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, θ	Slučajna varijabla
$\boldsymbol{\mathit{a}}$, $\boldsymbol{\theta}$	Slučajni vektor ili niz
$\boldsymbol{\mathit{A}}$, $\boldsymbol{\Theta}$	Slučajna matrica ili višedimenzionalni niz
$\mathbb{A}$	Slučajni skup ili multiskup
a, riječ	Oznaka koja ne predstavlja matematički objekt

Konstante

$\{\}$	Prazni skup
e	Konstanta za koju vrijedi $\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{e}^x = \mathrm{e}^x$
π	Opseg kruga promjera 1

$\mathbf{0}$	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
$(a_i)_i, (a_{i,j})_{i,j}, (a_{i,j,k})_{i,j,k}$	Višedimenzionalni niz s implicitnim ili neodređenim brojem elemenata
(a, b)	Otvoreni interval
$[a, b]$	Zatvoreni interval
$[x_1, .., x_n]$	Vektor redak

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 $\boldsymbol{a}_{[i]} = a_i$. Indeksi kod n-dimenzionalnih nizova mogu biti i vektori iz $\mathbb{N}^n$, ili kombinacije vektora manje dimenzije sa skalarima. Ako nije navedeno drugačije, najmanji indeks je 1.

$a^{\mathrm{g}}_{\mathrm{d}}$	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 $\boldsymbol{a}_{[i_1]}, \boldsymbol{a}_{[i_1+1]}, .., \boldsymbol{a}_{[i_2]}$
$\boldsymbol{a}_{[(i_1..i_n)]}$	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 $\boldsymbol{A}$
$\boldsymbol{A}_{[\boldsymbol{i}]}$	Element $\boldsymbol{A}_{[\boldsymbol{i}_{[1]},..,\boldsymbol{i}_{[n]}]}$ n-dimenzionalnog niza
$\boldsymbol{A}_{[\boldsymbol{i}_1:\boldsymbol{i}_2]}$	Podniz $\boldsymbol{A}_{[\boldsymbol{i}_{1\,[1]}:\boldsymbol{i}_{2\,[1]},..,\boldsymbol{i}_{1\,[n]}:\boldsymbol{i}_{2\,[n]}]}$ n-dimenzionalnog niza
$\boldsymbol{A}_{[\boldsymbol{i}_1:\boldsymbol{i}_2,:]}$	Podniz $\boldsymbol{A}_{[\boldsymbol{i}_{1\,[1]}:\boldsymbol{i}_{2\,[1]},..,\boldsymbol{i}_{1\,[n-1]}:\boldsymbol{i}_{2\,[n-1]},:]}$ n-dimenzionalnog niza

Operacije linearne algebre i operacije s nizovima

$\langle \boldsymbol{a} \vert \boldsymbol{b} \rangle$, $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{b}$	Skalarni produkt
$\boldsymbol{a}\boldsymbol{b}^{\mathsf{T}}$	Vanjski produkt
$\boldsymbol{a} \odot \boldsymbol{b}$	Umnožak po elementima; Hadamardov produkt
$\boldsymbol{a} \oslash \boldsymbol{b}$	Dijeljenje po elementima
$\boldsymbol{a}^{\odot b}$	Potenciranje po elementima; Hadamardovo potenciranje
$\boldsymbol{A}\boldsymbol{B}$	Matrično množenje (kompozicija linearnih operatora)
$\boldsymbol{A}^{-1}$	Inverz matrice (inverz linearnog operatora)
$\boldsymbol{A}^{\mathsf{T}}$	Transponiranje
$\mathrm{diag}(\boldsymbol{a})$	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 elementima
$\boldsymbol{a} +\!\!+ \boldsymbol{b}$	Konkatenacija vektora (stupaca) $\boldsymbol{a} \in \mathbb{R}^n$ i $\boldsymbol{b} \in \mathbb{R}^m$ u vektor iz $\mathbb{R}^{n+m}$
$\boldsymbol{A} +\!\!+ \boldsymbol{B}$	Konkatenacija višedimenzionalnih nizova po prvoj dimenziji
$\boldsymbol{A} +\!\!+' \boldsymbol{B}$	Konkatenacija višedimenzionalnih nizova po zadnjoj dimenziji
$\mathrm{vec}(\boldsymbol{A})$	Funkcija koja preslikava niz iz $\mathbb{R}^{d_1 \times \cdots \times d_n}$ u vektor iz $\mathbb{R}^{d_1 \ldots d_n}$
$\dim(\boldsymbol{a})$	Dimenzija vektora
$\dim(\boldsymbol{A})$	Vektor dimenzija $[d_1, .., d_n]$ ako $\boldsymbol{A} \in \mathbb{R}^{d_1 \times \cdots \times d_n}$

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 $x \in \mathcal{A}$
$\int f(x)\,\mathrm{d}x$, $\int_x f(x)$	Određeni integral funkcije $f(x)$ po $x \in \mathcal{A}$, gdje je $\mathcal{A}$ 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
$\mathcal{A}$	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 matematičke oznake

$\mathbb{A} \to \mathbb{B}$	Skup funkcija s domenom $\mathbb{A}$ i kodomenom $\mathbb{B}$
$f\colon \mathbb{A} \to \mathbb{B}$	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
$f + g$	Zbroj funkcija
fg	Umnožak funkcija
$f * g$	Konvolucija funkcija
$\langle f \vert g \rangle$	Skalarni produkt funkcija
$\vert\mathbb{A}\vert$	Kardinalitet skupa
$\delta(\cdot)$	Diracova delta
$[\![\cdot]\!]$	Iversonova uglata zagrada; $[\![P]\!] = \begin{cases} 1, & P \equiv \top \\ 0, & P \equiv \bot \end{cases}$
$\mathrm{mod}(a, b)$	Ostatak pri dijeljenju cijelih brojeva; $\mathrm{mod}(a, b) \coloneqq a - \lfloor a/b \rfloor b$

