SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1728

Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Ivan Grubišić

Zagreb, lipanj 2018.

Umjesto ove stranice umetnite izvornik Vašeg rada.

Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Procjena nesigurnosti predikcija vrlo je važan sastojak mnogih praktičnih primjena konvolucijskih modela računalnog vida. Do tog cilja možemo doći analizom višeznačnosti podataka, nesigurnosti odluke modela te vjerojatnosti da se podatak nalazi u distribuciji skupa za učenje. U ovom radu razmatramo pristupe koji procjenu nesigurnosti predikcija uče nadzirano, primjenom istih podataka na kojima se uči i promatrani model.

U okviru rada, potrebno je proučiti i ukratko opisati postojeće pristupe za procjenu nesigurnosti predikcija. Uhodati postupke procjene nesigurnosti dubokih konvolucijskih modela temeljene na nadziranom učenju. Validirati hiperparametre te prikazati i ocijeniti ostvarene rezultate na problemu semantičke segmentacije. Predložiti pravce budućeg razvoja. Radu priložiti izvorni i izvršni kod razvijenih postupaka, ispitne slijedove i rezultate, uz potrebna objašnjenja i dokumentaciju. Citirati korištenu literaturu i navesti dobivenu pomoć.

zahvala

Sadržaj

Oznake

Objekti

Varijable se označavaju kosim slovima sa serifima, većina konstanti uspravnim slovima sa serifima, a slučajne varijable kosim slovima bez serifa. Vektori se označavaju malim podebljanim slovima, matrice i višedimenzionalni nizovi (tenzori) velikim podebljanim slovima, a skupovi slovima s udvostručenim linijama. Za svaku vrstu objekta mogu se koristiti i latinska i grčka slova.

a, A, θ	Varijabla (najčešće skalar)
$\boldsymbol{a}$, $\boldsymbol{\theta}$	Vektor ili niz (najčešće vektor stupac)
$\boldsymbol{A}$, $\boldsymbol{\Theta}$	Matrica ili višedimenzionalni niz
$\mathbb{A}$	Skup ili multiskup
a, A, $\grave{}$	Konstanta
$\mathbf{a}$, $\grave{}$	Konstanta vektor ili niz
$\mathbf{A}$, $\mathbf{\Theta}$	Konstanta matrica ili višedimenzionalni niz
$\mathbb{A}$, $\nleq$	Kostanta skup
a, A, $\mathsf{\theta}$	Slučajna varijabla
$\mathsf{\boldsymbol{a}}$, $\mathsf{\boldsymbol{\theta}}$	Slučajni vektor ili niz
$\mathsf{\boldsymbol{A}}$, $\mathsf{\boldsymbol{\Theta}}$	Slučajna matrica ili višedimenzionalni niz
$\mathbb{A}$	Slučajni skup ili multiskup
a, oznaka	Oznaka koja ne predstavlja matematički objekt

Konstante

$\{\}$	Prazni skup
$\mathbf{0}$	Nul-vektor
$\mathbf{e}_i$	i-ti vektor kanonske baze
$\mathbf{1}$	Zbroj svih vektora kanonske baze
$\mathbf{I}$, $\mathbf{I}_n$	Matrica identiteta (s n redaka i stupaca)
$\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{C}$	Poznati skup
$\mathbb{R}_{\geq 0}$, $\mathbb{R}_{>0}$	Skup nenegativnih/pozitivnih realnih brojeva

Definiranje skupova i nizova

$a..b$	Kraći zapis za $a, .., b$
$\{a..b\}$	Skup cijelih brojeva od a do b

$\{f(a)\colon P(a)\}$, $\{f(a)\}_{P(a)}$	Skup čiji su elementi definirani preko funkcije f i predikata P
$\{f(a)\}_a$	Skup čiji su elementi definirani preko funkcije f i varijabli a iz implicitno određenog skupa
$\{a_1..a_n\}$, $\{a_i\}_{i=1..n}$	Skup s n elemenata
$(a_i)_i$, $\left(a_{i,j}\right)_{i,j}$, $\left(a_{i,j,k}\right)_{i,j,k}$	Višedimenzionalni niz s implicitnim ili neodređenim brojem elemenata
(a, b)	Otvoreni interval
$[a, b]$	Zatvoreni interval
$[x_1, .., x_n]$	Vektor redak
$(x_1, .., x_n)$	Vektor (stupac)

Donji i gornji indeks

U donjem indeksu oznake mogu biti oznake drugih matematičkih objekata. U donjem i gornjem indeksu oznake mogu biti oznake (slova ili riječi) koje ne predstavljaju matematičke objekte. Oznake koje ne predstavljaju matematičke objekte istog su stila kao tekst. Indeksi (redni brojevi) elemenata vektora ili višedimenzionalnih nizova se, ako nije određeno drugačije, pišu u donjem indeksu oznake vektora u uglatim zagradama. Npr. ako je definiran vektor $\boldsymbol{a} = (a_1, .., a_n)^\mathsf{T}$, onda je njegov i-ti element $\boldsymbol{a}_{[i]} = a_i$.

$a_{\mathsf{d}}^{\mathsf{g}}$	Oznaka varijable s oznakama u donjem i gornjem indeksu
$\boldsymbol{a}_{[i]}$	i-ti element vektora $\boldsymbol{a}$
$\boldsymbol{a}_{[i_1:i_2]}$	Vektor kojeg čine elementi $\boldsymbol{a}_{[i_1]}, \boldsymbol{a}_{[i_1+1]}, .., \boldsymbol{a}_{[i_2]}$
$\boldsymbol{a}_{\left[(i_1..i_n)\right]}$	Vektor kojeg čine elementi $\boldsymbol{a}_{[i_1]}, \boldsymbol{a}_{[i_2]}, .., \boldsymbol{a}_{[i_n]}$
$\boldsymbol{A}_{[i,j]}$	Element i, j matrice $\boldsymbol{A}$
$\boldsymbol{A}_{[i,:]}$	i-ti redak matrice $\boldsymbol{A}$
$\boldsymbol{A}_{[:,i_1:i_2,j]}$	2-D odsječak 3-D niza $\boldsymbol{A}$

Operacije linearne algebre i druge operacije s nizovima

$\langle \boldsymbol{a} \vert \boldsymbol{b} \rangle$	Skalarni produkt, može biti i $\boldsymbol{a}^\mathsf{T}\boldsymbol{b}$
$\boldsymbol{a} \odot \boldsymbol{b}$	Umnožak po elementima; Hadamardov produkt
$\boldsymbol{a} \oslash \boldsymbol{b}$	Dijeljenje po elementima
$\boldsymbol{AB}$	Matrično množenje

$\boldsymbol{A}^{-1}$	Inverz matrice
$\boldsymbol{A}^{\mathsf{T}}$	Transponiranje
$\operatorname{diag}(\boldsymbol{a})$	Dijagonalna matrica kojoj dijagonalu čini vektor $\boldsymbol{a}$
$\det \boldsymbol{A}$	Determinanta matrice $\boldsymbol{A}$
$\|\boldsymbol{a}\|_2$	L^2-norma vektora $\boldsymbol{a}$
$\|\boldsymbol{a}\|_p$	L^p-norma vektora $\boldsymbol{a}$
$\|\boldsymbol{A}\|_p$	Matrična L^p-norma matrice $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{\mathsf{F}}$	Frobeniusova norma matrice $\boldsymbol{A}$
$f(\boldsymbol{a})$	Ako f nije drugačije definirana i inače označava funkciju $\mathbb{R} \to \mathbb{R}$, onda se primjenjuje po svakom elementu vektora posebno
$\boldsymbol{a} \mathbin{+\!\!+} \boldsymbol{b}$	Konkatenacija vektora (stupaca) $\boldsymbol{a} \in \mathbb{R}^n$ i $\boldsymbol{b} \in \mathbb{R}^m$ u vektor iz $\mathbb{R}^{n+m}$
$\boldsymbol{A} \mathbin{+\!\!+} \boldsymbol{B}$	Konkatenacija višedimenzionalnih nizova po prvoj dimenziji
$\boldsymbol{A} \mathbin{+\!\!+'} \boldsymbol{B}$	Konkatenacija višedimenzionalnih nizova po zadnjoj dimenziji

Diferencijalni račun

$\frac{\mathrm{d}y}{\mathrm{d}x}$, $\frac{\mathrm{d}}{\mathrm{d}x} f(x)$	Derivacija $y = f(x)$ po x
$\frac{\partial y}{\partial x}$, $\frac{\partial}{\partial x} f(x)$	Parcijalna derivacija $y = f(x)$ po x
$\nabla_{\boldsymbol{x}} y$, $\nabla_{\boldsymbol{x}} f(x)$, $\left(\frac{\partial y}{\partial \boldsymbol{x}}\right)^{\mathsf{T}}$	Gradijent $y = f(\boldsymbol{x})$ po $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$, $\nabla_{\boldsymbol{X}} f(x)$	Gradijent $y = f(\boldsymbol{x})$ po $\boldsymbol{X}$
$\frac{\partial^2 y}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathsf{T}}}$, $\boldsymbol{H}_f(\boldsymbol{x})$, $\boldsymbol{H}$	Hessijan iz $\mathbb{R}^{n \times n}$ za $f \colon \mathbb{R}^n \to \mathbb{R}$ i $y = f(\boldsymbol{x})$
$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$, $\boldsymbol{J}_f(\boldsymbol{x})$, $\boldsymbol{J}$	Jakobijeva matrica iz $\mathbb{R}^{m \times n}$ za $f \colon \mathbb{R}^n \to \mathbb{R}^m$ i $\boldsymbol{y} = f(\boldsymbol{x})$
$\int_A f(x) \mathrm{d}x$, $\int_{x \in A} f(x)$	Određeni integral funkcije $f(x)$ po $x \in A$
$\int f(x) \mathrm{d}x$, $\int_x f(x)$	Određeni integral funkcije $f(x)$ po $x \in A$, gdje je A implicitan

Teorija vjerojatnosti

Svakoj slučajnoj varijabli a jednoznačno je dodijeljena jedna razdioba $\mathrm{p}(a)$ (ili $\mathrm{P}(a)$) i funkcija gustoće vjerojatnosti (koja može biti poopćena funkcija) $p_a(a) = \mathrm{p}(a = a)$. $\mathrm{P}(A)$ označava vjerojatnost događaja A, a P_a funkciju vjerojatnosti slučajne varijable a. Gustoća vjerojatnosti se još kraće može zapisati $\mathrm{p}(a)$, gdje se po slovu implicitno pretpostavlja slučajna varijabla označena istim slovom bez serifa. Isto tako, vjerojatnost elementarnog događaja se može zapisati $\mathrm{P}(a)$. Mogu se koristiti i druge oznake za funkciju vjerojatnosti ili funkciju gustoće vjerojatnosti.

$(a \mid b = b)$, $(a \mid b)$	Uvjetna slučajna varijabla
(a, b)	Združena slučajna varijabla
$a \perp b$	*Slučajne varijable a i b su nezavisne*
$a \not\perp b$	*Slučajne varijable a i b su zavisne*
$a \perp b \mid c$	*Slučajne varijable a i b su uvjetno nezavisne uz poznat ishod slučajne varijable c*
$a \not\perp b \mid c$	*Slučajne varijable a i b su uvjetno zavisne uz poznat ishod slučajne varijable c*
p, q	Razdioba ili funkcija gustoće vjerojatnosti
$\mathbb{A}$	Događaj
$\{R(a)\}$	Događaj definiran predikatorm slučajne varijable a
$\mathrm{P}(\{R(a)\})$, $\mathrm{P}(R(a))$	Vjerojatnost događaja $\{R(a)\}$
$\mathrm{P}(a)$, $\mathrm{p}(a)$, $\mathcal{D}$	Razdioba slučajne varijable a; P ako je a diskretna slučajna varijabla, a p ako nije ili ako se ne zna
$\mathrm{P}(a = a)$, $P_a(a)$, $\mathrm{P}(a)$	Vjerojatnost događaja $\{a = a\}$
$\mathrm{p}(a = a)$, $p_a(a)$, $\mathrm{p}(a)$	Gustoća vjerojatnosti događaja $\{a = a\}$
$p_{a\mid b}(a)$, $\mathrm{p}(a \mid b)$	Gustoća vjerojatnosti događaja $\{a = a \mid b = b\}$
$p_{a,b}(a, b)$, $\mathrm{p}(a, b)$	Gustoća vjerojatnosti događaja $\{a = a, b = b\}$
$a \sim q$, $\mathrm{p}(a) = q$	*Slučajna varijabla a ima razdiobu q*
$a \sim \mathbb{A}$	*Slučajna varijabla a ima takvu razdiobu da svi elementi (multi)skupa $\mathbb{A}$ imaju vjerojatnost proporcionalnu višestrukosti ($\frac{1}{\lvert\mathbb{A}\rvert}$ za običan skup)*
$a \sim q$	*a se izvlači iz razidiobe q*
$a \sim a$, $a \sim \mathrm{p}(a)$	*a se izvlači iz razidobe $\mathrm{p}(a)$*
$\mathbf{E}_{a\sim a}\, f(a)$, $\mathbf{E}_a\, f(a)$	Očekivanje funkcije slučajne varijable a
$\mathbf{D}_{a\sim a}\, f(a)$, $\mathbf{D}_a\, f(a)$	Disperzija (varijanca) funkcije slučajne varijable a
$\mathrm{Cov}(a, b)$	Kovarijanca
$\mathcal{N}(\mu, \sigma^2)$	Normalna razdioba s učekivanjem μ i varijancom σ^2
$\mathcal{U}(\mathbb{A})$	Uniformna razdioba nad skupom $\mathbb{A}$

Teorija informacije

$\mathrm{I}(\mathbb{A})$	Sadržaj informacije događaja $\mathbb{A}$
$\mathrm{H}(a)$	Entropija

$\mathrm{h}(a)$	Diferencijalna entropija
$\mathrm{I}(a, b)$	Međusobna informacija
$\mathrm{H}(a \mid b)$	Uvjetna entropija
$\mathrm{H}_b(a)$	Unakrsna entropija
$\mathrm{D}_{\mathrm{KL}}(a \parallel b)$	Kullback-Leiblerova divergencija (relativna entropija)

Grafovi

$\mathrm{pa}_G(a)$	Skup čvorova roditelja čvora a u grafu G
$\mathrm{ch}_G(a)$	Skup čvorova djece čvora a u grafu G
$\mathrm{pred}_G(a)$	Skup čvorova prethodnika čvora a u grafu G
$\mathrm{succ}_G(a)$	Skup čvorova nasljednika čvora a u grafu G

Ostale oznake

$f \colon \mathbb{A} \to \mathbb{B}$	Funkcija s domenom $\mathbb{A}$ i kodomenom $\mathbb{B}$
$x \mapsto g(x)$	Definicija funkcije; funkcija koja preslikava x iz domene u $g(x)$ iz kodomene
$f * g$	Konvolucija funkcija f i g
$\lvert\mathbb{A}\rvert$	Kardinalitet skupa $\mathbb{A}$
$\delta(\cdot)$	Diracova delta
$[\![\cdot]\!]$	Iversonova uglata zagrada; $[\![P]\!] = \begin{cases} 1, & P \equiv \top \\ 0, & P \equiv \bot \end{cases}$

1. Uvod

Uvod rada. Nakon uvoda dolaze poglavlja u kojima se obrađuje tema.

duboko učenje

neizvjesnost modela

primjene procjene nesigurnosti

primjena na semantičkoj segmentaciji i procjeni dubine

1.1. Struktura rada

2. Osnovni pojmovi

2.1. Teorija vjerojatnosti

Jako važan pojam u strojnom učenju je nesigurnost ili neizvjesnost. Ona dolazi od šuma u mjerenju i iz konačnosti skupa podataka (Bishop, 2006). Teorija vjerojatnosti nam omogućuje modeliranje nesigurnosti pronalaženje optimalnih zaključaka korištenjem dostupnih informacija.

