. Kovarijacijska matrica je **pozitivno semidefinitna**, tj. $\mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{x} \geqslant 0$. To znači da ne mora nužno imati inverz, tj. da može biti singularna. (Kada bi matrica bila **pozitivno definitna**, onda bi nužno imala inverz.) Konkretno, kovarijacijska matrica neće imati inverz akko je neki od elemenata na dijagonali jednak nuli ili ako postoji linearna zavisnost između redaka odnosno stupaca matrice matrice dizajna $\mathbf{X}$. Sada se možemo pitati kada se to zapravo događa? Kada će neki element dijagonale kovarijacijske matrice biti jednak nuli? Budući da su na dijagonali nalaze varijance za svaku značajku, to se dogoditi samo ako neka značajka ima varijancu jednaku nuli, a to znači da je značajka konstantna. Konstantna značajka očito nije varijabla (jer ne varira) i ona nam je potpuno beskorisna: naime, ako joj je vriijednost jednaka za sve primjere, onda očito nema utjecaja na predikciju pojedinačnog primjera. Što je s linearnom zavisnošću redaka odnosno stupaca? Nju ćemo imati ako postoji **savršena multikolinearnost** značajki, tj. ako je neka značajka **redundantna** jer se može predvidjeti iz drugih značajki. Znamo već i da je multikolinearnost problem i kod poopćenih linearnih modela, gdje ona uzrokuje nestabilnost rješenja uslijed loše kondicionirane matrice dizajna.

2.4 Nezavisnost varijabli

Kovarijanca dakle mjeri linearnu zavisnost između varijabli. Međutim, već smo napomenuli da varijable mogu biti zavisne, a da ta zavisnost nije linearna.

▸ **PRIMJER**

Prisjetimo se primjera ovisnosti slobodnog vremena o dobi:



Premda ovdje očito postoji ovisnost varijable Y o varijabli X (i obrnuto), tu ovisnost nije moguće iskazati korelacijom odnosno kovarijacijom. Naime, ovdje za varijable X i Y vrijedi $\text{Cov}(X,Y) = \rho_{X,Y} = 0$. (Zašto? Zato jer je ova krivulja simetrična po Y osi, pa će se očekivanja $(X - \mathbb{E}[X])$ poništiti. Čim je grafikon raspršenja simetričan po jednoj od osi, korelacija je jednaka nuli.)

Nas će u strojnom učenju zanimati jesu li varijable općenito zavisne, i to bilo linearno ili nelinearno. Zato uvodimo pojam **nezavisnosti slučajnih varijabli**. Dvije slučajne varijable X i Y su **(stohastički) nezavisne** ako i samo ako:

$$P(X,Y) = P(X)P(Y)$$

Intuitivno, varijable X i Y su nezavisne ako je vjerojanost zajedničkog ishoda jednaka umnošku vjerojatnosti pojedinačnih ishoda. Npr., hoće li sutra u Samoboru pasti snijeg i hoće li Kim Kardashian sutra objaviti fotografiju na Instagramu nezavisni su događaji, i vjerojatnost zajedničkog ishoda jednostavno je umnožak vjerojatnosti pojedinačnih ishoda.

Ako su varijable X i Y nezavisne, onda vrijedi $\text{Cov}(X,Y) = \rho_{X,Y} = 0$. Dakle, **nezavisne varijable su nekorelirane**. No, kao što smo već napomenuli i kao što smo vidjeli na gornjem primjeru (ovisnost prihoda o dobi), obrat općenito ne vrijedi: koeficijent korelacije može biti jednak nuli, a da su varijable ipak nelinearno zavisne (jer koeficijent korelacije mjeri isključivo linearnu zavisnost varijabli).

Ok, sad smo se prisjetili što je to slučajna varijabla i koja su njezina svojstva. Kao što smo već rekli, kod parametarskih probabilističkih modela mi ćemo pretpostaviti da se slučajne varijable pokoravaju nekoj distribuciji. Te distribucije nećemo izmišljati, nego ćemo koristiti poznate **teorijske distribucije**. Pa, pogledajmo koje su to konkretno distribucije s kojima ćemo raditi.

3 Osnovne vjerojatnosne distribucije

Odabir vjerojatnosne distribucije prvenstveno ovisi o tome s kakvim podatcima radimo: jesu li podatci diskretni ili kontinuirani, te jesu li jednodimenzijski ili višedimenzijski. U strojnom učenju tipično koristimo:

- Diskretna varijabla:
  - Jednodimenzijska:

        * Binarna: **Bernoullijeva distribucija**
        * Viševrijednosna: **Kategorička (multinulijeva) distribucija**
    - Višedimenzijska: Konkatenirani vektor binarnih/viševrijednosnih varijabli
- Kontinuirana varijabla:
    - Jednodimenzijska: **univarijatna Gaussova (normalna) distribucija**
    - Višedimenzijska: **multivarijatna Gaussova (normalna) distribucija**

Premda su ovo najčešće korištene distribucije, u strojnom učenju susrećemo i neke druge, npr., beta-distribuciju, Dirichletovu distribuciju i Laplaceovu distribuciju. 9

Ove se distribucije koriste kako za modeliranje primjera $\mathbf{x}$, tako i za modeliranje oznaka y. Ulazni primjer tipično je **višedimenzijska** slučajna varijabla, a oznaka je **jednodimenzijska** slučajna varijabla.