Fraze

dimenzija vektora	Broj komponenata ili kardinalitet baze vektorskog prostora
n-dimenzionalni vektor	Vektor s dimenzijom n
i-ta komponenta vektora $\boldsymbol{a}$	$\boldsymbol{a}_{[i]}$
n-dimenzionalni niz	Niz (engl. *array*) iz $\mathbb{R}^{d_1 \times \cdots \times d_n}$, tj. postoji $f\colon \{1..d_1\} \times \cdots \times \{1..d_n\} \to \mathbb{R}$ tako da za svaku n-torku $\boldsymbol{i}$ iz njene domene vrijedi $\boldsymbol{A}_{[\boldsymbol{i}]} = f(\boldsymbol{i})$
i-ta dimenzija niza	d_i, ako je niz iz $\mathbb{R}^{d_1 \times \cdots \times d_n}$

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 i 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 se označava $\mathrm{P}(\{x = x\})$, $\mathrm{P}(x = x)$ ili $\mathrm{P}(x)$. **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

$$P_x : \mathbb{X} \longrightarrow [0, 1]$$
$$x \longmapsto \mathrm{P}(x = x)$$

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*)

$$p_x : \mathbb{X} \longrightarrow [0, \infty)$$
$$x \longmapsto \mathrm{p}(x)$$

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

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

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

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

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:

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

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

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

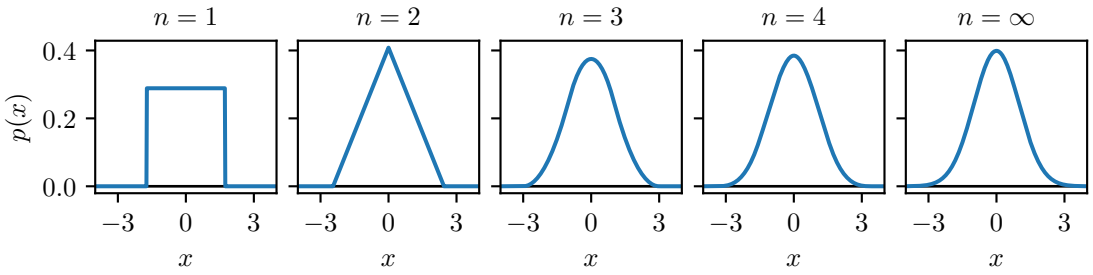
Laplaceova razdioba je kontinuirana razdioba definirana parametrima $\beta \in \mathbb{R}_{>0}$ i $\mu \in \mathbb{R}$ i ima ova svojstva:

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

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

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

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.

Višedimenzionalna Gaussova razdioba je kontinuirana razdioba definirana parametrima $\boldsymbol{\mu} \in \mathbb{R}^n$ i pozitivno definitnom matricom $\boldsymbol{\Sigma}$ i ima ova svojstva:

$$\mathrm{p}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\right) \tag{2.40}$$

$$\mathbf{E}\, \boldsymbol{x} = \boldsymbol{\mu}, \tag{2.41}$$

$$\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}. \tag{2.42}$$

Ako su $x_{[i]}$ nezavisne, $\boldsymbol{\Sigma}$ će biti dijagonalna matrica, ali mora vrijediti obrnuto.

2.2. Teorija informacije

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

$$\mathrm{I}(x \in A) \coloneqq \log_b \frac{1}{\mathrm{P}(x \in A)} = -\log_b \mathrm{P}(x \in A). \tag{2.43}$$

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, 2015b). 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**:

$$\mathrm{H}(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(x = x) = -\mathop{\mathbf{E}}_{x} \log_b \mathrm{P}(x). \tag{2.44}$$

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.

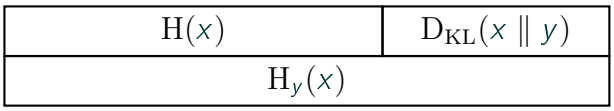
Entropija kontinuirane slučajne varijable je beskonačna. Ako se u izrazu (2.44) vjerojatnost zamijeni gustoćom vjerojatnosti, onda on predstavlja **diferencijalnu entropiju**, jedan od analoga[2] entropije za kontinuirane varijable koji nema neka od svojstava koja ima entropija.

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:

$$\mathrm{H}_y(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(y = x) = -\mathop{\mathbf{E}}_{x} \log_b P_y(x). \tag{2.45}$$

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 (2015b), ovdje koristimo oznaku $\mathrm{H}_y(x)$.

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



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

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

$$\mathrm{D}_{\mathrm{KL}}(x \parallel y) \coloneqq \mathrm{H}_y(x) - \mathrm{H}(x) = \mathop{\mathbf{E}}_{x} \log_b \frac{P_x(x)}{P_y(x)}. \tag{2.46}$$

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 (2.46) 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:

$$\mathrm{I}(x; y) \coloneqq \mathop{\mathbf{E}}_{x,y} \log_b \frac{P_{x,y}(x, y)}{P_x(x) P_y(y)}, \tag{2.47}$$

a može se i na ove načine izraziti:

$$\mathrm{I}(x; y) = \mathrm{H}(x) + \mathrm{H}(y) - \mathrm{H}(x, y) \tag{2.48}$$