Postoje dvije glavne interpretacije vjerojatnosti (Murphy, 2012). Jedna je **frekventistička interpretacija** prema kojoj vjerojatnosti predstavljaju učestalosti različitih događaja ako se pokus ponavlja velik broj puta. Druga je **bayesovska interpretacija** prema kojoj vjerojatnost izražava našu nesigurnost o ishodu pokusa.

Ovo poglavlje daje kratak i matematički ne potpuno precizan pregled nekih od osnovnih pojmova i pravila vezanih uz vjerojatnost. Na strukturu ovog poglavlja imaju utjecaj Goodfellow et al. (2016); Murphy (2012).

2.1.1. Slučajne varijable i razdiobe

Neizvjesnost neke pojave modeliramo **slučajnom varijablom**. Slučajnoj varijabli dodijeljena je **razdioba** koja definira skup vrijednosti koje slučajna varijabla može poprimiti i vjerojatnosti ostvarivanja tih vrijednosti. Skup mogućih vrijednosti neke slučajne varijable još se naziva i **prostor elementarnih događaja**. **Elementarni događaj** je element prostora elementarnih događaja i, ako je x slučajna varijabla za koju se u nekom eksperimentu opaža vrijednost x, taj događaj ima zapis $\{x = x\}$, a njegova vjerojatnost $\mathrm{P}(\{x = x\})$ ili $\mathrm{P}(x = x)$. **Događaj** je skup vrijednosti i obično se izražava predikatom nad slučajnom varijablom: $\{R(x)\} = \{x \colon R(x)\}$. Ako je $\mathbb{X}$

prostor elementarnih događaja slučajne varijable x, onda $\mathrm{P}(x \in \mathbb{X}) = 1$. Funkcija

$$\begin{aligned} P_x \colon \mathbb{X} &\to [0,1] \\ x &\mapsto \mathrm{P}(x = x) \end{aligned}$$

je **funkcija vjerojatnosti** (engl. *probability mass function*, *pmf*).

Razlikujemo diskretne i kontinuirane slučajne varijable. Prostor elementarnih događaja diskretne slučajne varijable je prebrojiv skup. Razdioba kontinuirane slučajne varijable x koja poprima vrijednosti iz skupa $\mathbb{X}$ je određena **funkcijom gustoće vjerojatnosti** (engl. *probability density function*, *pdf*)

$$\begin{aligned} p_x \colon \mathbb{X} &\to [0,\infty) \\ x &\mapsto \mathrm{p}(x) \end{aligned}$$

za koju vrijedi

$$\mathrm{P}(x \in A) = \int_A p_x(x)\mathrm{d}x \tag{2.1}$$

za svaki $A \subset \mathbb{X}$.

Funkciju gustoće vjerojatnosti možemo smatrati i **poopćenom funkcijom**[1]. To nam omogućuje da funkcijom gustoće predstavljamo razdiobe za koje neki elementarni događaji imaju vjerojatnost veću od 0. Diskretnu razdiobu onda možda možemo predstaviti funkcijom gustoće vjerojatnosti

$$p_x(x) = \sum_{x' \in \mathbb{X}} \mathrm{P}(x = x)\delta(x - x'), \tag{2.2}$$

gdje je $\mathbb{X}$ prostor elementarnih događaja slučajne varijable x, a δ Diracova delta, poopćena funkcija za koju vrijedi $\delta(x) = 0$ za $x \neq 0$ i $\int_x \delta(x)\mathrm{d}x = 1$. Diracova delta se može promatrati kao limes funkcije gustoće Gaussove razdiobe:

$$\delta(x) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{x^2}{2\sigma^2}\right).$$

Ako je x vektor $\boldsymbol{x} = (x_1, .., x_n)$, mora vrijediti

$$\delta(\boldsymbol{x}) \coloneqq \prod_i \delta(x_i). \tag{2.3}$$

Onda n-struki integral gustoće definirane izrazom (2.2) ima vrijednost 1.

[1] https://en.wikipedia.org/wiki/Distribution_(mathematics)

Razdioba slučajne varijable x će se u ovom radu označavati s $\mathrm{P}(x)$ ako je diskretna, a s $\mathrm{p}(x)$ ako je kontinuirana ili neodređena. Funkcija (gustoće) vjerojatnosti će se označavati bez oznake slučajne varijable u indeksu ako je po slovu vrijednosti jasno o kojoj se varijabli radi. Druge oznake koje se koriste opisane su u popisu oznaka na početku rada. Na nekim mjestima će, radi kratkoće, riječ *razdioba* imati značenje *funkcija gustoće* ili *funkcija vjerojatnosti*.

2.1.2. Združena, uvjetna i marginalna vjerojatnost i osnovna pravila vjerojatnosti

Dvije razdiobe su iste ako imaju iste funkcije gustoće vjerojatnosti. Dvije slučajne varijable, i ako imaju istu razdiobu, ne moraju biti iste jer se mogu razlikovati po odnosima s drugim slučajnim varijablama.

Možemo razmatrati više slučajnih varijable zajedno (združenu slučajnu varijablu) i njihovu **združenu razdiobu** $\mathrm{p}(x, y)$. Događaji onda imaju oblik $\{R(x, y)\}$. Elementarni događaj onda ima oblik $\{x = x, y = y\}$. Dalje će se izrazi pravila vjerojatnosti odnositi samo na elemntarne događaje. Npr. x, y će skraćeno označavati $\{x = x, y = y\}$ kada je jasno po slovima o kojim se slučajnim varijablama radi. Ista pravila vjerojatnosti vrijede i za općenitije događaje jer za svaki događaj možemo definirati indikatorsku slučajnu varijablu kojoj je taj događaj elementarni događaj: $e_i = [\![R_i(x, y)]\!]$. Takve slučajne varijable imaju skup elementarnih događaja $\{0, 1\}$ i za njih vrijede ista pravila.

Uvjetna vjerojatnost je vjerojatnost nekog događaja ako je poznato da se neki drugi događaj ostvario. Ovako je definirana uvjetna vjerojatnost događaja $\{x = x\}$ ako je poznato da se ostvario događaj $\{y = y\}$:

$$\mathrm{p}(x \mid y) \coloneqq \frac{\mathrm{p}(x, y)}{\mathrm{p}(y)}. \tag{2.4}$$

Združena vjerojatnost se može rastaviti **pravilom umnoška**:

$$\mathrm{p}(x, y) = \mathrm{p}(x \mid y)\,\mathrm{p}(y). \tag{2.5}$$

Općenitije, pravilo umnoška za n slučajnih varijabli $x_1, .., x_n$ izgleda ovako:

$$\mathrm{p}(x_1, .., x_n) = \mathrm{p}(x_1)\,\mathrm{p}(x_2 \mid x_1) \cdots \mathrm{p}(x_n \mid x_1, .., x_{n-1}) \tag{2.6}$$

$$= \mathrm{p}(x_1) \prod_{i=2..n} \mathrm{p}(x_i \mid x_1, .., x_{i-1}). \tag{2.7}$$

Marginalna vjerojatnost slučajne varijable x je $\mathrm{p}(x) = \mathrm{p}(x = x, y \in \mathbb{Y})$, gdje je $\mathbb{Y}$ prostor elementarnih događaja slučajne varijable y. Izraženo gustoćom vjerojatnosti (**pravilo zbroja**, **marginalizacija**):

$$\mathrm{p}(x) = \int_{\mathbb{Y}} \mathrm{p}(x, y)\mathrm{d}y = \int_{\mathbb{Y}} \mathrm{p}(x \mid y)\,\mathrm{p}(y)\mathrm{d}y. \tag{2.8}$$

Dvije slučajne varijable koje imaju istu razdiobu ne moraju biti u istom odnosu prema drugim slučajnim varijablama. Npr. ako $x_1 \sim q_1$, $x_2 \sim q_1$ i $y \sim q_2$, ne mora vrijediti $\mathrm{p}(x_1, y) = \mathrm{p}(x_2, y)$.

Rastavljanjem lijeve strane jednadžbe (2.6) na umnožak $\mathrm{p}(x \mid y)\,\mathrm{p}(y)$ dobivamo **Bayesovo pravilo**:

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(y \mid x)\,\mathrm{p}(x)}{\mathrm{p}(y)}, \tag{2.9}$$

što možemo i ovako zapisati:

$$\mathrm{p}(x \mid y) = \frac{\mathrm{p}(y \mid x)\,\mathrm{p}(x)}{\int \mathrm{p}(y \mid x)\,\mathrm{p}(x)\mathrm{d}x}, \tag{2.10}$$

gdje se nazivnik integrira po svim vrijednostima.

2.1.3. Nezavisnost, uvjetna nezavisnost i uvjetna zavisnost

Kada su dvije slučajne varijable x i y **zavisne**, što se označava $x \not\perp y$, znanje o ishodu jedne utječe na znanje o ishodu druge, tj. uvjetna razdioba $\mathrm{p}(x \mid y = y)$ ovisi o ishodu y. *Znanje o ishodu* ne mora značiti da je ishod poznat. Dovoljna je promjena znanja o razdiobi koja može biti posljedica opažanja neke treće slučajne varijable. Slučajne varijable x i y su **nezavisne**, što se označava $x \perp y$, akko za svaki par (x, y) vrijedi

$$\mathrm{p}(x, y) = \mathrm{p}(x)\,\mathrm{p}(y), \tag{2.11}$$

ili, ekvivalentno,

$$\mathrm{p}(x \mid y) = \mathrm{p}(x). \tag{2.12}$$

Znanje o ishodu jedne slučajne varijable onda ne utječe na znanje o ishodu druge.

Slučajne varijable x i y, koje mogu biti zavisne, su uz znanje o ishodu slučajne varijable z **uvjetno nezavisne**, što se označava $x \perp y \mid z$, akko su slučajne varijable $(x \mid z = z)$ i $(y \mid z = z)$ nezavisne za svaki mogući ishod z. Onda za svaku trojku (x, y, z) vrijedi

$$\mathrm{p}(x, y \mid z) = \mathrm{p}(x \mid z)\,\mathrm{p}(y \mid z), \tag{2.13}$$

ili, ekvivalentno,

$$\mathrm{p}(x \mid y, z) = \mathrm{p}(x \mid z). \tag{2.14}$$

Isto tako, slučajne varijable x i y koje su nezavisne mogu biti **uvjetno zavisne** uz znanje o ishodu neke slučajne varijable z. Općenito, dvije slučajne varijable ne moraju biti ni uvjetno zavisne ni uvjetno nezavisne jer neki ishodi treće slučajne varijable mogu utjecati na njihovu zavisnost, a neki ne. Također se može govoriti i o zavisnosti ili nezavisnosti pojedinih događaja.

2.1.4. Očekivanje, varijanca i kovarijanca

Očekivanje (prvi moment) slučajne varijable definirano je ovako:

$$\mathbf{E}\,x \coloneqq \int x\,\mathrm{p}(x)\mathrm{d}x, \tag{2.15}$$

gdje se integrira po prostoru elementarnih događaja. Još se označava ovako: μ_x. Očekivanje funkcije slučajne varijable zapisujemo ovako:

$$\mathop{\mathbf{E}}_{x \sim x} f(x) \coloneqq \mathbf{E}\,f(x) = \int f(x)\,\mathrm{p}(x)\mathrm{d}x. \tag{2.16}$$

Ako je po oznaci jasno o kojoj se slučajnoj varijabli radi, možemo kraće pisati $\mathbf{E}_x\,f(x)$. Očekivanje ima svojstvo linearnosti:

$$\mathbf{E}[\alpha f(x) + \beta g(x)] = \alpha\,\mathbf{E}\,f(x) + \beta\,\mathbf{E}\,g(x). \tag{2.17}$$

Varijanca (disperzija, drugi centralni moment) slučajne varijable definirana je

ovako:

$$\mathbf{D}\,x \coloneqq \mathbf{E}\Big[(x - \mathbf{E}\,x)^2\Big] = \int (x - \mathbf{E}\,x)^2\,\mathrm{p}(x)\mathrm{d}x. \tag{2.18}$$

Varijanca se može izraziti preko drugog momenta $\mathbf{E}\,x^2$ i kvadrata očekivanja $(\mathbf{E}\,x)^2$:

$$\mathbf{D}\,x = \mathbf{E}\Big[(x - \mathbf{E}\,x)^2\Big] = \mathbf{E}\Big[\Big(x^2 - 2x\,\mathbf{E}\,x + (\mathbf{E}\,x)^2\Big)\Big] \tag{2.19}$$

$$= \mathbf{E}\,x^2 - 2(\mathbf{E}\,x)^2 + (\mathbf{E}\,x)^2 = \mathbf{E}\,x^2 - (\mathbf{E}\,x)^2. \tag{2.20}$$

Drugi korijen varijance je standardna devijacija σ_x.

Kovarijanca para slučajnih varijabli definirana je ovako:

$$\mathrm{Cov}(x, y) \coloneqq \mathbf{E}[(x - \mathbf{E}\,x)(y - \mathbf{E}\,y)] = \mathbf{E}\,xy - (\mathbf{E}\,x)(\mathbf{E}\,y). \tag{2.21}$$

Kovarijacijska matrica slučajnog vektora $\boldsymbol{x} \in \mathbb{R}^n$ je matrica tipa $n \times n$ takva da:

$$\mathrm{Cov}(\boldsymbol{x})_{[i,j]} = \mathrm{Cov}\Big(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}\Big). \tag{2.22}$$

Dijagonalni elementi te matrice su $\mathrm{Cov}(\boldsymbol{x})_{[i,i]} = \mathbf{D}\,\boldsymbol{x}_{[i]}$.

2.1.5. Funkcije slučajnih varijabli

Neka je odnos između slučajnih varijabli x i y definiran funkcijom f koja ishode jedne slučajne varijable deterministički preslikava u ishode druge, što se označava ovako: $y = f(x)$. Ako su x i y diskretne slučajne varijable, onda je razdioba slučajne varijable y definirana ovako:

$$P_y(y) = \sum_{x\colon\, f(x)=y} P_x(x). \tag{2.23}$$

Ako su x i y kontinuirane slučajne varijable s vrijednostima iz $\mathbb{R}$ i f je injektivna, može se pokazati (Elezović, 2007) da vrijedi

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|. \tag{2.24}$$

To se može poopćiti i na vektore. Onda je $p_y(\boldsymbol{y}) = \left|\det \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right|$ (Murphy, 2012).