▸ **PRIMJER**

Da vidimo kako biste se snašli s odabirom distribucija.

- Recimo da radimo predviđanje prosjeka ocjena studenta četvrte godine na temelju ocjena iz prethodne tri godine. Očito, riječ je o regresijskom problemu. Koju distribuciju biste iskoristili za vektor značajki $\mathbf{x}$? Odgovor je: multivarijatna Gaussova distribucija, jer imamo više kontinuiranih značajki.
- Pretpostavite da, umjesto prosjeka prethodnih godina, imate ocjene za svaki predmet (od 2 do 5). Kako biste sada modelirali vektor $\mathbf{x}$? Odgovor je: opet multivarijatna Gaussova distribucija. Nema veze što su značajke cijeli brojevi, bitno je da su brojevi, a ne kategoričke vrijednosti. Čim radimo s brojevima, bilo cijelim ili realnim, znači da između njih postoji potpuni uređaj (1 je manje od 2, koji je manji od 3 itd.), i taj je uređaj bitan i želimo ga modelirati. Kada bismo koristili kategoričku distribuciju, ne bismo modelirali taj uređaj, jer između kategorija nema nikakvog uređaja.
- Sada pretpostavite da, umjesto ocjena za svaki predmet, imate podatak o tome u koju je srednju školu student/ica išao/la i podatak o tome iz kojeg je grada. Kako biste modelirali značajku $\mathbf{x}$? Odgovor je: sa dvije multinulijeve značajke. A koliko mogućih vrijednosti imaju te značajke? Odgovor je: onoliko koliko ima različitih škola odnosno gradova.
- A sad pretpostavite da, umjesto regresije (predviđanje prosjeka ocjena četvrte godine), radite klasifikaciju (tko će pasti godinu?). Kako biste modelirali varijablu oznake y? Odgovor je: Bernoullijevom varijablom, jer postoje dvije moguće vrijednosti.
- Konačno, pretpostavite da, umjesto predviđanja tko će pasti godinu, radite predviđanje tko će, u godinu dana nakon završetka studija, pronaći posao u HR, tko u inozemstvu, a tko neće pronaći posao. Kako biste sada modelirali varijablu oznake y? Odgovor je: multinulijeva varijablom s tri moguće vrijednosti.

Pogledajmo sada malo detaljnije svaku od ovih distribucija, premda smo se s većinom njih već susreli.

3.1 Bernoullijeva distribucija

Bernoullijeva distribucija modelira vjerojatnost diskretne slučajne varijable s dva moguća ishoda, $\{0, 1\}$ – dakle, **binarne varijable**. Bernoullijeva distribucija ima svega jedan parametar, μ, koji određuje koja je vjerojatnost da $x = 1$. Očito, vjerojatnost da $\mu = 0$ je onda $1 - \mu$. To pišemo ovako:

$$P(X = x|\mu) = \begin{cases} \mu & \text{ako } x = 1 \\ 1 - \mu & \text{inače} \end{cases} = \mu^x (1 - \mu)^{1-x}$$

Ako uvrstimo ovu distribuciju u definiciju za očekivanje odnosno varijance, dobivamo $\mathbb{E}[X] = \mu$ odnosno $\text{Var}(X) = \mu(1 - \mu)$.

3.2 Kategorička (aka "multinulijeva") distribucija

Kategorička distribucija (također, u zadnje vrijeme nazivana **"multinulijeva"**, po analogiji s Bernoulijevom) modelira slučajnu varijablu koja poprima jednu (i samo jednu) od K mogućih vrijednosti – dakle, **viševrijednosnu diskretnu varijabla**.

Ovu distribuciju smo već sreli kada smo pričali o mutlinomijalnoj logističkoj regresiji, gdje smo njome modelirali varijablu oznake, koja je mogla poprimiti vrijednost jedne od K klasa.

Kategoričku varijablu modeliramo kao binaran vektor **indikatorskih varijabli**:

$$\mathbf{x} = (x_1, x_2, \ldots, x_K)$$

Binarnih varijabli ima onoliko koliko kategorička varijabla ima mogućih različitih vrijednosti. Samo jedna binarna varijabla će imati vrijednost jedan, a sve ostale su nula. To se zove **vektor 1-od-K** ili **one-hot encoding**. Npr., trovrijednosna kategorička varijabla koja je poprimila drugu od tri vrijednosti bila bi predstavljena kao $\mathbf{x} = (0, 1, 0)$.

Vjerojatnost da kategorička varijabla poprimi neku vrijednost je općenito različita za svaku vrijednost. Drugim riječima, svaka binarna varijabla iz vektora indikatorskih varijabli ima svoju vjerojatnost da bude jednaka jedinici. To znači da imamo po jedan parametar μ za svaku binarnu varijablu. Te parametre možemo strpati u vektor parametara $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$$

pri čemu mora vrijediti $\sum_k \mu_k = 1$ i $\mu_k \geqslant 0$.