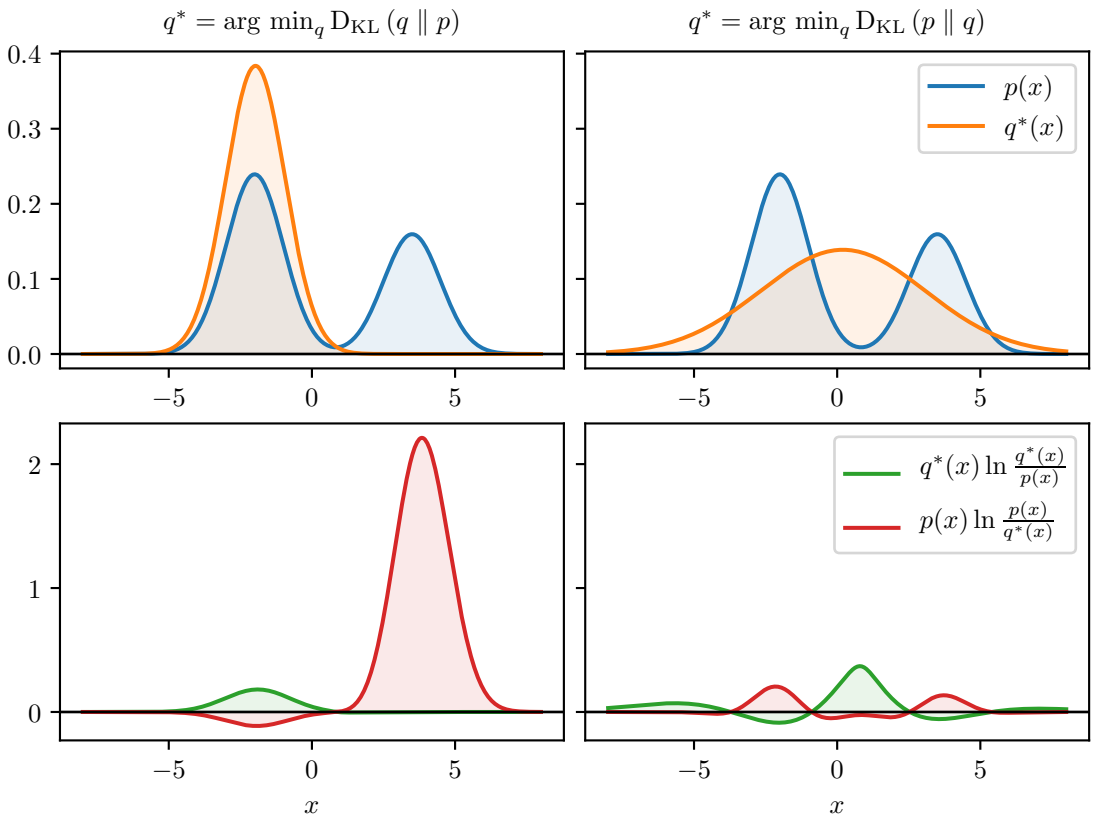
$$= \mathrm{H}(x) - \mathrm{H}(x \mid y) \tag{2.49}$$

$$= \mathrm{H}(y) - \mathrm{H}(y \mid x), \tag{2.50}$$

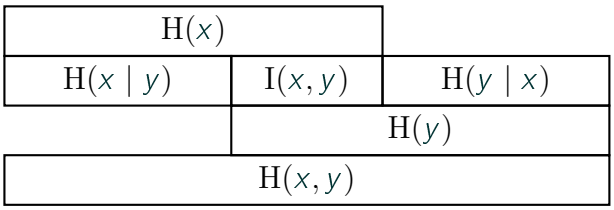
gdje je

$$\mathrm{H}(x \mid y) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{H}(y \mid x = x) \tag{2.51}$$

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 i neki izvedeni algoritmi koji koriste doaatne heuristike. Oni su bitni u strojnom učenju (poglavlje 4), posebno u dubokom učenju (poglavlje 5). Primjena optimizacijskih algoritama u dubokom učenju opisana je u pododjeljku 5.2.3.

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

$$f(\boldsymbol{x}+\boldsymbol{d}) = f(\boldsymbol{x}) + \frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{x})\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\mathsf{T} \frac{\partial^2 y}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\mathsf{T}} \boldsymbol{x} f(\boldsymbol{x})\boldsymbol{d} + \cdots. \tag{2.52}$$

S drugačijim oznakama:

$$f(\boldsymbol{x}+\boldsymbol{d}) = f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^\mathsf{T}\boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\mathsf{T}\boldsymbol{H}_f(\boldsymbol{x})\boldsymbol{d} + \cdots. \tag{2.53}$$

2.3.1. Gradijentni spust i još neki algoritmi koje se temelje na njemu

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:

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}), \tag{2.54}$$

gdje je i redni broj iteracije, a η **veličina koraka** (**stopa učenja** kod strojnog učenja) koja može biti konstanta ili može ovisiti o i.

Neka $\boldsymbol{g} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ i $\boldsymbol{H} = \boldsymbol{H}_f(\boldsymbol{x})$. Za dovoljno mal η

$$f(\boldsymbol{x} - \eta\boldsymbol{g}) = f(\boldsymbol{x}) - \eta\boldsymbol{g}^\mathsf{T}\boldsymbol{g} - \frac{1}{2}\eta^2\boldsymbol{g}^\mathsf{T}\boldsymbol{H}\boldsymbol{g} + \cdots \tag{2.55}$$

Uz neke blage uvjete koje mora zadovoljavati f i dovoljno mal η, gradijentni spust konvergira, tj. proizvoljno se 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:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| < \lambda \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{2.56}$$

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

Gradijentni spust s inercijom

Jedna heuristika koja je često korisna kod optimizacije funkcija koje su nam zanimljive je simuliranje inercije. Jedan korak **gradijentnog spusta s inercijom** (engl. *momentum gradient descent*) ima ovakav oblik:

$$\boldsymbol{v}_i = \beta \boldsymbol{v}_{i-1} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}), \tag{2.57}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta \boldsymbol{v}_i, \tag{2.58}$$

gdje je $\beta \in [0, 1)$ faktor kojim se određuje *otpor* proporcionalan *brzini* $\boldsymbol{v}$, tj. otpor je proporcionalan faktoru $(1 - \beta)$, što se bolje vidi ako se jednadžba (2.57) izrazi ovako:

$$\boldsymbol{v}_i = \boldsymbol{v}_{i-1} - (1 - \beta)\boldsymbol{v}_{i-1} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}). \tag{2.59}$$

β se obično odabire da bude bliže 1. Uz $\beta = 0$ dobiva se obični gradijentni spust, dobiva čestica koja klizi bez trenja. Uz dobro odabran β prigušuju se pomaci koji nisu u smjeru gibanja $\boldsymbol{v}$. To omogućuje bržu konvergenciju i izbjegavanje malih lokalnih optimuma i drugih stacionarnih točaka. Svojstva gradijentnog spusta s inercijom su dobro objašnjena u Goh (2017).

Jedno poboljšanje gradijentnog spusta s inercijom je **Nesterovljev postupak** (Nesterov, 2014; Sutskever, 2013):

$$\boldsymbol{v}_i = \beta \boldsymbol{v}_{i-1} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1} - \eta \boldsymbol{v}_{i-1}), \tag{2.60}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta \boldsymbol{v}_i. \tag{2.61}$$

Brzina se ažurira uz procjenu gradijenta u budućoj točki na temelju brzine iz prethodne iteracije. Onda se izračuna novi pomak uz tako ažuriranu brzinu.