Neka je z zbroj slučajnih varijabli x i y. Onda vrijedi

$$p_z(z) = \int p_{x,y}(x, z - x)\mathrm{d}x. \tag{2.25}$$

Ako su x i y nezavisne, onda to postaje konvolucija:

$$p_z(z) = \int p_x(x) p_y(z - x) \mathrm{d}x =: (p_x * p_y)(z). \tag{2.26}$$

2.1.6. Primjeri razdioba

Bernoullijeva razdioba je binarna razdioba s prostorom elementarnih događaja koji je obično $\{0, 1\}$. Ona je onda određena parametrom $\mu \in [0, 1]$ i ima ova svojstva:

$$\mathrm{P}(x) = \mu[\![x = 1]\!] + (1 - \mu)[\![x = 0]\!] = \mu^x (1 - \mu)^{1-x}, \tag{2.27}$$

$$\mathbf{E}\, x = \mu, \tag{2.28}$$

$$\mathbf{D}\, x = \mu(1 - \mu). \tag{2.29}$$

Kategorička razdioba je poopćenje Bernoullijeve razdiobe na konačan prostor elementarnih događaja koji može imati više od 2 vrijednosti. Ako prostor elementarnih događaja ima kardinalitet n, razdioba je određena vektorom $\boldsymbol{p} \in [0, 1]^{n-1}$ za koji vrijedi $\sum_i p_{[i]} \leq 1$. Prostor elementarnih događaja ne mora biti skup $\{1..n\}$ pa je kategorička razdioba najopćenitija diskretna razdioba nad konačnim skupom elementarnih događaja.

Eksponencijalna razdioba je kontinuirana razdioba s domenom $\mathbb{R}_{\geq 0}$. Ona je definirana parametrom $\lambda \in \mathbb{R}_{>0}$ ili $\beta = \lambda^{-1}$ i ima ova svojstva:

$$\mathrm{p}(x) = \lambda \exp(-\lambda x) \tag{2.30}$$

$$\mathbf{E}\, x = \lambda^{-1}, \tag{2.31}$$

$$\mathbf{D}\, x = \lambda^{-2}. \tag{2.32}$$

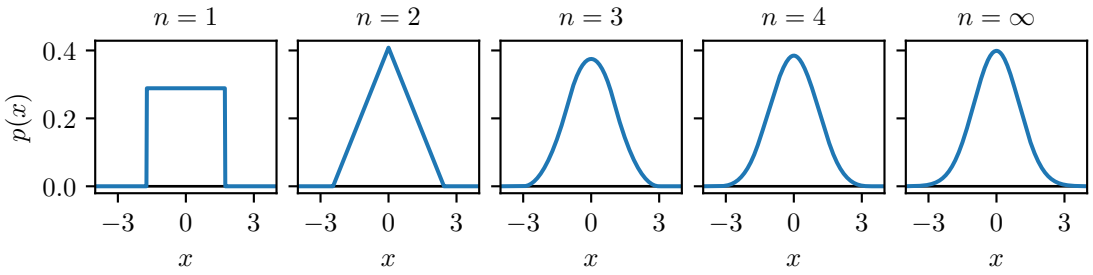
Laplaceova razdioba je kontinuirana razdioba definirana parametrima $\beta \in \mathbb{R}_{>0}$ i $\mu \in \mathbb{R}$ i ima ova svojstva:

$$\mathrm{p}(x) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta}\right) \tag{2.33}$$

$$\mathbf{E}\, x = \mu, \tag{2.34}$$

$$\mathbf{D}\, x = \beta^2. \tag{2.35}$$

Gaussova (normalna) razdioba je kontinuirana razdioba definirana



Slika 2.1: Ilustracija centralnog graničnog teorema. Grafovi za različite brojeve pribrojnika n prikazuju funkcije gustoće vjerojatnosti normaliziranih zbrojeva nezavisnih slučajnih varijabli s razdiobom prikazanom prvim grafom. Zadnji graf prikazuje funkciju gustoće Gaussove razdiobe s očekivanjem 0 i varijancom 1.

parametrima $\mu \in \mathbb{R}$ i $\sigma \in \mathbb{R}_{>0}$ i ima ova svojstva:

$$\mathrm{p}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.36}$$

$$\mathbf{E}\, x = \mu, \tag{2.37}$$

$$\mathbf{D}\, x = \sigma^2. \tag{2.38}$$

Neka je $z_n = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma\sqrt{n}}$ normalizirani zbroj n nezavisnih slučanih varijabli x_i koje imaju jednaku razdiobu s očekivanjem μ i varijancom σ^2. Prema centralnom graničnom teoremu, z_n u razdiobi konvergira prema Gaussovoj razdiobi kada $n \to \infty$, tj.

$$\lim_{n\to\infty} \mathrm{P}(z_n < z) = \int_{-\infty}^{z} p_{\mathcal{N}(0,1)}(z')\mathrm{d}z'. \tag{2.39}$$

$p_{\mathcal{N}(0,1)}$ označava funkciju gustoće normalne razdiobu s $\mu = 0$ i $\sigma = 1$. To je detaljnije objašnjeno i dokazano npr. u (Elezović, 2007). Centralni granični teorem je ilustriran na slici 2.1.

2.2. Teorija informacije

Jedan od osnovnih pojmova u teoriji informacije je **sadržaj informacije** koji događaj preslikava u nenegativan realni broj:

$$\mathrm{I}(x \in A) := \log_b \frac{1}{\mathrm{P}(x \in A)} = -\log_b \mathrm{P}(x \in A). \tag{2.40}$$

Događaji koji imaju manju vjerojatnost sadrže više informacije. Ako je vjerojatnost nekog događaja 1, njegov sadržaj informacije je 0. b je najčešće 2 ili e.

Sadržaj informacije odgovara minimalnom broju simbola (bitova ako $b = 2$) potrebnih za kodiranje elementarnih događaja prefiksnim kodom za koji je očekivanje duljine poruke minimalno (Olah, 2015). Kod prefiksnog koda nijedna kodna riječ nije prefiks neke druge kodne riječi. Takav kod se može prenositi kao niz združenih kodnih riječi bez posebnog simbola za označavanje granica između kodnih riječi. Donja granica očekivanja duljine poruke kod optimalnog koda naziva se **entropija**:

$$\mathrm{H}(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(x = x) = -\mathop{\mathbf{E}}_{x} \log_b \mathrm{P}(x). \tag{2.41}$$

Ona iskazuje neizvjesnost diskretne slučajne varijable. Entropija će biti 0 ako je vjerojatnost nekog elementarnog događaja 1, a najveća će biti kada svi elementarni događaji imaju istu vjerojatnost: $\mathrm{H}(x) = \log_b n$, gdje je n broj elementarnih događaja.

Entropija kontinuirane slučajne varijable je beskonačna. Ako se u izrazu (2.41) vjerojatnost zamijeni gustoćom vjerojatnosti, onda on predstavlja **diferencijalnu entropiju**, jedan od analoga[2] entropije za kontinuirane varijable koji nema neka od svojstava koja ima entropija.

Unakrsna entropija je mjera koja iskazuje donju granicu očekivanja duljine poruke kodirane optimalnim kodom za razdiobu $\mathrm{P}(y)$ dok izvor poruka ima razdiobu $\mathrm{P}(x)$. Ovako je definirana:

$$\mathrm{H}_y(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(y = x) = -\mathop{\mathbf{E}}_{x} \log_b P_y(x). \tag{2.42}$$

Za $\mathrm{P}(y) = \mathrm{P}(x)$ je $\mathrm{H}_y(x) = \mathrm{H}_x(x) = \mathrm{H}(x)$. Za unakrsnu entropiju se često koristi oznaka $\mathrm{H}(x, y)$, ali ista oznaka se koristi i za entropiju združene slučajne varijable (x, y). Po uzoru na Olah (2015), ovdje koristimo oznaku $\mathrm{H}_y(x)$.

Kao mjera razlike između dviju razdioba često se koristi **relativna entropija** ili **Kullback-Leiblerova divergencija** (KL-divergencija):

$$\mathrm{D}_{\mathrm{KL}}(x \parallel y) \coloneqq \mathrm{H}_y(x) - \mathrm{H}(x) = \mathop{\mathbf{E}}_{x} \log_b \frac{P_x(x)}{P_y(x)}. \tag{2.43}$$

Ona je uvijek pozitivna i mjeri koliko simbola više se u prosjeku koristi ako se opaža razdioba $\mathrm{P}(x)$, a događaji se kodiraju kodom optimalnim za razdiobu $\mathrm{P}(y)$. KL-divergencija će biti 0 akko x i y imaju iste razdiobe. To je ilustrirano slikom 2.2. KL-divergencija, kao ni unakrsna entropija, nije simetrična (slika 2.3), tj. općenito $\mathrm{D}_{\mathrm{KL}}(x \parallel y) \neq \mathrm{D}_{\mathrm{KL}}(y \parallel x)$ i $\mathrm{H}_y(x) \neq \mathrm{H}_x(y)$. KL-divergencija je izrazom (2.43)

[2] https://en.wikipedia.org/wiki/Differential_entropy

$\mathrm{H}(x)$	$\mathrm{D}_{\mathrm{KL}}(x \parallel y)$
$\mathrm{H}_y(x)$	

Slika 2.2: Odnos entropije, unakrsne entropije i KL-divergencije.

definirana i za kontinuirane slučajne varijable ako se funkcije vjerojatnosti zamijene funkcijama gustoće vjerojatnosti. Ona divergira kada postoji x za koji $P_x(x) > 0$ i $P_y(x) = 0$ ili, u slučaju kontinuiranih razdioba, $p_x(x) > 0$ i $p_y(x) = 0$.

Međusobna informacija je mjera zavisnosti između slučajnih varijabli. Definirana je ovako:

$$\mathrm{I}(x; y) \coloneqq \mathop{\mathbf{E}}_{x,y} \log_b \frac{P_{x,y}(x, y)}{P_x(x) P_y(y)}, \tag{2.44}$$

a može se i na ove načine izraziti:

$$\mathrm{I}(x; y) = \mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y) \tag{2.45}$$

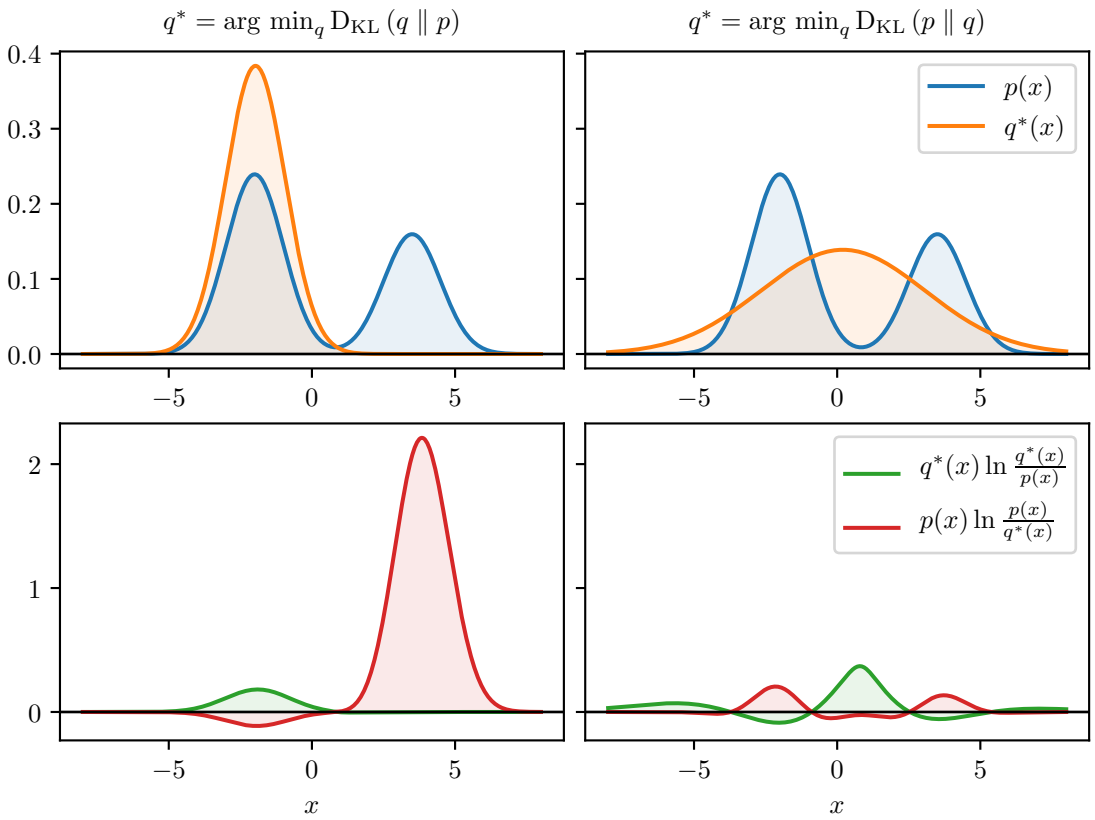
$$= \mathrm{H}(x) - \mathrm{H}(x \mid y) \tag{2.46}$$

$$= \mathrm{H}(y) - \mathrm{H}(y \mid x), \tag{2.47}$$

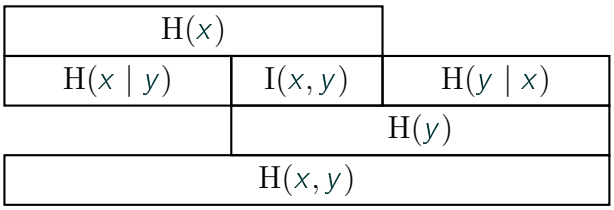
gdje je

$$\mathrm{H}(x \mid y) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{H}(y \mid x = x) \tag{2.48}$$

uvjetna entropija. Ako su x i y nezavisne, njihova međusobna informacija će biti 0. Ako npr. postoji surjekcija f tako da $y = f(x)$, tj. poznavanje ishoda varijable x jednoznačno određuje ishod varijable y, onda $\mathrm{H}(y \mid x) = 0$ i $\mathrm{I}(x; y) = \mathrm{H}(y)$. Ako je f bijekcija, onda $\mathrm{I}(x; y) = \mathrm{H}(x) = \mathrm{H}(y)$. Definirane veličine mogu se prikazati kao na slici 2.4. Isti odnosi vrijede ako se entropija zamijeni diferencijalnom entropijom.