Kako sada napisati vjerojatnost da kategorička slučajna varijabla poprimi neku svoju konkretnu vrijednost? To smo već vidjeli kada smo pričali o multinomijalnoj regresiji: ideja je da se matematički napiše primjena binarnog vektora $\mathbf{x}$ kao **maske** nad vektorom $\boldsymbol{\mu}$, tako da izabremo μ_k koji odgovara vrijednosti x_k koja je u vektoru $\mathbf{x}$ postavljena na jedinicu: [10]

$$P(X = \mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

▸ **PRIMJER**

Četverovrijednosna kategorička varijabla $\mathbf{x}$ neka je poprimila treću vrijednost, tj.

$$\mathbf{x} = x_3 \quad \Rightarrow \quad \mathbf{x} = (0, 0, 1, 0)$$

Vjerojatnosti pojedinačnih vrijednosti neka su:

$$\boldsymbol{\mu} = (0.2, 0.3, 0.4, 0.1)$$

Onda je vjerojatnost da je varijabla poprimila treću vrijednost jednaka:

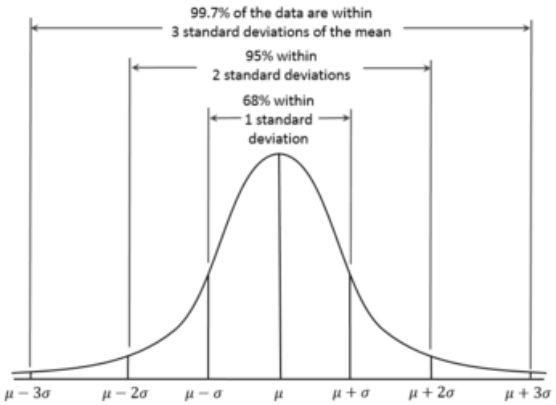
$$P\big(\mathbf{x} = (0, 0, 1, 0)\big) = \prod_{k=1}^{4} \mu_k^{x_k} = 1 \cdot 1 \cdot \mu_3 \cdot 1 = \mu_3 = 0.4$$

3.3 Gaussova distribucija

Gaussova (normalna) distribucija kontinuirana je distribucija koja se tipično koristi se za modeliranje kontinuirane varijable. Zašto? Prvo, zato što se pokazuje da mnogi prirodni fenomeni slijede Gaussovu distribuciju. Drugi razlog je što je Gaussova distribucija matematički elegantna i jednostavna. [11]

Univarijatna (jednodimenzijska) Gaussova distribucija definirana je sljedećom funkcijom gustoće vjerojatnosti (koju smo već vidjeli):

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Iz definicije Gaussove distribucije može se za očekivanje i varijancu slučajne varijable koja se pokorava toj distribuciji izvesti $\mathbb{E}[X] = \mu$ i $\mathrm{Var}(X) = \sigma^2$. Dakle, parametri distribucije μ i σ^2 direktno definiraju očekivanje i varijancu slučajne varijable koja se ravna po Gaussovoj distribuciji.

U strojnom učenju Gaussovu distribuciju koristit ćemo za modeliranje kontinuiranih varijabli uz prisustvo **šuma**. Tu je ideja sljedeća: očekivana vrijednost varijable zapravo je μ, ali zbog šuma dolazi do rasipanja, pa mi u podatcima opažamo neku malo drugačiju vrijednost. Što je šum veći, to je veće rasipanje, odnosno veća je varijanca σ^2.

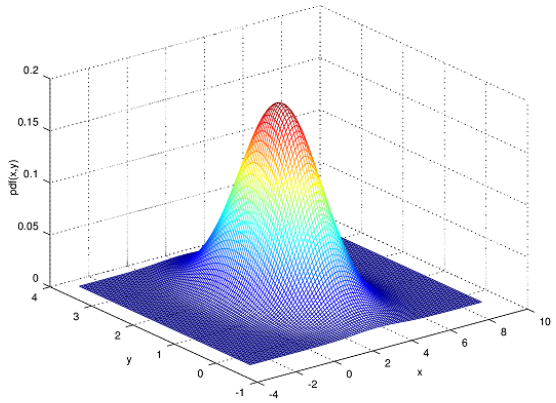
3.4 Multivarijatna Gaussova distribucija

U strojnom učenju često je primjer prikazan kao vektor realnih brojeva, npr., prosjek ocjena kroz četiri razreda srednje škole. Takav primjer modelirat ćemo **multivarijatnom (višedimenzijskom) Gaussovom distribucijom**. Gustoća vjerojatnosti multivarijatne distribucije definirana je ovako:

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

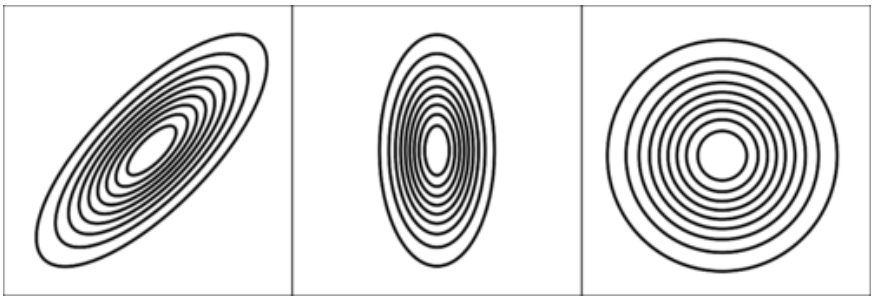
gdje su $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ srednja vrijednost odnosno kovarijacijska matrica, i oni čine parametri multivarijatne Gaussove distribucije.

U dvije dimenzije, imamo dvodimenzijsku odnosno **bivarijatnu** Gaussovu distribuciju, koja izgleda ovako:



Matrica $\mathbf{\Sigma}$ je već spomenuta **kovarijacijska matrica**. Vidimo da nam za izračun gustoće vjerojatnosti treba inverz i determinanta kovarijacijske matrice, te da determinanta ne smije biti nula, jer inače imamo dijeljeje s nulom. Kovarijacijska matrica će imati inverz i determinantu koja je veća od nule samo ako je matrica **pozitivno definitna**. A to, već smo rekli, znači da nema redundantnih značajki i da nema konstantnih značajki (koje su ionako beskorisne). Sjetite se također naših razmatranja u kontekstu invertiranja matrice dizajna kod linearne regresije: čak i ako značajke nisu savršeno multikolinearne, imat ćemo problema s izračunom inverza jer će matrica imati **visok kondicijski broj** i rješenje će biti nestabilno. Vrijede iste napomene kao i ranije: treba eliminirati redundantne značajke.

Eksponent koji se javlja u izrazu za Gaussovu gustoću vjerojatnosti je tzv. **kvadratna forma**: $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$. Korijen toga je **Mahalanobisova udaljenost** između $\mathbf{x}$ i $\boldsymbol{\mu}$, o kojoj smo već bili pričali kada smo pričali o jezgrenim funkcijama. Udaljenost između primjera bit će, dakle, definirana u ovisnosti o matrici kovarijacije $\mathbf{\Sigma}$. Isto tako, matrica $\mathbf{\Sigma}$ će određivati kako izgleda Gaussova gustoća vjerojatnosti. Razmotrimo kroz tri konkretna slučaja kako $\mathbf{\Sigma}$ utječe na oblik bivarijatne Gaussove gustoće vjerojatnosti:



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

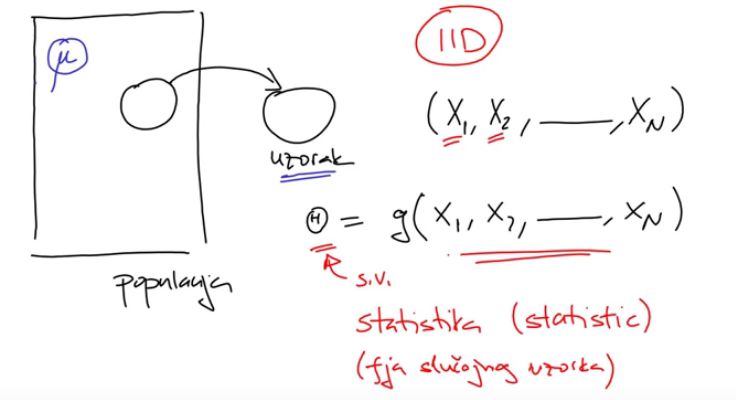
Prva slika pokazuje slučaj kada s porastom vrijednosti varijale X imamo i porast vrijednosti varijable Y, tj. varijable X i Y imaju pozitivnu kovarijaciju (također i: pozitivnu korelaciju). Zbog toga su izokonture Gaussove gustoće vjerojatnosti pozitivno zakošene elipse: gustoća vjerojatnosti za Y raste/pada kako raste/pada gustoća vjerojatnosti za X. Kada bi kovarijacija između X i Y bila negativna, elipse bi bile zakošene u drugu stranu. Taj slučaj, dakle, odgovara situaciji kada imamo punu matricu $\mathbf{\Sigma}$, tj. matricu koja nema nule izvan glavne dijagonale, što znači da postoje kovarijacije (odnosno korelacije) između varijabli. Srednja slika odgovara situaciji kada je kovarijacijska matrica dijagonalna: na dijagonali su varijance, koje su općenito različite za svaku značajku, a izvan dijagonale su nule, tj. nema kovarijacije. Budući da nema kovarijacije, izokonture su elipse koje nisu zakošene, već su poravnata s osima koordinatnog sustava. No, budući da varijance za svaku varijablu nisu identične, imamo elipse a ne kružnice. U ovom konkretnom slučaju varijanca za Y je veća od varijance za X, pa je elipsa vertikalno izdužena. Kada bi varijanca za X bila veća od varijance za Y, elipsa bi bila horizontalno izdužena. Konačno, na desnoj slici prikazan je slučaj kada je kovariacijska matrica izotropna: dijagonalna matrica sa jednakovrijednim varijancama na dijagonali. Izokonture gustoće vjerojatnosti sada odgovaraju kružnicama.