Primjeri algoritama koji koriste još neke heuristike

Kod opisanih algoritama konvergenciju mogu usporavati područja u kojima gradijent ima male vrijednosti. Način na koji se ta pojava može poništiti je npr. da se gradijent podijeli s njegovom normom. Na taj način će pomaci ovisiti samo o stopi učenja, a ne o strmosti funkcije koja se minimizira. Algoritam **RMSProp** (opisan npr. u Hinton (2012) ili Ruder (2016)) ostvaruje nešto slično. Iteracija RMSPropa izgleda ovako:

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}), \tag{2.62}$$

$$\boldsymbol{r}_i = \rho \boldsymbol{r}_{i-1} + (1-\rho)\boldsymbol{g}_i^{\odot 2}, \tag{2.63}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta(\epsilon \mathbf{1} + \boldsymbol{r}_i)^{\odot -\frac{1}{2}} \odot \boldsymbol{g}_i, \tag{2.64}$$

gdje je $\rho \in [0, 1)$ faktor koji određuje koliko se brzo ažurira pokretni prosjek gradijenta kvadriranog po elementima, a ϵ neka mala konstanta. ρ se obično odabire da bude blizu 1. Za $\rho = 0$, ako se ϵ zanemari, dobiva se iteracija algoritma **Rprop** (Hinton, 2012): $\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \operatorname{sgn}(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}))$. RMSPropu se još može dodati inercija.

Jedan algoritam koji često ubrzava učenje je **Adam** (Kingma i Ba, 2014). On uključuje i inerciju i skaliranje. Ime Adam izvedeno je iz *adaptive moment estimation*. Jedna iteracija tog algoritma izgleda ovako:

$$\boldsymbol{g}_i = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}), \tag{2.65}$$

$$\boldsymbol{v}_i = \beta_1 \boldsymbol{v}_{i-1} + (1-\beta_1)\boldsymbol{g}_i, \tag{2.66}$$

$$\boldsymbol{r}_i = \beta_2 \boldsymbol{r}_{i-1} + (1-\beta_2)\boldsymbol{g}_i^{\odot 2}, \tag{2.67}$$

$$\boldsymbol{v}_i' = \frac{1}{1-\beta_1^i}\boldsymbol{v}_i, \tag{2.68}$$

$$\boldsymbol{r}_i' = \frac{1}{1-\beta_1^i}\boldsymbol{r}_i, \tag{2.69}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta\left(\epsilon \mathbf{1} + \boldsymbol{r}_i^{\odot \frac{1}{2}}\right)^{\odot -1} \odot \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_{i-1}), \tag{2.70}$$

gdje je $\boldsymbol{v}_i$ pomični prosjek gradijenta, $\boldsymbol{r}_i$ pokretni prosjek kvadrata gradijenta po elementima, a ϵ mala konstanta. Parametar β_1 ima ulogu kao β kao gradijentnog spusta s inercijom, a β_2 kao ρ kod RMSPropa. Brzina $\boldsymbol{v}_i$ i pokretni prosjek kvadrata $\boldsymbol{r}_i$ se inicijaliziraju u $\mathbf{0}$ i u svakom koraku se skaliraju obrnuto proporcionalno udjelu gradijenta u odnosu na inicijalnu vrijednost $\mathbf{0}$ u pokretnom prosjeku radi poništavanja pristranosti procjena. Za velik i ti faktori skaliranja približavaju se 1.

2.3.2. Postupci drugog reda

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

Ako koristimo kvadratnu aproksimaciju (2.55), 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:

$$\eta = \frac{\boldsymbol{g}^\mathsf{T} \boldsymbol{g}}{\boldsymbol{g}^\mathsf{T} \boldsymbol{H} \boldsymbol{g}}. \tag{2.71}$$

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 (2.53) po $\boldsymbol{d}$ i izjednačavanjem s $\mathbf{0}$ dobiva se:

$$\mathbf{0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^\mathsf{T} + \boldsymbol{d}^\mathsf{T} \boldsymbol{H}_f(\boldsymbol{x}) + \cdots. \tag{2.72}$$

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

$$\boldsymbol{d} = \boldsymbol{H}^{-1} \boldsymbol{g}. \tag{2.73}$$

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

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i - \eta \boldsymbol{H}_f(\boldsymbol{x}_i) \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i) \tag{2.74}$$

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):

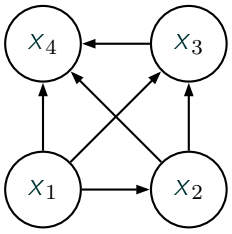
$$\mathrm{p}(\boldsymbol{x}_\mathsf{q} \mid \boldsymbol{x}_\mathsf{o}) = \frac{\mathrm{p}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{o})}{\mathrm{p}(\boldsymbol{x}_\mathsf{o})} = \frac{\int \mathrm{p}(\boldsymbol{x}_\mathsf{q}, \boldsymbol{x}_\mathsf{n}, \boldsymbol{x}_\mathsf{o})\, \mathrm{d}\boldsymbol{x}_\mathsf{n}}{\int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

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 (3.3) 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**:

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

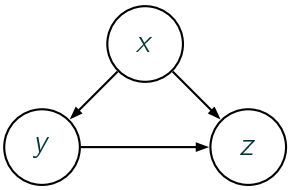
Jednadžba (3.3) onda prelazi u

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

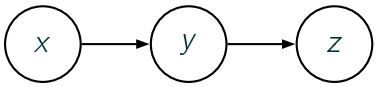
To omogućuje primjenu efikasnijih algoritama za zaključivanje (Murphy, 2012). Na slici 3.2 prikazani su osnovni slučajevi odnosa između triju slučajnih varijabli povezanih zavisnostima koje mogu biti dio većeg grafa. Oni su detaljnije objašnjeni npr. u Bishop (2006); Alpaydin (2014).