Slika 2.3: Asimetričnost KL-divergencije. p je fiksna razdioba (funkcija gustoće), a $q*$ je Gaussova razdioba koja ju aproksimira minimizacijom KL-divergencije $\mathrm{D_{KL}}(q \parallel p)$ (lijevo) ili $\mathrm{D_{KL}}(p \parallel q)$ (desno). U donjem retku grafovi prikazuju podintegralne funkcije odgovarajućih KL-divergencija. Kod njih zbrojevi predznačenih površina obojanih područja odgovaraju KL-divergencijama $\mathrm{D_{KL}}(q \parallel p)$ (zeleno) ili $\mathrm{D_{KL}}(p \parallel q)$ (crveno). Optimalna aproksimirajuća razdioba desno ima veliku gustoću gdje god razdioba p ima veliku gustoću. Lijevo optimalna aproksimirajuća razdioba nema veliku gustoću gdje razdioba p nema veliku gustoću. Da je razmak između komponenata razdiobe p malo manji, i lijeva razdioba q^* bi pokrila oba moda i bila sličnija desnoj. Slika je napravljena po uzoru na sliku 3.6 u Goodfellow et al. (2016).



Slika 2.4: Odnosi informacijsko-teorijskih veličina dviju slučajnih varijabli.

2.3. Optimizacija temeljena na gradijentu

U ovom odjeljku su opisani osnovni optimizacijski algoritmi temeljeni na gradijentu. Oni su bitni u strojnom učenju (poglavlje 4), posebno u dubokom učenju (poglavlje 5). Izvedeni algoritmi koji se primjenjuju u dubokom učenju opisani su u pododjeljku 5.3.2.

Neka je $f\colon \mathbb{R}^n \to \mathbb{R}$ funkcija čiji minumom želimo naći s obziroma na parametre $\boldsymbol{x}$. Ona se u okolini točke $\boldsymbol{x}$, ako je dovoljno (beskonačno) puta derivabilna može izraziti Taylorovim redom:

$$f(\boldsymbol{x}+\boldsymbol{d}) = f(\boldsymbol{x}) + \tfrac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\mathsf{T} \tfrac{\partial^2 y}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\mathsf{T}} \boldsymbol{x} f(\boldsymbol{x})\boldsymbol{d} + \cdots. \tag{2.49}$$

S drugačijim oznakama:

$$f(\boldsymbol{x}+\boldsymbol{d}) = f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^\mathsf{T}\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\mathsf{T}\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{d} + \cdots. \tag{2.50}$$

2.3.1. Gradijentni spust

Ako je $\boldsymbol{d}$ ima malu normu, funkciju f u okoline neke točke možemo dobro aproksimirati s prvih nekoliko članova Taylorovog reda. **Gradijentni spust** je optimizacijski algoritam koji koristi linearnu aproksimaciju i iterativnim ažuriranjem parametara u smjeru gradijenta (*najstrmijem* smjeru) traži minimum. Iteracija gradijentnog spusta ima ovakav oblik:

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i), \tag{2.51}$$

gdje je i redni broj iteracije, a η **veličina koraka** (**stopa učenja** kod strojnog učenja) koja može biti konstanta ili može ovisiti o broju iteracije i. Neka $\boldsymbol{g} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ i $\boldsymbol{H} = \boldsymbol{H}_f(\boldsymbol{x})$. Za dovoljno mal η

$$f(\boldsymbol{x} - \eta\boldsymbol{g}) \approx f(\boldsymbol{x}) - \eta\boldsymbol{g}^\mathsf{T}\boldsymbol{g} - \frac{1}{2}\eta^2\boldsymbol{g}^\mathsf{T}\boldsymbol{H}\boldsymbol{g} \tag{2.52}$$

Uz neke blage uvjete koje mora zadovoljavati f i dovoljno mal η, gradijentni spust konvergira, tj. proizvoljno se blizu približi nekom lokalnom minimumu (ili stacionarnoj točki koja nije lokalni minimum, gdje $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{0}$) ovisno o η. Jedan blagi uvjet može biti **Lipschitz kontinuiranost** funkcije f ili njene derivacije (Goodfellow et al., 2016). Funkcija f je Lipschitz kontinuirana ako postoji

konstanta λ za koju za svaki par $(\boldsymbol{x}, \boldsymbol{y})$ vrijedi:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| < \lambda \|\boldsymbol{x} - \boldsymbol{y}\| . \tag{2.53}$$

Najmanji takav λ naziva se **Lipschitzova konstanta**.

2.3.2. Postupci drugog reda

Ovaj pododjeljak se temelji na Goodfellow et al. (2016).

Ako koristimo kvadratnu aproksimaciju (2.52), možemo pokušati pronaći optimalni η koji ju minimizira. η za koji $\frac{\partial}{\partial \eta} f(\boldsymbol{x} - \eta \boldsymbol{g}) = 0$ će, ako $\boldsymbol{g}^\mathsf{T} \boldsymbol{H} \boldsymbol{g} > 0$ dati minimum u smjeru gradijenta kvadratne aproksimacije funkcije f u točki $\boldsymbol{x}$. Dobije se:

$$\eta = \frac{\boldsymbol{g}^\mathsf{T} \boldsymbol{g}}{\boldsymbol{g}^\mathsf{T} \boldsymbol{H} \boldsymbol{g}}. \tag{2.54}$$

Ako je $f \colon \mathbb{R}^n \to \mathbb{R}$ konveksna (pozitivno definitna) kvadratna funkcija, izmijenjeni algoritam gradijentnog spusta, koji ovako određuje veličinu koraka, minimum pronalazi u najviše n koraka.

Postupak drugog reda koji se ne ograničava na pomake u smjeru gradijenta je **Newton-Raphsonov postupak**. Deriviranjem desne strane jednadžbe (2.50) po $\boldsymbol{d}$ i izjednačavanjem s $\mathbf{0}$ dobiva se:

$$\mathbf{0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^\mathsf{T} + \boldsymbol{d}^\mathsf{T} \boldsymbol{H}_f(\boldsymbol{x}) + \cdots . \tag{2.55}$$

Uz kvadratnu aproksimaciju i kraće oznake $\boldsymbol{g} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ i $\boldsymbol{H} = \boldsymbol{H}_f(\boldsymbol{x})$: $\boldsymbol{H}\boldsymbol{d} = \boldsymbol{g}$. Slijedi da je pomak $\boldsymbol{d}$ koji daje stacionarnu točku aproksimacije

$$\boldsymbol{d} = \boldsymbol{H}^{-1} \boldsymbol{g}. \tag{2.56}$$

Za nekvadratne funkcije, koje imaju pozitivno definitnu Hesseovu matricu u svakoj točki, može se iterativno primjenjivati

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \eta \boldsymbol{H}_f(\boldsymbol{x}_i) \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i) \tag{2.57}$$

s $\eta < 1$.

3. Statističko modeliranje

3.1. Probabilistički grafički modeli

Neka su $x_1, .., x_n$ slučajne varijable čiju združenu razdiobu razmatramo. Želimo na temelju opežanih varijabli korištenjem pravila vjerojatnosti **zaključivati** o razdiobama nekih neopažanih varijabli. Općenito, zaključivanje se provodi uvjetovanjem po opažanim varijablama i marginalizacijom po varijablama koje nas ne zanimaju izravno (Murphy, 2012):

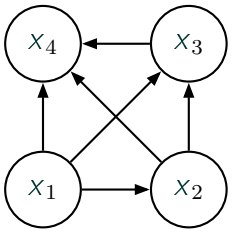
$$\mathrm{p}(\boldsymbol{x}_\mathsf{q} \mid \boldsymbol{x}_\mathsf{o}) = \frac{\mathrm{p}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{o})}{\mathrm{p}(\boldsymbol{x}_\mathsf{o})} = \frac{\int \mathrm{p}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{n}, \boldsymbol{x}_\mathsf{o}) \mathrm{d}\boldsymbol{x}_\mathsf{n}}{\int \mathrm{p}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{n}, \boldsymbol{x}_\mathsf{o}) \mathrm{d}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{n})}. \tag{3.1}$$

Ovdje je $\boldsymbol{x}_\mathsf{q}$ niz varijabli o kojima želimo zaključivati (varijable upita), $\boldsymbol{x}_\mathsf{o}$ niz opažanih varijabli, a $\boldsymbol{x}_\mathsf{n}$ niz varijabli *smetnje* (*nuisance*).

Zavisnosti između slučajnih varijabli otežavaju modeliranje i zaključivanje – potrebno je više podataka i zaključivanje je računski zahtjevnije. Obično možemo pretpostaviti uvjetne zavisnosti između slučajnih varijabli, što se može predstaviti neusmjerenim ili usmjerenim grafom. Prema definiciji na Wikipediji [1], **probabilistički grafički model** ili **grafički model** je probabilistički model koji se može prikazati grafom koji izražava strukturu uvjetnih zavisnosti među slučajnim varijablama. U tom grafu čvorovi označavaju slučajne varijable, a bridovi zavisnosti. Umjereni bridovi označavaju modeliranje uvjetne zavisnosti, a neusmjereni združeno modeliranje. Ako je graf grafičkog modela usmjeren i acikličan, on se naziva **Bayesova mreža** ili **Bayesovski model**, a ako je neusmjeren, naziva se **Markovljeva mreža** ili **Markovljevo slučajno polje** (engl. *Markov random field*, *MRF*). U nastavku ovog odjeljka naglasak će biti na Bayesovim mrežama.

Združena razdioba se prema pravilu umnoška (jednadžba 2.6) može npr. ovako

[1] https://en.wikipedia.org/wiki/Graphical_model



Slika 3.1: Prikaz grafičkog modela s faktorizacijom $\mathrm{p}(x_1, x_2, x_3, x_4) = \mathrm{p}(x_1)\,\mathrm{p}(x_2 \mid x_1)\,\mathrm{p}(x_3 \mid x_1, x_2)\,\mathrm{p}(x_4 \mid x_1, x_2, x_3)$.

izraziti:

$$\mathrm{p}(x_1, .., x_n) = \mathrm{p}(x_1)\,\mathrm{p}(x_2 \mid x_1) \cdots \mathrm{p}(x_n \mid x_1, .., x_{n-1}) \tag{3.2}$$

$$= \prod_i \mathrm{p}(x_i \mid x_1, .., x_{i-1}). \tag{3.3}$$

Prema tome, svaki probabilistički grafički model ima ekvivalentnu Bayesovu mrežu. Ako uzmemo $n = 4$, graf koji odgovara faktorizaciji u jednadžbi (3.5) prikazan je na slici 3.1.

Pretpostavljanjem uvjetnih nezavisnosti, neki bridovi grafa G se mogu ukloniti pa za varijable (čvorove grafa) vrijedi **uređajno Markovljevo svojstvo**:

$$x \perp \mathrm{pred}_G(x) \setminus \mathrm{pa}_G(x) \mid \mathrm{pa}_G(x). \tag{3.4}$$

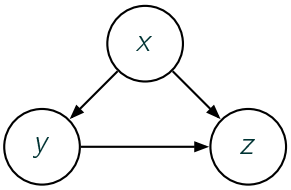
Jednadžba (3.5) onda prelazi u

$$\mathrm{p}(x_1, .., x_n) = \prod_i \mathrm{p}\left(x_i \,\middle|\, \bigcap_{x_j \in \mathrm{pa}_G(x_i)} \left\{ x_j = x_j \right\} \right). \tag{3.5}$$

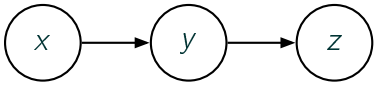
To omogućuje primjenu efikasnijih algoritama za zaključivanje (Murphy, 2012). Na slici 3.2 prikazani su osnovni slučajevi odnosa između triju slučajnih varijabli povezanih zavisnostima koje mogu biti dio većeg grafa. Oni su detaljnije objašnjeni npr. u Bishop (2006) i Alpaydin (2014).

Na slici 3.3 prikazan je primjer na kojemu se koriste još neke oznake: sivi čvorovi označavaju opažane varijable, četverokut označava veći broj podgrafova s istom strukturom.

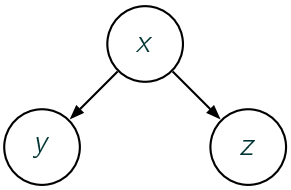
Općenitije, o uvjetnoj nezavisnosti podskupova varijabli govori svojstvo **d-separacije**. Kažemo da je staza (podgraf sa strukturom lanca) P grafa G **d-odvojena** skupom čvorova $\mathbb{E}$ akko P sadrži barem jedno od sljedećeg (Murphy, 2012):



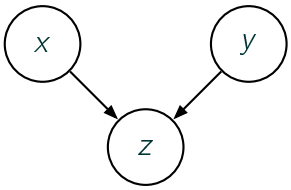
(a) Grafički model s faktorizacijom $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y \mid x)\,\mathrm{p}(z \mid x, y)$.



(b) Uz $x \perp z \mid y$ faktorizacija postaje $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y \mid x)\,\mathrm{p}(z \mid y)$ (lanac).

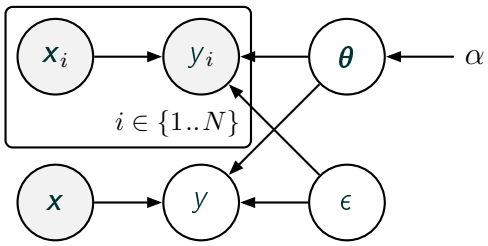


(c) Uz $y \perp z \mid x$ faktorizacija postaje $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y \mid x)\,\mathrm{p}(z \mid x)$ (račvanje).



(d) Uz $x \perp y$ faktorizacija postaje $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y)\,\mathrm{p}(z \mid x, y)$ (sraz). Ovdje također vrijedi $x \not\perp y \mid z$.

Slika 3.2: Osnovni slučajevi uvjetne nezavisnosti. Slike b, c i d prikazuju grafove dobivene uvođenjem pretpostavki uvjetne nezavisnosti za grafički model s 3 slučajne varijable prikazan na slici a.



Slika 3.3: Primjer grafičkog modela s faktorizacijom $\mathrm{p}(\boldsymbol{x}, y, \boldsymbol{x}_1..\boldsymbol{x}_N, y_1..y_N, \boldsymbol{\theta}, \epsilon) = \mathrm{p}(\boldsymbol{\theta})\,\mathrm{p}(\epsilon) p_{\boldsymbol{x}}(\boldsymbol{x}) p_{y|\boldsymbol{x},\boldsymbol{\theta},\epsilon}(y \mid \boldsymbol{x}, \boldsymbol{\theta}, \epsilon) \prod_i \Big(p_{\boldsymbol{x}}(\boldsymbol{x}_i) p_{y|\boldsymbol{x},\boldsymbol{\theta},\epsilon}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \epsilon) \Big)$. Graf prikazuje model regresije, gdje su $\boldsymbol{\theta}$ nepoznati parametri, $\boldsymbol{x}_i$ i y_i opažani parovi ulaza i izlaza, $\boldsymbol{x}$ opažani ulaz s nepoznati izlazom, a ϵ homoskedastički šum, tj. šum koji ne ovisi o ulazu. Na slici je još eksplicitno prikazana deterministička varijabla α koja je parametar razdiobe $\mathrm{p}(\boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{\theta} \mid \alpha)$. Slika je napravljena po uzoru na sliku 14.7 u Alpaydin (2014).