Sada smo se upoznali – ili prisjetili, kako tko – osnovnih vjerojatnosnih distribucija. Vidimo da svaka distribucija ima neke svoje **parametre**: npr., Bernoullijeva distribucija ima samo jedan parametar (μ), dok multivarijatna Gaussova distribucija ima dva parametra (kovarijacijsku matricu $\mathbf{\Sigma}$ i vektor srednje vrijednosti $\boldsymbol{\mu}$).

Vratimo se sada na osnovno današnje pitanje: ako pretpostavimo da se podatci pokoravaju nekoj teorijskoj distribuciji, kako izračunati njezine parametre na temelju tih podataka? To nas vodi do **procjene parametara**.

4 Procjena parametara

Iz perspektive statistike, podatci koje imamo na raspolaganju jesu **uzorak** iz neke **populacije**. Populacija su svi podatci (kojih može biti konačno ili beskonačno mnogo), a uzorak je samo jedan konačan **podskup**. Taj uzorak treba biti **slučajan**, jer slučajan uzorak (ako je dovoljno velik) je **reprezentativan uzorak**. To onda znači da su podatci u tom uzorku očekivano distribuirani kao i podatci u cijeloj populaciji. Ideja je onda da na temelju tog slučajnog uzorka napravimo **procjenu (estimaciju) parametra** modela koji objašnjava cijelu **populaciju**. Shematski to možemo prikazati ovako:



Što je uzorak? Uzorak je niz slučajnih varijabli: $(X_1, X_2, \ldots, X_N)$ – N-torka. Naša pretpostavka su varijable koje čine uzorak **i.i.d. – nezavisno i identično distribuirane** (engl. *independently and identically distributed*). To znači da su slučajne varijable međusobno nezavisne (ishod jedne ne utječe na ishod druge) i da dolaze iz iste distribucije (budući da je cijeli uzorak iz iste populacije). Ova je pretpostavka centralna za mnoge algoritme strojnog učenja, i za sve algoritme koje proučavamo na ovom predmetu.

Na temelju takvog uzorka možemo onda izračunati različite vrijednosti. Označimo to ovako:

$$\Theta = g(X_1, X_2, \ldots, X_N)$$

g je dakle neka funkcija koja izračunava nešto na temelju uzorka. Npr., g bi mogla biti funkcija sume, pa bi funkcija računala zbroj vrijednosti u uzorku. Rezultat funkcije g je Θ. Budući da funkcija g radi na slučajnom uzorku, tj. na vektoru slučajnih varijabli, to će rezultat Θ također biti u određenoj mjeri slučajan, tj. Θ je i sama **slučajna varijabla**. Preciznije, Θ je slučajna varijabla čija je vrijednost izračunata na temelju slučajnog uzorka. Takva slučajna varijabla naziva se **statistika** (engl. *statistic*). Dakle, statistika je nekakva **funkcija slučajnog uzorka**. Od beskonačno mnogo različitih statistika koje možemo izmisliti, nas zanimaju one koje odgovaraju nekom **parametru θ** našeg pretpostavljenog modela, koji predstavlja podatake (tj. populaciju). Vrijednost tog parametra nam je nepoznata i nju želimo izračunati, da bismo mogi raditi predikciju. Takvu statistiku, koja nam daje vrijednost parametra populacije, zovemo **procjenitelj (estimator)** parametra θ. Vrijednost procjenitelja $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ naziva se **procjena** (uočite "šešir" na θ). Parametar θ može biti konkretno parametar populacije (npr., srednja vrijednost μ) ili neki drugi parametar koji imamo u našem modelu (npr., vektor težina $\mathbf{w}$), a koji upravlja izgledom distribucije podataka.

12

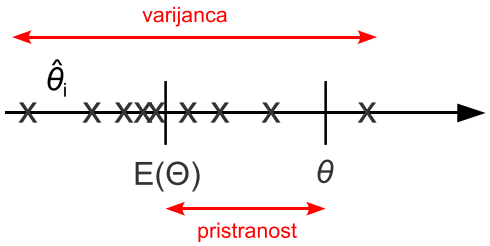
4.1 Pristranost procjenitelja

Budući da je procjenitelj slučajna varijable, on i sam, kao i svaka slučajna varijabla, ima svoje očekivanje i varijancu. Razlika između očekivane vrijednosti procjenitelja i parametra populacije koji želimo procijeniti tim procjeniteljem naziva se **pristranost** (engl. *bias*). Formalno,

pristranost procjenitelja Θ kao procjenitelja parametra θ jednaka je:

$$b_\theta(\Theta) = \mathbb{E}[\Theta] - \theta$$

Grafički to možemo prikazati ovako:



Prava vrijednost parametra populacije je θ, a križići su procjene, odnosno konkretne vrijednosti $\hat{\theta}_i$ procjenitelja Θ na različitim uzorcima iz te populacije. Dakle, to su vrijednednosti procjene koje bismo dobivali kada bismo ponavljali izvlačenje uzorka (to je nešto što u praksi ne radimo; u praksi imamo samo jedan uzorak). Očekivano, budući da su uzorci slučajni, dobili bismo svaki puta malo drugačiju procjenu. Srednja vrijednost procjena $\hat{\theta}_i$ jednaka je očekivanju procjenitelja, $\mathbb{E}[\Theta]$ (zapravo, srednja vrijednost procjena je i sama procjena za očekivanje procjenitelja, ali to sad nije bitno). Odstupanje očekivanja procjenitelja $\mathbb{E}[\Theta]$ od prave vrijednosti parametra populacije θ je pristranost procjenitelja Θ.

Procjenitelja parametara populacije ima dobrih i manje dobrih. Željeli bismo procjenitelj koji je što točniji, tj. procjenitelj čija je greška što manja. To često znači da želimo da je procjenitelj čija je pristranost jednaka nuli. Kažemo da je procjenitelj Θ **nepristran procjenitelj** (engl. *unbiased estimator*) parametra θ ako i samo ako: 13

$$\mathbb{E}[\Theta] = \theta$$

tj. ako je pristranost jednaka nuli. (NB: Pristranost procjenitelja nije ista kao i induktivna pristranost algoritma strojnog učenja, premda je povezana s njom.)

▸ **PRIMJER**

Neka je X slučajna varijabla sa vrijednostima iz $x \in \mathbb{R}$. Označimo $\mathbb{E}[X] = \mu$ (srednja vrijednost) i $\mathrm{Var}(X) = \sigma^2$ (varijanca) te varijable. Zanimaju nas parametri μ i σ^2 populacije. Ti parametri su nam nepoznati.

Parametre μ i σ^2 možemo procijeniti na temelju uzorka $\{x^{(i)}\}_{i=1}^N$ pomoću **procjenitelja**. Koje procjenitelje upotrijebiti za ove parametre? Mogli bismo probati s ovima:

$$\hat{\mu} = \frac{1}{N}\sum_i x^{(i)} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^N (x^{(i)} - \hat{\mu})^2$$

Naravno, sad se pitamo jesu li ovo dobri procjenitelji za naše parametre. Drugim riječima, pitamo se je li $\hat{\mu}$ nepristran procjenitelj za μ i je li $\hat{\sigma}^2$ nepristran procjenitelj za σ^2, tj. je li:

$$\mathbb{E}[\hat{\mu}] = \mu$$
$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$$

Može se pokazati, ali to ćemo ostaviti za domaću zadaću, da $\mathbb{E}[\hat{\mu}] = \mu$, tj. $\hat{\mu}$ jest **nepristran** procjenitelj srednje vrijednosti populacije. Međutim, isto se tako može pokazati da $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$, tj. $\hat{\sigma}^2$ **nije nepristran** procjenitelj varijance!

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$$

Vidimo da očekivana vrijednost procjenitelja nije jednaka parametru populacije σ^2, pa je dakle ovaj procjenitelj pristran. Možemo izračunati njegovu pristranost:

$$b(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{N}$$

Budući da je pristranost negativna, to znači da procjenitelj **podcjenjuje** (engl. *underestimates*) pravu varijancu! Primjenom ovog procjenitelja, dobit ćemo varijancu koja je manja od prave varijance populacije.

Možemo li to ispraviti, tj. učiniti procjenitelj nepristranim? Naravno da možemo! Sve što trebamo napraviti jest korigirati vrijednost koju nam daje procjenitelj tako da pristranost svedemo na nulu. Iz izraza za pristranost vidimo da će ona biti nula ako pomnožimo vrijednost procjenitelja sa $\frac{N}{N-1}$. Tako dobivamo **nepristran procjenitelj varijance**:

$$\hat{\sigma}^2_{\text{nepr.}} = \frac{N}{N-1}\hat{\sigma} = \frac{N}{N-1}\frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \hat{\mu})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)} - \hat{\mu})^2$$

Ova korekcija se treba raditi kada je N (uzorak) malen. Ako je N velik, onda je gotovo svejedno dijelimo li sa N ili $N-1$. Ili, drugim riječima, ako $N \to \infty$, onda pristranost $-\frac{\sigma^2}{N} \to 0$.

4.2 Metode za izvođenje procjenitelja

Sada znamo što je procjenitelj i što je nepristrani procjenitelj. Međutim, pitanje je odakle nam uopće formula za procjenitelj? U gornjim primjerima ptičica nam je došapnula formule za procjenitelj srednje vrijednosti i formulu za procjenitelja varijance, koji su se pokazali nepristranim, odnosno pristranim ali ispravljivim. Kako općenito možemo doći do toga? Kako izvesti procjenitelje za bilo koji parametar vjerojatnosne distribucije?