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

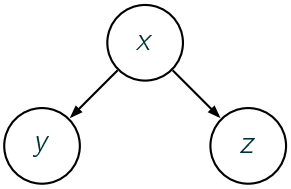
Općenitije, o uvjetnoj nezavisnosti podskupova varijabli govori svojstvo **d-separacije**. Kažemo da je staza (podgraf sa strukturom lanca) P grafa G **d-odvojena** skupom čvorova $\mathbb{E}$ akko P sadrži barem jedno od sljedećeg (Murphy, 2012):



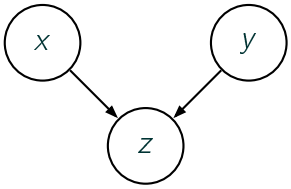
(a) Grafički model s faktorizacijom $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y \mid x)\,\mathrm{p}(z \mid x, y)$.



(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).

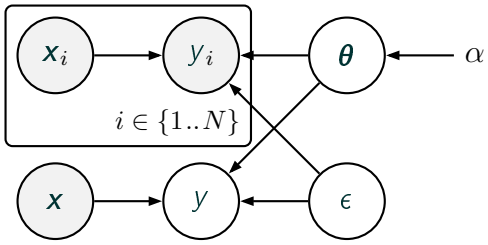


(c) Uz $y \perp z \mid x$ faktorizacija postaje $\mathrm{p}(x, y, z) = \mathrm{p}(x)\,\mathrm{p}(y \mid x)\,\mathrm{p}(z \mid x)$ (račvanje).



(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 y, 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).

1. lanac $a \to b \to c$, gdje $b \in E$
2. račvanje $a \leftarrow b \to c$, gdje $b \in E$
3. 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, 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:

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

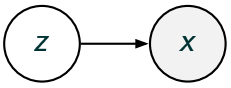
Procjenitelj $\hat{s}_n$ je nepristran ako su x_i nezavisne i imaju istu razdiobu kao x i valjan ako su varijance $f(x_i)$ ograničene. Vrijedi $\mathbf{D}\, \hat{s}_n = \frac{1}{n} \mathbf{D}\, f(x)$.

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. Varijacijsko zaključivanje

Za razliku od uzorkovanja kod MCMC-postupaka, osnovna ideja kod varijacijskog zaključivanja je optimizacija. Prvo se odabire familija razdioba $\mathbb{Q} = \{\mathrm{p}(\tilde{\boldsymbol{z}})\}_{\tilde{z}} = \{q_{\boldsymbol{\phi}}\}_{\boldsymbol{\phi}}$ 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: $q \coloneqq p_{\tilde{z}}$.

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

Ako ciljnu funkciju ovako izrazimo:

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

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*):

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

Ona se može i ovako izraziti:

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

Maksimiziranje takve ciljne funkcije s obzirom na varijacijske parametre daje razdiobu $q^* = \mathrm{p}(\tilde{\mathsf{z}}^*)$ 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 $\ln \mathrm{p}(\boldsymbol{x}) \geq L_{\boldsymbol{x}}(\tilde{\mathsf{z}})$ preko Jensenove nejednakosti. Ta nejednakost slijedi i iz prethodne jednadžbe i nenegativnosti KL-divergencije:

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

3.5.1. 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); 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 $\boldsymbol{\theta}$ kojem 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 4.4.

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

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

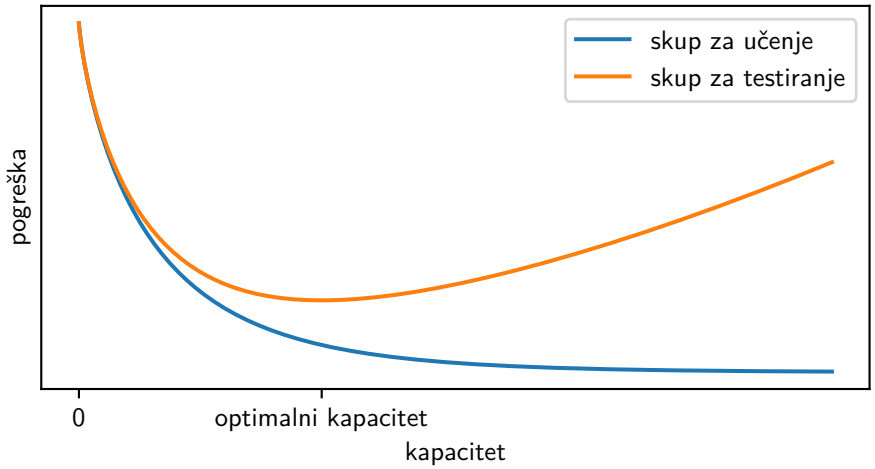
   Kod nekih jednostavnijih modela minimum možemo odrediti analitički. Inače moramo koristiti neki iterativni optimizacijski postupak. Kod nekih složenijih modela, kao što su neuronske mreže, funkcija pogreške nije unimodalna i vjerojatnost pronalaska globalnog optimuma je zanemariva, ali ipak se mogu pronaći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

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

4.4.1. 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**.

4.4.2. 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 $C_2 = -\ln(\int_{\boldsymbol{\theta}'} p_0(\boldsymbol{\theta}') \, \mathrm{d}\boldsymbol{\theta}')$ 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.

Optimalni hiperparametri modela se tražiti postupcima odabira modela (odjeljak 4.5) kod kojih se za procjenu generalizacije koristi skup podataka koji nije korišten za učenje.

4.5. Odabir modela

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

Performansa modela se mjeri nekom evaluacijskom mjerom. Ona omogućuje usporedbu hipoteza ili modela na nekom skupu podataka. Budući da nas zanima generalizacija, za procjenu generalizacije trebamo koristiti podatke koji nisu korišteni

za učenje. Odabir modela se obično svodi na traženje optimalnih **hiperparametara** modela.

4.5.1. Unakrsna validacija

Najjednostavniji način procjenjivanja generalizacije je **unakrsna validacija**. Kod unakrsne validacije, skup podatakama dijelima na **skup za učenje** i **skup za validaciju**. Ako se unakrsna validacija ne koristi za odabir modela, nego za konačnu procjenu generalizacije, onda se skup na kojem se model evaluira naziva **skup za testiranje**.