- lanac $a \to b \to c$, gdje $b \in E$
- račvanje $a \leftarrow b \to c$, gdje $b \in E$
- sraz $a \to b \leftarrow c$, gdje $\forall b' \in \{b\} \cup \mathrm{succ}_G(b), b' \notin E$.

Kažemo da skup čvorova $\mathbb{E}$ d-odvaja čvorove x i y akko su sve staze između njih d-odvojene. Vrijedi $x \perp y \mid \mathbb{E}$ akko skup čvorova $\mathbb{E}$ d-odvaja čvorove x i y. To se može poopćiti na skupove čvorova. Skup čvorova opažanjem kojega neki čvor postaje neovisan o ostatku grafa naziva se **Markoveljev pokrivač** (engl. *Markov blanket*). Markovljev pokrivač čvora x je

$$\mathrm{pa}_G(x) \cup \mathrm{ch}_G(x) \cup \bigcup_{y \in \mathrm{ch}_G(x)} \mathrm{pa}_G(y) \tag{3.6}$$

U navedenim knjigama opisani su algoritmi koji se koriste za efikasno zaključivanje iskorištavanjem strukture grafa.

3.2. Procjena parametara i zaključivanje

3.2.1. Procjenitelji i točkaste procjene parametara

Ovaj pododjeljak se temelji na Elezović (2007).

Neka je x slučajna varijabla i $\mathrm{p}(x)$ njena razdioba s nama nepoznatim parametrom θ. Taj parametar možemo procijeniti na temelju opaženih vrijednosti $x_1, .., x_n$ slučajne varijable x, za što definiramo funkciju g koja daje procjenu parametara

$$\hat{\theta} = f(x_1, .., x_N). \tag{3.7}$$

Ako kao parametre takve funkcije uzmemo **uzorak**, tj. skup slučajnih varijabli $\mathbb{D} = (x_1, .., x_N)$, gdje pretpostavljamo da su $x_1, .., x_N$ međusobno nezavisne i imaju istu razdiobu kao x, dobivamo slučajnu varijablu

$$\hat{\theta} = f(\mathbb{D}). \tag{3.8}$$

Takva slučajna varijabla naziva se **statistika**. Ako je θ nepoznati parametar razdiobe $\mathrm{p}(x)$, onda kažemo da je ta statistika $\hat{\theta}$ **procjenitelj** parametra θ, a njen ishod $\hat{\theta}$ **procjena** parametra θ.

3.2.2. Svojstva i pogreška procjenitelja

Pristranost procjenitelja $\hat{\theta}$ je definirana izrazom $\mathbf{E}\,\hat{\theta} - \theta$, gdje je θ stvarna vrijednost parametra koji se procjenjuje. Ona mjeri koliko procjenitelj griješi neovisno o ishodu uzorka. Kažemo da je procjenitelj parametra θ **nepristran** ako vrijedi

$$\mathbf{E}\,\hat{\theta} = \theta. \tag{3.9}$$

Varijanca procjenitelja $\hat{\theta}$ je definirana izrazom $\mathbf{D}\,\hat{\theta}$. Ona mjeri koliko procijenitelj griješi ovisno variranju uzorka. Neka N u oznaci $\mathbb{D}_N$ označava veličinu uzorka. Nepristrani procjenitelj $\hat{\theta}$ je **valjan** ako

$$\lim_{N\to\infty} \mathbf{D}\Big[\hat{\theta}(\mathbb{D}_N)\Big] = 0. \tag{3.10}$$

Može se pokazati da je očekivanje srednje kvadratne pogreške procjenitelja jednaka zbroju njegove varijance i kvadrata njegove pristranosti (Šnajder i Dalbelo Bašić, 2014), tj.

$$\mathbf{E}\Big[\big(\hat{\theta} - \theta\big)^2\Big] = \mathbf{D}\,\hat{\theta} + \big(\mathbf{E}\,\hat{\theta}\big)^2. \tag{3.11}$$

3.2.3. Procjenitelj maksimalne izglednosti

Procjenitelj maksimalne izglednosti (**ML-procjenitelj**, engl. *maximum likelihood*) uzorku dodjeljuje parametre maksimiziraju vjerojatnost uzorka, tj. imaju najveću **izglednost**:

$$\boldsymbol{\theta}_{\mathsf{ML}} = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}). \tag{3.12}$$

Zbog pretpostavke međusobne nezavisnosti primjera vrijedi

$$\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{\boldsymbol{d}\in\mathbb{D}} \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{3.13}$$

Za razliku od generativnih, diskriminativni modeli ne modeliraju razdiobu ulaznih primjera, nego samo uvjetnu razdiobu $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D})$ pa kod njih razdioba ulaznih primjera ne ovisi o $\boldsymbol{\theta}$, tj. $\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{x})$. Onda je izglednost

$$\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{x}) \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{3.14}$$

Faktor $\mathrm{p}(\boldsymbol{x})$ ne ovisi o parametrima i može se zanemariti pri optimizaciji.

3.2.4. Procjenitelj maksimalne aposteriorne vjerojatnosti

Procjenitelj maksimalne aposteriorne vjerojatnosti (**MAP-procjenitelj**, engl. *maximum a posteriori estimator*) u obzir uzima **apriornu razdiobu** $\mathrm{p}(\boldsymbol{\theta})$ koja predstavlja dodatne pretpostavke za razdiobu parametara. Apriorna razdioba parametara pojednostavljuje model dajući prednost nekim hipotezama i posebno je korisna kada ima malo podataka. Apriorna razdioba može biti definirana nekim hiperparametrima ali oni ovdje nisu prikazani. Po Bayesovom pravilu, **aposteriorna vjerojatnost** parametara je

$$\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) = \frac{\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})}{\mathrm{p}(\mathbb{D})} = \frac{\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta})}{\int \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}')\,\mathrm{p}(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'}. \tag{3.15}$$

Maksimizacijom aposteriorne vjerojatnosti dobivaju se parametri

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})\,\mathrm{p}(\boldsymbol{\theta}). \tag{3.16}$$

Ovdje nije potrebno normalizirati aposteriornu vjerojatnost izračunavanjem **marginalne izglednosti** (engl. *marginal likelihood*, *evidence*) $\mathrm{p}(\mathbb{D})$ u nazivniku na desnoj strani jednadžbe (3.15) jer ona ne ovisi $\boldsymbol{\theta}$, nego samo o modelu $\mathbb{H}$. Odabirom uniformne (neinformativne) apriorne razdiobe MAP-procjenitelj postaje ekvivalentan ML-procjenitelju.

Poželjno je da $\mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})$ i $\mathrm{p}(\boldsymbol{\theta})$ kao funkcije parametra $\boldsymbol{\theta}$ imaju takav algebarski oblik da njihov umnožak ima sličan oblik i može se analitički izračunati. Ako $\mathrm{p}(\boldsymbol{\theta})$ i $\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D})$ imaju isti algebarski oblik definiran nekim parametrima, nazivaju se **konjugatne razdiobe** (Šnajder i Dalbelo Bašić, 2014).

3.2.5. Bayesovski procjenitelj i zaključivanje

Prethodno opisani procjenitelji daju točkastu procjenu parametara i ne izražavaju nesigurnost procjene kojoj uzrok može biti npr. nedovoljna količina podataka ili šum u podacima za učenje. **bayesovski procjenitelj** kao procjenu daje razdiobu nad hipotezama $\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D})$ za koju je integriranjem po svim mogućim parametrima potrebno izračunati marginalnu izglednost $\mathrm{p}(\mathbb{D}) = \int \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}')\,\mathrm{p}(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'$ iz nazivnika na desnoj strani jednadžbe (3.23).

Kod složenijih modela često ne možemo odabrati konjugatnu apriornu razdiobu, a i funkcija izglednosti je sama po sebi već dovoljno složena da se, neovisno o apriornoj razdiobi, marginalna izglednost $\mathrm{p}(\mathbb{D})$ ne može ni analitički ni numerički traktabilno računati.

Vjerojatnost nekog primjera $\boldsymbol{d}$ procjenjuje se marginalizacijom po parametrima (Neal, 1995):

$$\mathrm{p}(\boldsymbol{d} \mid \mathbb{D}) = \int \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta}), \tag{3.17}$$

gdje je korištena uvjetna nezavisnost $\boldsymbol{d} \perp \mathbb{D} \mid \boldsymbol{\theta}$.

Kada se parametri točkasto procjenjuju, npr. MAP-procjeniteljem, točkasta procjena parametara $\hat{\boldsymbol{\theta}}$ aproksimira cijelu aposteriornu razdiobu, tj. $\mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \approx \delta(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Onda je

$$\mathrm{p}(\boldsymbol{d} \mid \mathbb{D}) \approx \int \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta}) \delta(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathrm{p}(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}). \tag{3.18}$$

Za diskriminativne modele se bayesovsko zaključivanje može izraziti ovako:

$$\begin{aligned} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) &= \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{y} \mid \mathbb{D})}{\mathrm{p}(\boldsymbol{x} \mid \mathbb{D})} \\ &= \frac{\int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta}}{\int \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta}} \\ &= \frac{\mathrm{p}(\boldsymbol{x}) \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta}}{\mathrm{p}(\boldsymbol{x}) \int \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta}}. \end{aligned}$$

Poništavanjem $\mathrm{p}(\boldsymbol{x})$ i integriranjem nazivnika dobiva se

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) = \int \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{3.19}$$

Kod regresije je često, ako pretpostavljamo da pogreška izlaza ima Gaussovu razdiobu, najbolja procjena hipoteze očekivanje po naučenoj razdiobi parametara (Neal, 1995):

$$h(\boldsymbol{x}) = \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} h(\boldsymbol{x}; \boldsymbol{\theta}) = \int h(\boldsymbol{x}; \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \mathrm{d}\boldsymbol{\theta}. \tag{3.20}$$

U tom slučaju se nesigurnost može izraziti disperzijom $\mathbf{D}_{\boldsymbol{\theta} \mid \mathbb{D}}\, h(\boldsymbol{x}; \boldsymbol{\theta})$.

3.3. Monte Carlo aproksimacija

Ovaj odjeljak se temelji na Goodfellow et al. (2016).

Monte Carlo aproksimacija je postupak procjenjivanja vrijednosti koje se mogu izraziti kao očekivanje neke funkcije neke slučajne varijable na temelju uzoraka. Ponekad nije moguće analitički ili numerički traktabilno ili efikasno izračunati neki integral (ili zbroj). Ako se on može ovako izraziti:

$$s = \int \mathrm{p}(x) f(x) \mathrm{d}x = \mathbf{E}\, f(X), \tag{3.21}$$

može se procijeniti uzorkovanjem:

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{3.22}$$

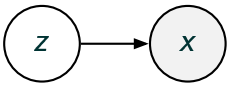
Procjenitelj $\hat{s}_n$ je nepristran ako su x_i nezavisne i imaju istu razdiobu kao x i valjan ako su varijance $f(x_i)$ ograničene. Vrijedi $\mathbf{D}\, \hat{s}_n = \frac{1}{n} \mathbf{D}\, f(x)$.

U širem smislu, postupci *Monte Carlo* obuhvaćaju i generiranje uzoraka slučajne varijable čije se očekivanje procjenjuje.

3.4. Aproksimacija razdioba i aproksimacijsko zaključivanje

Ovaj odjeljak se uglavnom temelji na Blei et al. (2017) i malo na Yang (2017).

Važan problem u bayesovskoj statistici, gdje se zaključivanje temelji na izračunima koji uključuju aposteriornu razdiobu, je aproksimacija razdioba koje su zahtjevne za računanje. Kod složenijih Bayesovskih modela aposteriorna razdioba se ne može lako izračunati i treba koristiti aproksimacijske postupke od kojih su glavni **varijacijski** postupci (Jordan et al., 1999) i postupci **Monte Carlo** aproksimacije s uzorkovanjem pomoću **Markovljevog lanca** (MCMC, engl. *Markov chain Monte Carlo*). MCMC-postupci temelje se na definiranju stohastičkog procesa koji ima stacionarnu razdiobu jednaku razdiobi koja se aproksimira, omogućuju asimptotski egzaktno uzorkovanje velike klase razdioba. Varijacijski postupci temelje se na aproksimaciji razdiobe nekom jednostavnijom koja se pronalazi rješavanjem optimizacijskog problema, brži su i jednostavniji za ostvariti za složenije modele.



Slika 3.4: Prikaz grafičkog modela sa skrivenom varijablom z i opažanom varijablom x.

Razmatramo bayesovski model koji ima latentnu varijablu z i vidljivu varijablu x. Model je prikazan na slici 3.4 i opisan je ovom jednadžbom združene vjerojatnosti:

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{z}) = \mathrm{p}(\boldsymbol{z})\,\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{z}).$$

Zaključivanjem se određuje aposteriorna razdioba latentne varijable:

$$\mathrm{p}(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{z})}{\mathrm{p}(\boldsymbol{x})} = \frac{\mathrm{p}(\boldsymbol{x}, \boldsymbol{z})}{\int \mathrm{p}(\boldsymbol{x}, \boldsymbol{z})\mathrm{d}\boldsymbol{z}}. \tag{3.23}$$

na temelju opažanih vrijednosti slučajne varijable x (podataka). Kod složenijih modela integriranje marginalne izglednosti u nazivniku nije traktabilno i aposteriorna razdioba se mora aproksimirati **približnim (aproksimacijskim) zaključivanjem**.

3.5. Postupci uzorkovanja

TODO

1.3.1 Neal (1995).

3.6. Varijacijsko zaključivanje

Za razliku od uzorkovanja kod MCMC-postupaka, osnovna ideja kod varijacijskog zaključivanja je optimizacija. Prvo se odabire familija razdioba $\mathbb{Q} = \{\mathrm{p}(\tilde{\boldsymbol{z}})\}_{\tilde{\boldsymbol{z}}} = \left\{q_{\boldsymbol{\phi}}\right\}_{\boldsymbol{\phi}}$ koje su lakše za računanje. Razdiobe iz $\mathbb{Q}$ su parametrizirane tzv. **varijacijskim parametrima** $\boldsymbol{\phi}$. Cilj je na temelju podataka kao zamjenu za aposteriornu razdiobu $\mathrm{p}(\boldsymbol{z} \mid \boldsymbol{x})$ pronaći razdiobu iz $\mathbb{Q}$ koja ju što bolje aproksimira. To možemo ostvariti minimizacijom Kullback-Leiblerove (KL) divergenciju s obzirom na stvarnu aposteriornu razdiobu po varijacijskim parametrima $\boldsymbol{\phi}$:

$$q^* = \underset{\mathrm{p}(\tilde{\boldsymbol{z}}) \in \mathbb{Q}}{\arg\min}\, \mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\boldsymbol{z} \mid \boldsymbol{x})). \tag{3.24}$$

Naziv **varijacijsko zaključivanje** dolazi od varijacijskog računa[2], gdje se koriste varijacije, tj. male promjene u funkcijama i funkcionalima, kako bi se pronašli minimumi ili maksimumi funkcionala, preslikavanja iz skupa funkcija u $\mathbb{R}$, koji su često izraženi kao integrali koji uključuju funkcije i njihove derivacije.