Srećom, u statistici su razvijene općeniti postupci za izvođenje procjenitelja. Konkretno, postoje tri glavne vrste procjenitelja:

- **Procjenitelj najveće izglednosti** (engl. *Maximum Likelihood Estimator; MLE*);
- **Procjenitelj maximum aposteriori (MAP)**;
- **Bayesovski procjenitelj** (engl. *Bayesian estimator*), koji zapravo motivira jedan još općenitiji pristup, naime bayesovsku statistiku.

Mi ćemo na ovom predmetu raditi prva dva procjenitelja: MLE i MAP, dok Bayesovski procjenitelj (i cijelu Bayesovsku statistiku) ostavljamo za bolje dane. Idući sat, dakle, nastavljamo sa MLE i MAP procjeniteljima.

Sažetak

- Učenje probalističkih modela svodi se na **procjenu parametara** distribucije
- Primjere i oznake modeliramo kao **slučajne varijable**
- Slučajne varijable imaju **očekivanje**, **varijancu** i **kovarijancu**
- **Nezavisne** slučajne varijable su **nekorelirane**, ali obrat ne vrijedi
- Koristimo teorijske distribucije: Bernoullijevu, Multinulijevu, Gaussovu
- **Statistika** je funkcija slučajnog uzorka, a **procjenitelj** je statistika koja odgovara parametru distribucije koji nas zanima
- Svaki procjenitelj ima svoju **pristranost** i **varijancu**

Bilješke

[1] Prepostavka je, naravno, da suvereno vladate osnovama teorije vjerojatnosti. Premda će nam u ovom predmetu zapravo trebati vrlo malo toga iz teorije vjerojatnosti, dobro je osvijestiti činjenicu da strojno učenje, jednako kao i statistika, ima temelje u teoriji vjerojatnosti. Mnogo je dobrih udžbenika o teoriji vjerojatnosti, međutim ako želite dobiti širu sliku te naučiti nešto o fundamentalnim aspektima te teorije te vezama s drugim znanstvenim disciplinama, posebice logikom, toplo preporučam da pogledate Jaynesov "Probability Theory" (Jaynes, 2003).
Usput, Edwin Thompson Jaynes bio je fizičar koji se prvenstveno bavio statističkom mehanikom, ali i temeljima teorije vjerojatnosti i statističkog zaključivanja. Također je poznat po konceptu *pogreške projekcije uma* (engl. *mind projection fallacy*), do koje dolazi kada na stvarnost projiciramo neke karakteristike koje nisu dijelom te stvarnosti. Prema Jaynesu, i teorija vjerojatnosti može dovesti do te pogreške, jer vjerojatnost nije inherentno objektivnoj stvarnosti već manifestacija neizvjesnosti koja proizlazi iz našeg neznanja o toj stvarnosti. Prema Jaynesu, način da se takve pogreške izbjegnu jest usvajanje epistemičke skromnosti (engl. *epistemic humility*).

[2] Mnogo detaljniju (i mnogo bolju) ekspoziciju ove teme možete naći u (Elezovic, 2010).

[3] Budući da ćemo, uglavnom, umjesto $P(X = x)$ pisati $P(x)$, to znači da u formulama nećemo razlikovati između slučajne varijable i njezine vrijednosti. Tako će, na primjer, '$\mathbf{x}$' označavati i slučajnu varijablu primjera (dakle, bilo koji primjer) i vrijednost (tj. realizaciju) te slučajne varijable (dakle, jedan konkretan primjer). U kontekstima gdje je ta distinkcija bitna, vratit ćemo se na notaciju koja razlikuje slučajnu varijablu od njezine vrijednosti (na primjer, pisat ćemo $P(\mathbf{x} = \mathbf{x}^{(i)})$).

[4] U matematici je **funkcija gustoće vjerojatnosti** tipično označena sa f. Mi ćemo koristiti p, jer je to uobičajeno u literaturi iz strojnog učenja. Dakle, veliko P ćemo koristiti za vjerojatnost, a malo p za gustoću vjerojatnosti. Međutim, često ćemo si pojednostaviti život i u oba slučaja govoriti o "vjerojatnosti" ili "distribuciji" (što nije isto, ali iz konteksta će biti jasno mislimo li na vjerojatnost, gustoću vjerojatnosti, ili distribuciju koje one definiraju).

[5] Prisjetite se da je **vjerojatnost** da kontinuirana slučajna varijabla X poprimi vrijednost iz intervala $[a, b]$ $(a, b \in \mathbb{R},\ a \leqslant b)$ jednaka:

$$P(a \leqslant X \leqslant b) = \int_a^b p(x)\,\mathrm{d}x\,.$$

Primijetite da za kontinuiranu varijablu X vrijedi $P(X = a) = 0$, tj. vjerojatnost da kontinuirana varijabla poprimi bilo koju pojedinačnu vrijednost jednaka je nuli.