Za dobivanje bolje procjene generalizacije često se koristi K-struka unakrsna validacija. Kod K**-struke unakrsne validacije** skup podataka $\mathbb{D}$ se podijeli na K dijelova $\mathbb{D}_i$, $i = 1..K$. Model se uči K puta tako da se u i-toj iteraciji za skup za validaciju odabere $\mathbb{D}_i$, a za skup za učenje ostali podaci, $\mathbb{D} \setminus \mathbb{D}_i$. Kao konačna procjena generalizacije uzima se prosjek evaluacija iz svih iteracija.

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.

4.6.1. Primjeri evaluacijskih mjera

Klasifikacija

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 $\boldsymbol{s} = \boldsymbol{W}\phi(\boldsymbol{x})$:

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

2. 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 $y \in 0, 1$.

3. Višeklasna logistička regresija:

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

   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): $\mathbf{e}_k \coloneqq [[\![i=k]\!]]_{i=1..C}^\mathsf{T}$.

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

Na ovaj odjeljak imaju utjecaj Goodfellow et al. (2016) i predavanja iz predmeta *Duboko učenje*.

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. Ni jezgrene metode kod kojih se preslikavanje temelji na pretpostavci sličnosti primjera bliskih u ulaznom prostoru ne generaliziraju dobro zbog **prokletstva dimanzionalnosti** (Bengio et al., 2005). Kod **dubokog učenja** (LeCun et al., 2015; Goodfellow et al., 2016) transformacija ϕ se uči.

Odabirom

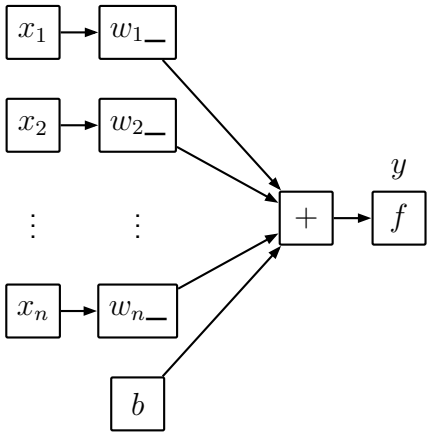
$$\phi(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{\theta}_\mathsf{h}) = f(\boldsymbol{W}_\mathsf{h}\boldsymbol{x} + \boldsymbol{b}_\mathsf{h}), \tag{5.1}$$

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** kojem 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,

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



Slika 5.1: Grafički prikaz umjetnog neurona. w_ označava da se u w množi s ulazom čvora.

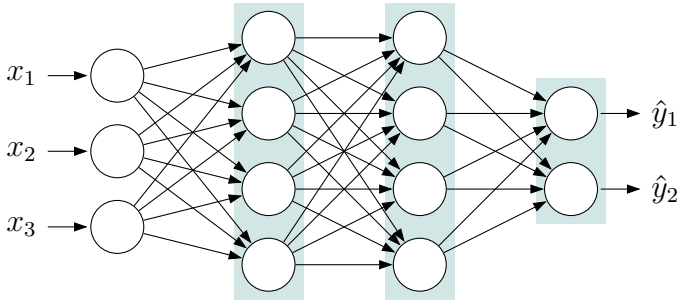
gdje $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathsf{h}}, \boldsymbol{b}_{\mathsf{h}}, \boldsymbol{W}_{\mathsf{o}}, \boldsymbol{b}_{\mathsf{o}})$. Jedinice neuronske mreže kojima odgovaraju operacije oblika $\boldsymbol{x} \mapsto f(\boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{x} + b_i)$ nazivaju se **umjetni neuroni**. Uz taj naziv, ovdje će se još koristiti naziv **jedinica**. Model umjetnog neurona prikazan je na slici 5.1.

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 5.2.

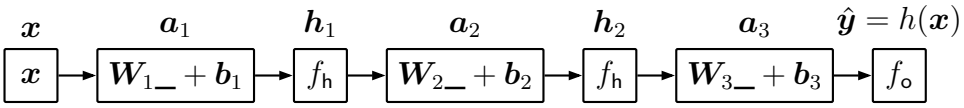
5.1. Duboke unaprijedne mreže

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 5.2 ili apstraktnije, kao na slici 5.3. Povećavanjem broja skrivenih slojeva svaka jedinica u nekom sloju može koristiti izlaze svih jedinica prethodnog



Slika 5.2: Prikaz primjera troslojne mreže. Svakom bridu odgovara jedna težina (pomaci nisu prikazani). Slojevi su označeni plavim četverokutima. Čvorovi koji su unutar četverokuta mreže predstavljaju umjetne neurone. Slika je napravljena prema http://www.texample.net/tikz/examples/neural-network/.

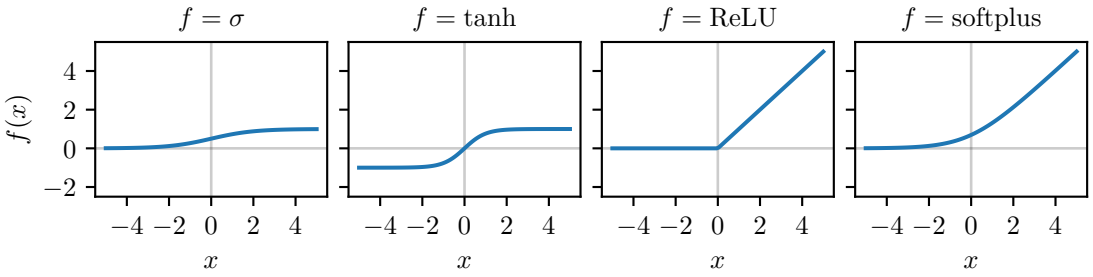


Slika 5.3: Prikaz troslojne mreže kao računskog grafa. Čvorovi predstavljaju funkcije s parametrima, a bridovi podatke (vektore) čije su oznake prikazane uz neke od čvorova iz kojih izlaze. Funkcije su označene oznakom funkcije (aktivacijska funkcija) ili definicijom (afina transformacija). Ulaz sloja označen je s _, a oznake varijabli koje pripadaju čvorovima (ulaz u ulaznom čvoru i parametri u čvorovima afine transformacije) nisu podvučene.