Neka je q oznaka funkcije gustoće vjerojatnosti aproksimacijske razdiobe: $q \coloneqq p_{\tilde{z}}$. Ako ciljnu funkciju ovako izrazimo:

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel (z \mid \boldsymbol{x})) &= \mathop{\mathbf{E}}_{\tilde{z}} \ln \frac{q(\tilde{\boldsymbol{z}})}{p_{z|x}(\tilde{\boldsymbol{z}})} \\ &= \mathop{\mathbf{E}}_{\tilde{z}} \ln q(\tilde{\boldsymbol{z}}) - \mathop{\mathbf{E}}_{\tilde{z}} \ln \mathrm{p}(z = \tilde{\boldsymbol{z}}, \boldsymbol{x}) + \ln \mathrm{p}(\boldsymbol{x}), \end{aligned} \tag{3.25}$$

vidi se da se ona ne može lako izračunati jer zahtijeva računanje marginalne izglednosti $\mathrm{p}(\boldsymbol{x})$ iz nazivnika u jednadžbi (3.23) marginalizacijom po z. Marginalna izglednost ne ovisi o varijacijskim parametrima pa ju možemo zanemariti i maksimiziramo funkciju koja se naziva **varijacijska donja granica** (engl. *variational lower bound*) ili **donja granica (logaritma) marginalne izglednosti** (engl. *(log) evidence lower bound*, *ELBO*):

$$L_{\boldsymbol{x}}(\tilde{z}) \coloneqq \ln \mathrm{p}(\boldsymbol{x}) - \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel (z \mid \boldsymbol{x})) = \mathop{\mathbf{E}}_{\tilde{z}} \ln \mathrm{p}(z = \tilde{\boldsymbol{z}}, \boldsymbol{x}) - \mathop{\mathbf{E}}_{\tilde{z}} \ln q(\tilde{\boldsymbol{z}}). \tag{3.26}$$

Ona se može i ovako izraziti:

$$L_{\boldsymbol{x}}(\tilde{z}) = \mathop{\mathbf{E}}_{\tilde{z}} \ln \mathrm{p}(\boldsymbol{x} \mid z = \tilde{\boldsymbol{z}}) - \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel z). \tag{3.27}$$

Maksimiziranje takve ciljne funkcije s obzirom na varijacijske parametre daje razdiobu $q^* = \mathrm{p}(\tilde{z}^*)$ koja dobro objašnjava podatke jer se potiče veće očekivanje logaritma izglednosti (prvi član), a ne razlikuje se previše od apriorne razdiobe jer se potiče manja KL-divergencija između varijacijske razdiobe i apriorne razdiobe (Gal i Ghahramani, 2015).

Naziv *donja granica marginalne izglednosti* dolazi od toga što su Jordan et al. (1999) izveli nejednakost $\ln \mathrm{p}(\boldsymbol{x}) \geq L_{\boldsymbol{x}}(\tilde{z})$ preko Jensenove nejednakosti. Ta nejednakost slijedi i iz prethodne jednadžbe i nenegativnosti KL-divergencije:

$$\ln \mathrm{p}(\boldsymbol{x}) = L_{\boldsymbol{x}}(\tilde{z}) + \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel (z \mid x)) \geq L_{\boldsymbol{x}}(\tilde{z}). \tag{3.28}$$

[2] https://en.wikipedia.org/wiki/Calculus_of_variations

3.6.1. Metoda polja sredina

Dodatno pojednostavljenje koje pomaže u modeliranju i optimizaciji je pretpostavljanje nezavisnosti između latentnih varijabli. Onda za varijacijsku razdiobu vrijedi ovakva faktorizacija:

$$q(\tilde{\boldsymbol{z}}) = \prod_i q_i(\tilde{z}_i), \tag{3.29}$$

gdje su q_i funkcije gustoće pojedinih slučajnih varijabli, a $\tilde{z}_i = \tilde{\boldsymbol{z}}_{[i]}$. Kod **metode polja sredina** pretpostalja se takva razdioba i obično se primjenjuje koordinatni spust za optimizaciju, s čime ima veze ime metode. To je detaljnije objašnjeno u Murphy (2012).

4. Nadzirano strojno učenje

Ovo poglavlje se uglavnom temelji na Šnajder i Dalbelo Bašić (2014) i Goodfellow et al. (2016).

Zadatak algoritama **nadziranog učenja** je preslikavanje **ulaznih primjera** $\boldsymbol{x}$ iz **ulaznog prostora** $\mathbb{X}$ u **izlaze** (**oznake**) $\boldsymbol{y} \in \mathbb{Y}$ na temelju konačnog skupa označenih primjera $\mathbb{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_i$. Algoritmima strojnog učenja pretražuje se **model** ili **prostor hipoteza** u cilju pronalaska **hipoteze** koja osim primjera iz skupa za učenje, u izlaze dobro preslikava i primjere koji nisu u skupu za učenje. Sposobnost postizanja dobre performanse na neviđenim primjerima naziva se **generalizacija**.

Neka je $\mathbb{D} = \{\boldsymbol{d}_i\}_i$ skup nezavisnih primjera izvučenih iz neke razdiobe $\mathcal{D}$. Možemo definirati **probabilistički model** $\mathbb{H}$ s nepoznatim parametrima $\boldsymbol{\theta}$ kojemu je cilj što bolje modelirati tu razdiobu pronalaskom najbolje hipoteze na temelju podataka: $\mathrm{p}(\boldsymbol{d} \mid \mathbb{D}, \mathbb{H})$. Model koji modelira razdiobu primjera nazivamo **generativnim modelom**. U nastavku ćemo izostavljati oznaku modela radi kraćeg zapisa.

Ako su primjeri parovi $\boldsymbol{d}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{X} \times \mathbb{Y}$, može nam biti cilj ulaznim primjerima iz $\mathbb{X}$ dodjeljivati oznake iz $\mathbb{Y}$. Ako je problem koji rješavamo dodjeljivanje oznaka ulaznim primjerima, onda su često prikladniji **diskriminativni modeli**. Probabilistički diskriminativni modeli izravno modeliraju uvjetne razdiobe $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$ hipotezom koja ulazni primjer $\boldsymbol{x}$ preslikava u razdiobu $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D})$. Neprobabilistički diskriminativni modeli modeliraju funkciju dodjeljivanja oznaka hipotezom $h \colon \mathbb{X} \to \mathbb{Y}$. Modeliranje zajedničke razdiobe $\mathrm{p}(\boldsymbol{x}, \boldsymbol{y})$ obično zahtijeva više računalnih resursa i podataka (Bishop, 2006).

Modeli se još mogu podijeliti na **parametarske** i **neparametarske**. Kod parametarskih modela broj parametara je unaprijed određen, dok kod neparametarskih on ovisi o podacima za učenje.

4.1. Induktivna pristranost

Uz zadani skup hipoteza koji dopušta model, **algoritam strojnog učenja** traži parametre koji definiraju jednu hipotezu. Učenje hipoteze je loše definiran (engl. *ill-posed*) problem jer skup podataka $\mathbb{D}$ nije dovoljan za jednoznačan odabir hipoteze. Osim dobrog opisivanja podataka za učenje, naučena hipoteza mora dobro generalizirati. Kako bi učenje i generalizacija bili mogući, potreban je skup pretpostavki koji se naziva **induktivna pristranost**. Razlikujemo dvije vrste induktivne pristranosti (Šnajder i Dalbelo Bašić, 2014):

1. **pristranost ograničavanjem** ili **pristranost jezika** – ograničavanje skupa hipoteza koje se mogu prikazati modelom,
2. **pristranost preferencijom** ili **pristranost pretraživanja** – dodjeljivanje različitih prednosti različitim hipotezama.

Većina algoritama strojnog učenja kombinira obje vrste induktivne pristranosti.

4.2. Komponente algoritma strojnog učenja

Prema Šnajder i Dalbelo Bašić (2014), kod većine algoritama strojnog učenja možemo razlikovati 3 osnovne komponente, od kojih prva predstavlja pristranost ograničavanjem, a druge dvije obično pristranost preferencijom:

1. **Model** ili prostor hipoteza. Model $\mathbb{H}$ je skup funkcija h parametriziranih parametrima $\boldsymbol{\theta}$: $\mathbb{H} = \{h(\boldsymbol{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$.

2. **Funkcija pogreške** ili ciljna funkcija. Funkcija pogreške $E(\boldsymbol{\theta}, \mathbb{D})$ na temelju parametara modela (hipoteze) i skupa podataka izračunava broj koji izražava procjenu dobrote hipoteze. Obično pretpostavljamo da su primjeri iz skupa za učenje nezavisni i definiramo **funkciju gubitka** $L\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$, kojoj je prvi parametar predikcija hipoteze, a drugi ciljna oznaka koja odgovara ulaznom primjeru. Funkciju pogreške možemo definirati kao prosječni gubitak na skupu za učenje:

$$E(\boldsymbol{\theta}, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{D}} L(\boldsymbol{y}, h(\boldsymbol{x}; \boldsymbol{\theta})). \tag{4.1}$$

   Obično joj dodajemo **regularizacijski** član kojim unosimo dodatne

pretpostavke radi postizanja bolje generalizacije. Više o funkciji pogreške u smislu smanjivanja empirijskog i strukturnog rizika piše u odjeljku 4.5.

3. **Optimizacijski postupak**. Optimizacijski postupak je algoritam kojim pronalazimo hipotezu koja minimizira pogrešku:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}, \mathbb{D}). \tag{4.2}$$

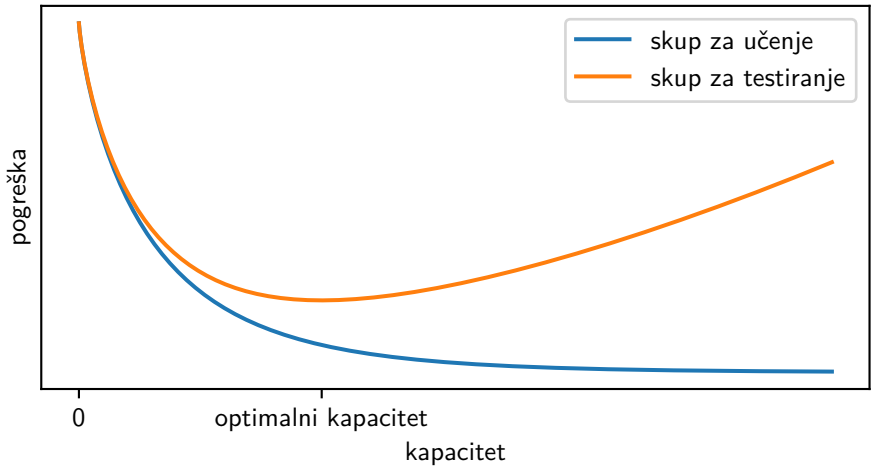
   Kod nekih jednostavnijih modela minimum možemo odrediti analitički. Inače moramo koristiti neki iterativni optimizacijski postupak. Kod nekih složenijih modela, kao što su neuronske mreže, funkcija pogreške nije unimodalna i vjerojatnost pronalaska globalnog optimuma je zanemariva, ali ipak se mogu pronaći dobra rješenja.

U literaturi riječ *model* često ima šire značenje. Uz skup hipoteza obično obuhvaća i induktivnu pristranost ili dio nje. Model u tom smislu bi se formalno mogao definirati kao par $(\mathbb{H}, B)$, gdje je $\mathbb{H}$ skup mogućih hipoteza, a B induktivna pristranost koja hipotezama dodjeljuje različite važnosti. Ovdje će se u nastavku koristiti takvo značenje riječi *model*, a riječ *prostor hipoteza* će se koristiti sa značenjem modela u užem smislu.

4.3. Kapacitet modela, podnaučenost i prenaučenost

Cilj algoritama strojnog učenja je postići malu **pogrešku generalizacije**, tj. malo očekivanje pogreške na primjera koji nisu korišteni za učenje i odabir modela. Generalizacijska pogreška se procjenjuje pogreškom na **skupu za testiranje**. Obično pretpostavljamo da su skupovi primjera koje koristimo za učenje, odabir modela i testiranje generirani međusobno nezavisno i iz iste razdiobe.

Kapacitet ili složenost modela je svojstvo koje opisuje njegovu sposobnost prilagodbe podacima. Model koji se previše prilagođava podacima za učenje (i statističkom šumu u njima) obično ima slabu prediktivnu moć. Treba odabrati model (ili hipotezu) koji dobro objašnjava podatke, ali nije previše složen. O tome govori i načelo **Occamove oštrice** prema kojemu među hipotezama konzistentnima s opažanjem treba odbaciti sve osim najjednostavnije od njih. Postoje formalizacije Occamove oštrice (Blumer et al., 1987, 1989; Grünwald, 2005; Rathmanner i



Slika 4.1: Ovisnost pogrešaka na skupovima za učenje i testiranje o kapacitetu modela. Povećavanjem kapaciteta povećava se razlika između pogrške ne skupu za testiranje i pogreške na skupu za učenje.

Hutter, 2011). Na ograničavanje složenosti modela možemo utjecati ograničavanjem prostora hipoteza i regularizacijom (*mekim* ograničavanjem).

Model s većim kapacitetom (složeniji model) može postići manju pogrešku na skupu za učenje. Prevelik kapacitet povećava pogrešku generalizacije. Za model koji daje veliku pogrešku generalizacije kažemo da je **prenaučen**. Kod takvog modela hipoteze će jako varirati u ovisnosti o skupu za učenje i zato kažemo da složeni modeli imaju visoku varijancu. Model premalog kapaciteta (prejednostavan model) ima manju razliku između pogreške na skupu za učenje i pogreške na skupu za testiranje, ali su obje pogreške veće od optimalnih. Za model koji ne postiže malu pogrešku na skupu za učenje kažemo da je **podnaučen**. U jednostavan model ugrađene su jače pretpostavke i kažemo da on ima jaču pristranost. Uobičajena ovisnost pogrešaka na skupovima za učenje i testiranje o kapacitetu ilustrirana je slikom 4.1.

4.4. Odabir modela

TODO

Murray i Ghahramani (2005) MacKay

4.5. Funkcija pogreške

Dijelovi ovog odjeljka temelje se na (Murphy, 2012).