[6] **Marginalna vjerojatnost** je vjerojatnost pojedinačne varijable koju dobivamo iz zajedničke vjerojatnosti više varijabli. Marginalnu vjerojatnost dobivamo **marginalizacijom**. Kod diskretnih varijabli marginalizacija se svodi na zbrajanje vrijednosti zajedničke vjerojatnosti po svim preostalim varijablama. Npr., marginalna vjerojatnost $P(X)$ iz zajedničke se vjerojatnosti $P(X, Y)$ dobiva kao:

$$P(X) = \sum_Y P(X, Y)$$

Više o tome u predavanju broj 15, kada ćemo pričati o Bayesovom klasifikatoru.

[7] Preuzeto sa https://en.wikipedia.org/wiki/Correlation_and_dependence.
Primijetite da grafikoni raspršenja prikazuju konkretne vrijednosti varijabli X i Y, tj. prikazujemo njihove realizacije u uzorku (svaka realizacija je jedna točka u grafikonu raspršenja). Međutim, očekivanje, varijanca i kovarijanca (posljedično, i korelacija) definirane su nad slučajnim varijablama (tj. nad njihovim distribucijama), a ne nad njihovim konkretnim realizacijama (tj. nad uzorkom). Međutim, u strojnom učenju mi raspolažemo uzorkom (skupom podataka), a ne distribucijama. Morat ćemo, dakle, nekako od uzorka doći do distribucije. Upravo to radimo pomoću **procjenu parametara**. Ideja je, dakle, da iz podataka kojima raspolažemo procijenimo očekivanje varijancu, korelaciju, kovarijaciju i ostale nama važne veličine koje opisuju varijablu od interesa.

[8] Ovdje ste možda uočili da je Pearsonov koeficijent korelacije nekako povezan s pogreškom modela linearne regresije, budući da oba modela pretpostavljaju linearnu ovisnost varijabli. To je točno.

Naime, za model jednostavne regresije, $h(x) = w_0 + w_1 x$, kvadrat Pearsonovog koeficijenta korelacije procijenjenog iz podataka, r^2, između zavisne varijable y i nezavisne varijable (značajke) x jednak je **koeficijentu determinacije** R^2. Slično kao i pogreška kvadratnog odstupanja, koeficijent determinacije mjeri koliko regresijski pravac (odnosno općenito hiperravnina) dobro modelira podatke, ali, za razliku od kvadratnog odstupanja, koeficijent determinacije je (na skupu za učenje) normaliziran na interval $[0, 1]$. Veći R^2 odgovara većem r^2, tj. sve većoj linearnoj zavisnosti između varijabli X i Y.

[9] Primijetite da sve ove distribucije koje ovdje spominjemp pripadaju **eksponencijalnoj familiji distribucija**, koju smo već bili spomenuli u kontekstu poopćenih linearnih modela. To znači da sve ove distribucije imaju neka zajednička svojstva, i da bismo ih mogli promatrati unificirano kroz ta svojstva (i isto tako izvesti unificirane procjenitelje za parametre tih distribucija). Međutim, to je malo kompliciranije, pa nećemo to raditi, nego ćemo svaku distribuciju promatrati zasebno.

[10] Legitimno pitanje ovdje jest zašto umnožak $\prod \mu_k^{x_k}$ a ne zbroj $\sum x_k \mu_k$? Zato što ćemo kasnije željeti raditi s logaritmima vjerojatnosti, a onda želimo imati umnožak, koji će se logaritmiranjem pretvoriti upavo u ovu sumu.

[11] Ovdje ponavljam napomenu iz skriptice 3, da pogledate zanimljivu diskusiju o sveprisutnosti i opravdanosti normalne distribucije na https://stats.stackexchange.com/q/204471/93766.

[12] U engleskom jeziku kaže se **statistic** (jednina), dok je **statistics** (množina) naziv za matematičku disciplinu. U hrvatskome jeziku nemamo tu razliku!

[13] Ovdje gadno pojednostavljujemo stvari. Pristranost je poželjno svojstvo procjenitelja, ali stvari nisu tako jednostavne. Kvaliteta procjenitelja sastoji se od mnogo komponenti, jedna od kojih je pogreška procjenitelja. Ta se može rastaviti na kvadrat pristranosti i varijancu. U pravilu, želimo da je pogreška procjenitelja što manja, ali to onda podrazumijeva **kompromis između pristranosti i varijance** (engl. *bias-variance tradeoff*). Ako je procjenitelj nepristran i k tome još ima najmanju moguću varijancu, onda govorimo o **nepristranom procjenitelju najmanje varijance** (engl. *minimum variance unbiased estimator, MVUE*). Pritom je "najmanja moguća varijanca" definirana Cramér-Raovom ogradom. Premda je ovo zapravo vrlo fundamentalno i važno za strojno učenje, pogotovo ideja rastava na pristranost i varijancu, mi ipak u to nećemo ulaziti. Zainteresirane upućujem na https://en.wikipedia.org/wiki/Estimator. Ovo je također dobar tekst: https://stats.stackexchange.com/a/207764/93766.

Literatura

N. Elezovic. Vjerojatnost i statistika. *Slucajne varijable, Element, Zagreb*, 2010.

E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.