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). Posebne vrste dubokih modela koji uz to iskorištavaju još neke pretpostavke su zato jako uspješne u zadacima u vezi slika, zvuka, teksta i drugih signala. Niži slojevi služe višim slojevima kao značajke transformiranjem kojih se dobivaju značajke više razine.

Kao prijenosna funkcija skrivenih slojeva često se koristi **zglobnica** (**ReLU**, engl. rectified linear unit) $\mathrm{ReLU}(x) = \max(0, s)$ za koju se empirijski pokazalo da ima prednosti nad funkcijama koju su prije bile češće korištene (Glorot et al., 2011). 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. Na slici 5.4 prikazani su grafovi nekih prijenosnih funkcija.

Dosad opisivane mreže nazivaju se **unaprijedne mreže** (engl. *feedforward networks*) zato što se pri izračunu informacija propagira od ulaza prema izlazu, bez



Slika 5.4: Primjeri prijenosnih funkcija.

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 varijable ili računske operacije i njihove izlaze, a bridovi označavaju ovisnosti, tj. koji čvor je ulaz kojeg čvora. Svaki čvor koji nije ulazni predstavlja funkciju koju ostvaruje podgraf koji čine njegovi preci pa ga možemo poistovjetiti s funkcijom čiji su parametri svi ulazni čvorovi iz skupa čvorova predaka. Čvorovi koji nemaju roditelje su varijable koje čine ulazi i parametri. Parametri se mogu dijeliti, tj. mogu biti ulaz većem broju čvorova kao i svi drugi čvorovi. Čvorovi koji nemaju djecu su izlazi računskog grafa. Na slici 5.1 i 5.3 su prikazani takvi grafovi s različitim razinama apstrakcije. U njima, radi sažetosti, parametri nemaju zasebne čvorove, nego su označeni unutar čvorova koji o njima ovise.

5.2. Učenje

Cilj učenja je pronaći parametre $\boldsymbol{\theta}$ koji minimiziraju pogrešku

$$E(\boldsymbol{\theta}; \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{D}} L(\boldsymbol{y}_i, h(\boldsymbol{x}_i; \boldsymbol{\theta})) + \lambda R_{\mathrm{R}}(\boldsymbol{\theta}) \tag{5.4}$$

i postići dobru generalizaciju. Duboki modeli se obično uče algoritmima koji se temelje na gradijentnom spustu. Gradijent pogreške s obzirom na parametre je

$$\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}; \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{D}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{y}_i, h(\boldsymbol{x}_i; \boldsymbol{\theta})) + \lambda \nabla_{\boldsymbol{\theta}} R_{\mathrm{R}}(\boldsymbol{\theta}). \tag{5.5}$$

Kod dubokih mreža, tj. usmjerenih acikličkih računskih grafova, gradijent se računa **algoritmom propagacije pogreške unatrag** (Rumelhart et al., 1986) koji se temelji na **pravilu deriviranja kompozicije funkcija** i **dinamičkom**

programiranju.

U ovom odjeljku su kratko opisane ideje korištene za efikasno računanje gradijenta i optimizacijski postupci koji se koriste za pronalaženje dobrih parametara kod dubokih mreža.

5.2.1. Algoritam propagacije pogreške unatrag

Na ovaj pododjeljak ima utjecaj Olah (2015a).

Gradijent gubitka nije potrebno analitički računati za svaki parametar posebno. Primjenom pravila deriviranja složene funkcije, derivacija vrijednosti nekog čvora b s obzirom na čvor vrijednost nekog čvora $a \in \mathrm{pred}(b)$ jednaka je zbroju umnožaka parcijalnih derivacija *dijete-roditelj* čvorova po svakom putu između a i b. Pri tome je put definiran kao niz takav da sljedeći (usmjereni) brid počinje u čvoru u kojem je prethodni završio. Svakom bridu (p, c) odgovara derivacija $\frac{\partial c}{\partial p}$. Derivacije između susjednih čvorova ne moraju se računati za svaki put posebno. Već izračunate derivacije se mogu ponovo koristiti. Isto vrijedi ako su vrijednosti čvorova vektori (ako su višedimenzionalni nizovi, možemo ih svesti na vektore) i ako računamo Jakobijeve matrice. Dalje će se za Jakobijeve matrice isto koristiti riječ *derivacija*.

Algoritam propagacije pogreške unatrag se tako naziva zato što se izračun gradijenta širi od izlaznog čvora (ili čvora koji predstavlja gubitak ili funkciju pogreške) prema njegovim roditeljima, pa prema roditeljima roditelja itd. uz primjenu pravila deriviranja kompozicije funkcija. Pri tome se ne moraju računati gradijenti s obzirom na čvorove koji ne ovise o varijablama za koje se računa gradijent.

Neka je L vrijednost čvora gubitka, a $\boldsymbol{\theta}$ neki parametar. Želimo izračunati gradijent $\nabla_{\boldsymbol{\theta}_i} L = \frac{\partial L}{\partial \boldsymbol{\theta}}^{\mathsf{T}}$. Derivacija gubitka s obzirom na čvoru u se može rekurzivno izraziti:

$$\frac{\partial L}{\partial u} = \sum_{c \in \mathrm{ch}(u)} \frac{\partial L}{\partial c} \frac{\partial c}{\partial u}, \tag{5.6}$$

gdje su c djeca čvora u. Ako $\frac{\partial L}{\partial c}$ nije izračunat za trenutni ulaz, izračuna se, a inače se koristi prethodno izračunata vrijednost. Ista jednadžba vrijedi za čvorove parametara.