4.5.1. Rizik i empirijski rizik

Zadatak nadziranog strojnog učenja može se formulirati kao optimizacijski problem minimizacije **rizika**. Neka su $\boldsymbol{\theta}$ odabrani parametri. Definiramo **funkciju gubitka** $L\colon \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ koja kažnjava neslaganje izlaza sa stvarnom oznakom. **Rizik** definiramo kao očekivanje funkcije gubitka:

$$R(\boldsymbol{\theta}; \mathcal{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} L(\boldsymbol{y}, h(\boldsymbol{x}; \boldsymbol{\theta})). \tag{4.3}$$

Razdioba koja generira podatke nije poznata pa se koristi **empirijski rizik** koji **prirodnu razdiobu** $\mathcal{D}$ procjenjuje **empirijskom razdiobom**, tj. uzorkom $\mathbb{D}$:

$$R_{\mathsf{E}}(\boldsymbol{\theta}; \mathbb{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}} L(\boldsymbol{y}, h(\boldsymbol{x}; \boldsymbol{\theta})) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} L(\boldsymbol{y}, h(\boldsymbol{x}; \boldsymbol{\theta})). \tag{4.4}$$

Što je uzorak veći $\mathbb{D}$, sličniji je prirodnoj razdiobi i procjena rizika je bolja. U slučaju nenadziranog učenja, kada se hipoteza sastoji od kodera E i dekodera D, tj. $h(\boldsymbol{x}; \theta) = E(D(\boldsymbol{x}; \theta); \theta)$, ili generativnog modela, kada je $h(\boldsymbol{x}; \theta) = \mathrm{p}(\boldsymbol{x} \mid \theta)$, gubitak mjeri **pogrešku rekonstrukcije** i izraz za rizik je (Murphy, 2012):

$$R(\boldsymbol{\theta}; \mathcal{D}) = \mathop{\mathbf{E}}_{\boldsymbol{d}\sim\mathcal{D}} L(\boldsymbol{d}, h(\boldsymbol{d}; \boldsymbol{\theta})). \tag{4.5}$$

Kod probabilističkih modela empirijski rizik se može definirati kao **negativni logaritam izglednosti** parametara:

$$R_{\mathsf{E}}(\boldsymbol{\theta}; \mathbb{D}) = -\frac{1}{|\mathbb{D}|} \ln \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta}) = -\frac{1}{|\mathbb{D}|} \sum_{\boldsymbol{d}\in\mathbb{D}} \ln \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta}), \tag{4.6}$$

gdje je korištena pretpostavka međusobne nezavisnosti primjera. Gubitak je onda $L(\boldsymbol{d}, h(\boldsymbol{d}; \boldsymbol{\theta})) = -\ln \mathrm{p}(\boldsymbol{d} \mid \boldsymbol{\theta})$. U slučaju diskriminativnog modela, uz zanemarivanja faktora izglednosti koji ne ovisi o $\boldsymbol{\theta}$ (jednadžba (3.14)), vrijedi $L(\boldsymbol{d}, h(\boldsymbol{x}; \boldsymbol{\theta})) = -\ln \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$. Minimizacija gubitka definiranog kao negativni logaritam izglednosti ekvivalentna je minimizaciji KL-divergencije ili unakrsne entropije (odjeljak 2.2) s obzirom na empirijsku razdiobu. Zbog toga se takav gubitak još naziva **gubitak unakrsne entropije**.

4.5.2. Strukturni rizik i regularizacija

Kada ima malo podataka ili je model previše složen, minimizacija empirijskog rizika dovodi do velike varijance i slabe generalizacije. Procjenitelj koji minimizira empirijski rizik ne uzima u obzir apriornu razdiobu parametara. Radi postizanja bolje generalizacije, funkciji pogreške dodaje se **regularizacijski gubitak** $\lambda R_{\mathsf{R}}(\boldsymbol{\theta})$, $\lambda \geq 0$, koji predstavlja **strukturni rizik** koji daje prednost jednostavnijim hipotezama. Funkcija pogreške onda ima ovakav oblik:

$$E(\boldsymbol{\theta}; \mathbb{D}) = R_{\mathsf{E}}(\boldsymbol{\theta}; \mathbb{D}) + \lambda R_{\mathsf{R}}(\boldsymbol{\theta}). \tag{4.7}$$

Regularizacijski gubitak obično ovisi samo o parametrima, ali može ovisiti i o podacima (Goodfellow et al., 2016).

Ako kao funkciju pogreške koristimo negativni logaritam aposteriorne vjerojatnosti uz pretpostavku međusobne nezavisnosti primjera, funkcija pogreške je

$$E(\boldsymbol{\theta}; \mathbb{D}) = -\frac{1}{|\mathbb{D}|} \ln \mathrm{p}(\boldsymbol{\theta} \mid \mathbb{D}) \tag{4.8}$$

$$= \underbrace{-\frac{1}{|\mathbb{D}|} \ln \mathrm{p}(\mathbb{D} \mid \boldsymbol{\theta})}_{R_{\mathsf{E}}(\boldsymbol{\theta}; \mathbb{D})} \underbrace{-\frac{1}{|\mathbb{D}|} \ln \mathrm{p}(\boldsymbol{\theta})}_{\lambda R_{\mathsf{R}}(\boldsymbol{\theta})} + C_1, \tag{4.9}$$

gdje je $C_1 = \frac{1}{|\mathbb{D}|} \ln \mathrm{p}(\mathbb{D})$ konstanta koja ne ovisi o $\boldsymbol{\theta}$. Hiperparametar λ je onda parametar apriorne razdiobe. Možemo ga ovako izlučiti:

$$\ln \mathrm{p}(\boldsymbol{\theta}) = \lambda \ln p_0(\boldsymbol{\theta}) + C_2 = \ln p_0(\boldsymbol{\theta})^{\lambda} + C_2, \tag{4.10}$$

gdje je $C_2 = -\ln\left(\int_{\boldsymbol{\theta}'} p_0(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'\right)$ konstanta koja ne ovisi o $\boldsymbol{\theta}$. Vidi se da λ određuje koncentraciju apriorne razdiobe. Povećanje λ smanjuje entropiju apriorne razdiobe. Ona postaje koncentriranija i regularizacija jača. Jačom regularizacijom se povećava pristranost i smanjuje varijanca.

4.6. Osnovni zadaci nadziranog učenja

Osnovni zadaci nadziranog učenja su **klasifikacija** i **regresija**. Zadatak klasifikacije je svakom ulaznom primjerima dodjeljivati oznake iz konačnog skupa oznaka, npr. $\{1..C\}$, gdje svaka oznaka predstavlja jednu **klasu** (**razred**). Zadatak regresije je ulaznim primjerima dodjeljivati vrijednosti iz kontinuiranog skupa (obično

$\mathbb{R}$ ili $\mathbb{R}^n$). Ulazni primjeri su obično realni vektori. Klasifikacijska hipoteza se može definirati preko funkcije s kontinuiranom kodomenom. Ako $C = 2$, ta funkcija može biti $h\colon \mathbb{X} \to \mathbb{R}$, a hipoteza $h_{\mathsf{c}}(\boldsymbol{x}) = [\![h(\boldsymbol{x}) > 0]\!]$. Ako $C > 2$, onda to može biti npr. $h_{\mathsf{c}}(\boldsymbol{x}) = \arg\max_i h_i(\boldsymbol{x})$, gdje $h\colon \mathbb{X} \to \mathbb{R}^C$ i $h(\boldsymbol{x}) = [h_i(\boldsymbol{x})]_{i=1..C}^{\mathsf{T}}$. Kod nekih zadataka ulazi ili izlazi imaju složeniju strukturu i ona se može razlikovati između različitih primjera.

4.7. Primjeri modela: poopćeni linearni modeli

Ovaj odjeljak se temelji na (Šnajder i Dalbelo Bašić, 2014).

Linearni modeli su modeli kod kojih je hipoteza definirana afinom transformacijom:

$$h(\boldsymbol{x}) = h(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + b, \tag{4.11}$$

gdje je $\boldsymbol{w}$ vektor **težina**, b **pomak** (engl *bias*), a $\boldsymbol{\theta} = (\boldsymbol{w}, b)$. Kod linearnih modela je, u slučaju klasifikacije, granica $(n-1)$-dimenzionalna hiperravnina s normalom $\boldsymbol{w}$. Obično se na ulazne primjere primjenjuje neka nelinearna transformacija

$$\begin{aligned} \phi\colon \mathbb{R}^n &\to \mathbb{R}^m \\ \boldsymbol{x} &\mapsto [\phi_1(\boldsymbol{x}), .., \phi_m(\boldsymbol{x})]^{\mathsf{T}} \end{aligned}$$

koja predstavlja preslikavanje ulaznog prostora u **prostor značajki**. Funkcije $\phi_i\colon \mathbb{R}^n \to \mathbb{R}$ nazivaju se **bazne funkcije**. Hipoteza linearnog modela onda ima oblik

$$h(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x}). \tag{4.12}$$

Ovdje je izostavljen pomak b jer jedan od izlaza transformacije ϕ može biti konstanta 1 koja se množi s jednom težinom iz $\boldsymbol{w}$.

Poopćeni linearni modeli su modeli kod kojih je hipoteza ovako definirana:

$$h(\boldsymbol{x}) = f(\boldsymbol{w}^{\mathsf{T}}\phi(\boldsymbol{x})). \tag{4.13}$$

U odnosu na linearne modele, oni još imaju **prijenosnu** (**aktivacijsku**) funkciju $f\colon \mathbb{R} \to \mathbb{R}$. Ako je f nelinearna, model je nelinearan u parametrima, ali granica klasifikacijskog modela je i dalje hiperravnina.

Slijedi pregled nekih linearnih modela prema Šnajder (2017) uz oznake

$s = \boldsymbol{w}^\mathsf{T}\phi(\boldsymbol{x})$ i $\boldsymbol{s} = \boldsymbol{W}\phi(\boldsymbol{x})$:

– Linearna regresija:

$$\begin{aligned} h(\boldsymbol{x};\boldsymbol{w}) &= f(s) = s, \\ \mathrm{p}(y \mid \boldsymbol{x},\boldsymbol{w}) &= \mathcal{N}(h(\boldsymbol{x}),\sigma^2)(y), \\ L(y,h(\boldsymbol{x})) &= (h(\boldsymbol{x})-y)^2, \\ \nabla_{\boldsymbol{w}} L(y,h(\boldsymbol{x})) &= (h(\boldsymbol{x})-y)\phi(\boldsymbol{x}), \end{aligned}$$

gdje $y \in \mathbb{R}$.

– Logistička regresija:

$$\begin{aligned} h(\boldsymbol{x};\boldsymbol{w}) = f(s) &= \frac{1}{1+\exp(s)} = \mathrm{P}(y=1 \mid \boldsymbol{x},\boldsymbol{w}), \\ \mathrm{P}(y \mid \boldsymbol{x},\boldsymbol{w}) &= h(\boldsymbol{x})^y (1-h(\boldsymbol{x}))^{1-y}, \\ L(y,h(\boldsymbol{x})) &= -y \ln h(\boldsymbol{x})^y - (1-y)\ln(1-h(\boldsymbol{x})), \\ \nabla_{\boldsymbol{w}} L(y,h(\boldsymbol{x}) &= (h(\boldsymbol{x})-y)\phi(\boldsymbol{x}), \end{aligned}$$

gdje $y \in 0, 1$.

– Višeklasna logistička regresija:

$$\begin{aligned} h(\boldsymbol{x};\boldsymbol{W}) = f(\boldsymbol{s}) &= \frac{1}{\mathbf{1}^\mathsf{T}\exp(\boldsymbol{s})}\exp(\boldsymbol{s}) = [\mathrm{P}(y=k \mid \boldsymbol{x},\boldsymbol{w})]_{k=1..C}^\mathsf{T}, \\ \mathrm{P}(y \mid \boldsymbol{x},\boldsymbol{w}) &= h_y(\boldsymbol{x}) = \prod_k h_k(\boldsymbol{x})^{[\![y=k]\!]}, \\ L(y,h(\boldsymbol{x})) &= -\sum_k [\![y=k]\!] \ln h(\boldsymbol{x})^{[\![y=k]\!]}, \\ \nabla_{\boldsymbol{W}_{[k,:]}^\mathsf{T}} L(y,h_k(\boldsymbol{x})) &= (h_k(\boldsymbol{x})-y_k)\phi(\boldsymbol{x}) \\ \nabla_{\boldsymbol{W}} L(y,h(\boldsymbol{x})) &= \phi(\boldsymbol{x})^\mathsf{T}\Big(h(\boldsymbol{x}) - \mathbf{e}_y\Big), \end{aligned}$$

gdje $y \in 1..C$, $h(\boldsymbol{x}) = [h_k(\boldsymbol{x})]_{k=1..C}^\mathsf{T}$, $h_i \colon \mathbb{R}^n \to \mathbb{R}$, a $\mathbf{e}_k$ označava jednojedinični vektor (vektor kanonske baze): $\mathbf{e}_k = [[\![i=k]\!]]_{i=1..C}^\mathsf{T}$.

Funkcije gubitka su definirane kao negativni logaritam izglednosti, $L(y,h(\boldsymbol{x})) = -\ln \mathrm{P}(y \mid \boldsymbol{x},\boldsymbol{w})$, i konveksne su. Optimalne težine linearne regresije mogu se analitički izračunati, logistička regresija i višeklasna logistička regresija se obično uče optimizacijskim postupcima temeljenim na gradijentu.

Razdiobe $\mathrm{P}(y \mid \boldsymbol{x},\boldsymbol{w})$ poopćenih linearnih modela spadaju u **eksponencijalnu familiju razdioba**. Može se pokazati da je to jedina familija razdioba za koje

postoje konjugatne apriorne razdiobe, što pojednostavljuje računanje aposteriorne razdiobe (Murphy, 2012). Opći oblik ekponencijalne familije i više o njenima svojstvima i svojstvima poopćenih linearnih modela može se naći u Murphy (2012).

5. Duboko učenje i konvolucijske mreže

Klasični (plitki) modeli strojnog učenja (npr. poopćeni linearni modeli) oslanjaju se na kvalitetne značajke, tj. funkciju ϕ koja transformira ulazne primjere u vektore značajki. Za neke zadatke koji uključuju visokodimenzionalne primjere sa složenom strukturom (npr. slike, tekst i zvuk), ručno konstruiranje transformacije koja bi bila dovoljno dobra nije izvedivo, npr. jezgrene metode kod kojih se preslikavanje temelji na pretpostavci sličnosti primjera bliskih u ulaznom prostoru ne generaliziraju dobro. Kod **dubokog učenja** (LeCun et al., 2015; Goodfellow et al., 2016) transformacija ϕ se uči.

Odabirom

$$\phi(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{\theta}_\mathsf{h}) = f(\boldsymbol{W}_\mathsf{h}\boldsymbol{x} + \boldsymbol{b}_\mathsf{h}), \tag{5.1}$$

gdje je $\boldsymbol{W}_\mathsf{h}$ matrica težina, $\boldsymbol{b}_\mathsf{h}$ vektor pomaka, $\boldsymbol{\theta}_\mathsf{h} = (\boldsymbol{W}_\mathsf{h}, \boldsymbol{b}_\mathsf{h})$ a f nelinearna prijenosna (aktivacijska) funkcija koja se primjenjuje na svaki element vektora posebno, dobiva se jednostavna **umjetna neuronska mreža** (ovdje će se koristiti kraći nazivi: *neuronska mreža* ili *mreža*) s jednim **skrivenim slojem** kojemu odgovara funkcija ϕ. Ako to uvrstimo u jednadžbu poopćenog linearnog modela (jednadžba (4.13)):

$$h(\boldsymbol{x}; \boldsymbol{\theta}) = f(\boldsymbol{w}^\mathsf{T} f(\boldsymbol{W}_\mathsf{h}\boldsymbol{x} + \boldsymbol{b}_\mathsf{h}) + b), \tag{5.2}$$

ili, ako je izlaz vektor,

$$h(\boldsymbol{x}; \boldsymbol{\theta}) = f(\boldsymbol{W}_\mathsf{o} f(\boldsymbol{W}_\mathsf{h}\boldsymbol{x} + \boldsymbol{b}_\mathsf{h}) + \boldsymbol{b}_\mathsf{o}), \tag{5.3}$$

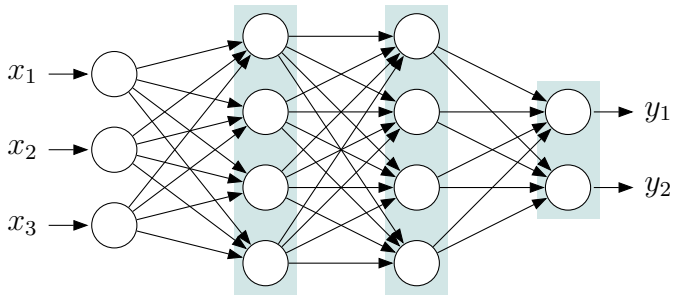
gdje $\boldsymbol{\theta} = (\boldsymbol{W}_\mathsf{h}, \boldsymbol{b}_\mathsf{h}, \boldsymbol{W}_\mathsf{o}, \boldsymbol{b}_\mathsf{o})$. Jedinice neuronske mreže kojima odgovaraju operacije oblika $\boldsymbol{x} \mapsto f(\boldsymbol{w}_i^\mathsf{T}\boldsymbol{x} + b_i)$ nazivaju se **umjetni neuroni**.

Za razliku od modela opisanih u odjeljku 4.7, za ovakav i dublje modele opisane u sljedećim odjeljcima ciljna funkcija nije konveksna (ni unimodalna) pa nije garantirano da će postupak učenja pronaći dobru hipotezu. Empirijski rezultati ipak pokazuju da duboke mreže uz neke prilagodbe uspješno uče i generaliziraju. Algoritmi koji se koriste za učenje modela dubokog učenja temelje se na gradijentnom spustu. Oni su opisani u odjeljku 5.3.

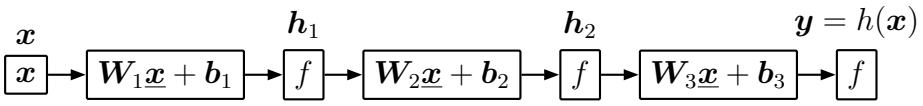
5.1. Duboke unaprijedne mreže

Može se pokazati da model mreže s jednim skrivenim slojem opisan jednadžbom (5.3), ako je dimenzija skrivenog sloja dovoljno velika, može s proizvoljno malom greškom aproksimirati svaku neprekinutu funkciju kojoj je domena konveksni podskup od $\mathbb{R}$. O tome govori teorem o univerzalnoj aproksimaciji (Cybenko, 1989; Leshno et al., 1993). Aktivacijska funkcija mora biti nelinearna jer kompozicija linearnih funkcija je linearna funkcija. Teorem o univerzalnoj aproksimaciji ne govori o tome hoće li takav model generalizirati. Dodavanjem jedinica u skriveni sloj povećava se kapacitet modela.

Obična neuronska mreža može imati više skrivenih slojeva, što se može prikazati kao na slici 5.1 ili slici 5.2. Povećavanjem broja skrivenih slojeva svaka jedinica u nekom sloju može koristiti izlaze svih jedinica prethodnog sloja kao značajke, što mreži omogućuje da, u odnosu na mrežu s 1 skrivenim slojem, s manje jedinica modelira funkcije u kojima postoje uzorci koji se ponavljaju i imaju hijerarhijsku strukturu (Goodfellow et al., 2016). Niži slojevi služe višim slojevima kao značajke transformiranjem kojih se dobivaju značajke više razine.



Slika 5.1: Prikaz primjera troslojne mreže. Svakom bridu odgovara jedna težina (pomaci nisu prikazani). Slojevi su označeni plavim četverokutima. Čvorovi koji su unutar slojeva mreže predstavljaju umjetne neurone. Slika je napravljena na temelju http://www.texample.net/tikz/examples/neural-network/.



Slika 5.2: Prikaz troslojne mreže kao računskog grafa. gdje čvorovi predstavljaju funkcije s parametrima, a bridovi podatke (vektore) čije su oznake prikazane uz neke od čvorova iz kojih izlaze. Funkcije su označene oznakom funkcije (aktivacijska funkcija) ili definicijom (afina transformacija). Ulaz sloja označen je s $\underline{\boldsymbol{x}}$, a oznake varijabli koje pripadaju čvorovima (ulaz u ulaznom čvoru i parametri u čvorovima afine transformacije) nisu podvučene.

Kao prijenosna funkcija skrivenih slojeva često se koristi **zglobnica** (**ReLU**, engl. rectified linear unit) $\mathrm{ReLU}(x) = \max(0, s)$, a prije su češće bile korištene **logistička funkcija**, $\sigma(s) = \frac{\exp(s)}{1+\exp(s)}$, i **tangens hiperbolni**, $\tanh(x) = \frac{\exp(s)-\exp(-s)}{\exp(s)+\exp(-s)}$. U izlaznom sloju obično se koriste funkcije korištene kod poopćenih linearnih modela (odjeljak 4.7) – identitet za regresiju, logistička funkcija za binarnu klasifikaciju, a **softmaks**, $\mathrm{softmax}(\boldsymbol{s}) = \frac{1}{\mathbf{1}^\mathrm{T}\exp(\boldsymbol{s})}\exp(\boldsymbol{s})$, koji kao izlaz daje normalizirani vektor koji predstavlja razdiobu, za višeklasnu klasifikaciju.

Dosad opisivane mreže nazivaju se **unaprijedne mreže** (engl. *feedforward networks*) zato što se pri izračunu informacija propagira od ulaza prema izlazu, bez povratnih veza. Za mrežu kažemo da je duboka ako ime veći broj slojeva. Struktura duboke unaprijedne mreže ne mora se sastojati samo od niza afinih transformacija i nelinearnosti. Općenito, mrežu možemo predstaviti **računskim grafom**, tj. usmjerenim acikličkim grafom kod kojega čvorovi predstavljaju proizvoljne računske operacije i njihove izlaze, a bridovi označavaju izlazi kojih čvorova su ulazi kojih čvorova. Takav graf se naziva **računski graf**. Čvorovi koji nemaju roditelje su varijable koje čine ulazi i parametri Parametri se mogu dijeliti, tj. mogu biti ulaz većem broju čvorova. Čvorovi koji nemaju djecu su izlazi (ili gubitak). Na slici 5.2 je prikazan jedan takav graf, ali na njemu, radi sažetosti, parametri nemaju zasebne čvorove, nego su označeni unutar čvorova koji o njima ovise.

5.2. Konvolucijske mreže

5.3. Učenje

5.3.1. Algortiam propagacije pogreške unatrag

5.3.2. Optimizacijski algoritmi

5.3.3. Algoritam propagacija pogreške unatrag

5.3.4. Isključivanje neurona - dropout

5.3.5. Normalizacija po grupama

6. Procjenjivanje nesigurnosti

6.1. Aleatorna i epistemička nesigurnost

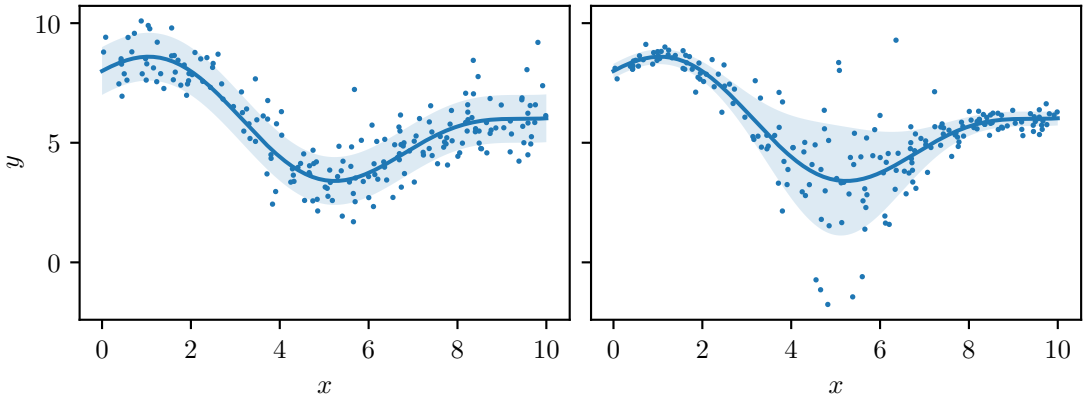
Kod bayesovskih modela nesigurnost zaključivanja izražava se razdiobom po vrijednostima varijable čija vrijednost se procjenjuje, a može se izraziti i entropijom ili varijancom kada je prikladno.

Postoje različiti izvori nesigurnosti (C. Kennedy i O'Hagan, 2002), ali nesigurnost općenito možemo podijeliti na dvije vrste: **aleatornu nesigurnost** i **epistemičku nesigurnost** (Kiureghian i Ditlevsen, 2009). Riječ *aleatorna* izvedena je vjerojatno od latinske riječi *aleator* (Gal, 2016) koja znači *kockar*, a riječ *epistemička* izvedena je od grčke riječi *epistēmē* koja znači *znanje*. Aleatorna nesigurnost je nesigurnost koju model ne može smanjiti neovisno o znanju i količini dostupnih podataka. Ona dolazi od nedeterminizma samog procesa koji generira podatke, nedostupnosti dijela informacija ili ograničenja modela. Epistemička nesigurnost je nesigurnost u strukturu i parametre modela (Gal, 2016). Ona dolazi od neznanja i može se smanjiti uz više podataka.

Granica između aleatorne i epistemičke nesigurnosti ovisi o modelu. Nešto što je kod jednostavnijeg modela aleatorna nesigurnost, kod složenijeg modela može biti epistemičkog karaktera. Ako su neke pojave po prirodi nasumične ili se ne mogu ili ne žele modelu dati informacije koje bi ih mogle objasniti, nesigurnost zaključivanja u vezi tih pojava će, neovisno o ograničenosti modela, biti aleatorna.

6.2. Homoskedastička i heteroskedastička nesigurnost

Aleatorna nesigurnost može biti homoskedastička i heteroskedastička.



Slika 6.1: Homoskedastički (lijevo) i heteroskedastički (desno) Gaussov šum. Crta prikazuje očekivanje $f(x)$, svjetloplava površina standardnu devijaciju šuma $s(x)$, a točke slučajne uzorke. Točke su generirane prema $(y \mid x) \sim \mathcal{N}(f(x), s(x)^2)$. Na lijevoj slici je $s(x) = 1$.

TODO: homoskedastička, heteroskedastička nesigurnost

TODO: model uncertainty, predictive uncertainty Gal-thesis 1.2

5.33331pt

11.74983pt

small 10.95pt

footnotesize 10.0pt

412.56497pt

7. Bayesovske neuronske mreže

8. Procjenjivanje nesigurnosti kod konvolucijskih mreža

9. Eksperimentalni rezultati

9.1. Programska izvedba

9.2. Skupovi podataka

10. Zaključak

Zaključak.

Literatura

Ethem Alpaydin. **Introduction to Machine Learning**. 2014.

Christopher M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. 2006.

David M. Blei, Alp Kucukelbir, i Jon D. McAuliffe. Variational Inference: A Review for Statisticians. **Journal of the American Statistical Association**, 2017. URL http://arxiv.org/abs/1601.00670.

Alselm Blumer, Andrzej Ehrenfeucht, David Haussler, i Manfred K. Warmuth. Occam's razor. **Inf. Process. Lett.**, 24(6):377–380, Travanj 1987. ISSN 0020-0190. doi: 10.1016/0020-0190(87)90114-1. URL http://dx.doi.org/10.1016/0020-0190(87)90114-1.

Anselm Blumer, A. Ehrenfeucht, David Haussler, i Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. **J. ACM**, 36(4):929–965, Listopad 1989. ISSN 0004-5411. doi: 10.1145/76359.76371. URL http://doi.acm.org/10.1145/76359.76371.

Marc C. Kennedy i Anthony O'Hagan. Bayesian calibration of computer models. 2002.

G. Cybenko. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems (MCSS)**, stranice 303–314, 1989. ISSN 0932-4194. doi: 10.1007/BF02551274. URL http://dx.doi.org/10.1007/BF02551274.

Neven Elezović. **Vjerojatnost i statistika: Slučajne varijable**. 2007.

Yarin Gal. **Uncertainty in Deep Learning**. Doktorska disertacija, University of Cambridge, 2016.

Yarin Gal i Zoubin Ghahramani. Dropout as a Bayesian Approximation: Appendix. 2015. URL https://arxiv.org/abs/1506.02157.

Ian Goodfellow, Yoshua Bengio, i Aaron Courville. **Deep Learning**. MIT Press, 2016. http://www.deeplearningbook.org.

Peter Grünwald. A tutorial introduction to the minimum description length principle. U **Advances in Minimum Description Length: Theory and Applications**, 2005.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, i Lawrence K. Saul. An introduction to variational methods for graphical models. 1999.

Armen Der Kiureghian i Ove Ditlevsen. Aleatory or epistemic? Does it matter? 2009.

Yann LeCun, Yoshua Bengio, i Geoffrey E. Hinton. Deep learning. **Nature**, 521 (7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, i Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. **Neural Networks**, stranice 861–867, 1993. URL http://dblp.uni-trier.de/db/journals/nn/nn6.html#LeshnoLPS93.

Kevin P. Murphy. **Machine Learning: A Probabilistic Perspective**. 2012.

Iain Murray i Zoubin Ghahramani. A note on the evidence and Bayesian Occam's razor. 2005.

Jan Šnajder. Strojno učenje: 7. logistička regresija ii, 2017. URL http://www.fer.unizg.hr/_download/repository/SU-2017-07-LogistickaRegresija2.pdf.

Jan Šnajder i Bojana Dalbelo Bašić. **Strojno učenje**. 2014.

Radford M. Neal. Bayesian learning for neural networks, 1995.

Christopher Olah. Visual information theory, 2015. URL http://colah.github.io/posts/2015-09-Visual-Information/.

Samuel Rathmanner i Marcus Hutter. A philosophical treatise of universal induction. **CoRR**, abs/1105.5721, 2011. URL http://arxiv.org/abs/1105.5721.

Xitong Yang. Understanding the Variational Lower Bound, 2017. URL http://legacydirs.umiacs.umd.edu/~xyang35/files/understanding-variational-lower.pdf.

Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.