Neka je zadatak nadzirano učenje, $\mathbb{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_i$ skup podataka za učenje, a

Operacija	Derivacije
$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{W}$ $\frac{\partial \boldsymbol{y}}{\partial (\boldsymbol{W}_{[i,:]})^\mathsf{T}} = \boldsymbol{x}^\mathsf{T}$ $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}} = \mathbf{I}$
$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{b}$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}} = \operatorname{diag}\left(\mathbf{1}_{\dim(\boldsymbol{a})} + \mathbf{0}_{\dim(\boldsymbol{b})}\right)$ $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}} = \operatorname{diag}\left(\mathbf{0}_{\dim(\boldsymbol{a})} + \mathbf{1}_{\dim(\boldsymbol{b})}\right)$
$\boldsymbol{y} = \mathrm{ReLU}(\boldsymbol{x})$	$\frac{\partial \boldsymbol{y}_{[i]}}{\partial \boldsymbol{x}_{[j]}} = [\![i = j]\!]\,[\![\boldsymbol{x}_{[j]} > 0]\!]$
$\boldsymbol{y} = \sigma(\boldsymbol{x})$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \operatorname{diag}(\boldsymbol{y} \odot (\mathbf{1} - \boldsymbol{y}))$
$\boldsymbol{y} = \tanh(\boldsymbol{x})$	$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \operatorname{diag}(1 - \boldsymbol{y} \odot \boldsymbol{y})$
$\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{x})$	$\frac{\partial \boldsymbol{y}_{[i]}}{\partial \boldsymbol{x}_{[j]}} = \boldsymbol{y}_{[j]}\left([\![i = j]\!] - \boldsymbol{y}_{[j]}\right)$
$y = -t \ln \sigma(x) - (1-t)\ln(1-\sigma(x))$	$\frac{\partial y}{\partial x} = \sigma(x) - t$
$y = -\ln \mathrm{softmax}(\boldsymbol{x})_{[t]}$	$\frac{\partial y}{\partial \boldsymbol{x}} = (\mathrm{softmax}(\boldsymbol{x}) - \mathbf{e}_t)^\mathsf{T}$

Tablica 5.1: Parcijalne derivacije (Jakobijeve matrice) nekih operacija po njihovim ulazima. Zadnja sva retke predstavljaju gubitak unakrsne entropije (negativni logaritam izglednosti) za binarnu i višeklasnu klasifikaciju, gdje je t indeks ciljne klase. $\mathbf{e}_t$ označava jednojedinični vektor s elementima $\mathbf{e}_{\mathbf{t}[i]} = [\![i = t]\!]$.

$L_i = L(\boldsymbol{y}_i, h(\boldsymbol{x}_i, \boldsymbol{\theta}))$ gubitak para $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Neka je pogreška npr. $E = \sum_i L_i + R_\mathsf{R}(\boldsymbol{\theta})$. Onda

$$\frac{\partial E}{\partial \boldsymbol{\theta}} = \sum_i \frac{\partial L_i}{\partial \boldsymbol{\theta}} + \frac{\partial R_\mathsf{R}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{5.7}$$

gdje se izrazi na desnoj strani računaju rekurzivno uz pamćenje izračunatih gradijenata (ili unaprijed izračunate gradijente koji odgovaraju bridovima u podgrafu koji se sastoji od čvorova potomaka) prema jednadžbi 5.6.

5.2.2. Gradijenti nekih osnovnih operacija

U tablici 5.1 prikazane su parcijalne derivacije (Jakobijeve matrice) nekih operacija s obzirom na njihove ulaze. Korištenjem pravila deriviranja kompozicije funkcija mogu se izračunati derivacije složenijih funkcija. Radi efikasnosti se izračunavanje vrijednosti u računskom grafu i algoritam propagacije pogreške unatrag provodi paralelno za više ulaza odjednom. Izvodi gradijenata nekih operacija uz višestruke ulaze mogu vidjeti npr. ovdje: http://www.zemris.fer.hr/~ssegvic/du/lab0.shtml.

5.2.3. Stohastička optimizacija

U pododjeljku 2.3.1 opisan je gradijentni spust i neki izvedeni algoritmi koji koriste neke dodatne heuristike. U ovom pododjeljku opisana je primjena tih algoritama u dubokom učenju. Kod učenja dubokih modela se obično koristi puno podataka i iteracija optimizacije se provodi procjenjivanjem gradijenta funkcije pogreške na temelju manjeg dijela skupa za učenje.

Kod **stohastičkog gradijentnog spusta** se u nekoj iteraciji gradijentnog spusta umjesto gradijenta pogreške koristi gradijent procjene pogreške na temelju nekog podskupa skupa za učenje ili samo jednog primjera. Takav algoritam naziva se **stohastički gradijentni spust**. Moguće je podskupove u svakoj iteraciji ponovo slučajno izvlačiti iz cijelog skupa za učenje $\mathbb{D}$, ali obično se iteracije podijele na **epohe** od kojih se svaka sastoji od B iteracija, u svakoj epohi se skup za učenje slučajno podijeli na B nepreklapajućih podskupova $\mathbb{D}_i$ jednake veličine, od kojih se svaki koristi u jednoj iteraciji unutar epohe. Skupove $\mathbb{D}_i$ nazivamo **mini-grupe**. U iteraciji i u nekoj epohi koristi se procjena gradijenta

$$\nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}; \mathbb{D}_i) = \frac{1}{|D_i|} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{D}_i} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{y}_i, h(\boldsymbol{x}_i; \boldsymbol{\theta})) + \lambda \nabla_{\boldsymbol{\theta}} R_{\mathrm{R}}(\boldsymbol{\theta}). \tag{5.8}$$

i iteracija (prema jednadžbi (2.54)) ima oblik

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \eta \nabla_{\boldsymbol{\theta}} E(\boldsymbol{\theta}; \mathbb{D}_i), \tag{5.9}$$

gdje je e broj epohe, a $i+1$ broj trenutne iteracije unutar epohe.

Prema Goodfellow et al. (2016), na odabir veličine mini-grupe utječu:

1. Kvaliteta procjene gradijenta. Veće minigrupe daju točniju procjenu gradijenta.
2. Računska efikasnost. Premale mini-grupe ne iskorištavaju potpuno mogućnost paralelizacije izračuna, a prevlike grupe ne stanu u memoriju.
3. Optimizacija s manjim mini-grupama ima učinak regularizacije (Wilson i Martinez, 2003), ali zahtijeva manju stopu učenja i sporije konvergira.

Kako bi optimizacijski algoritam konvergirao, treba se smanjivati stopa učenja ovisno o iteraciji (epohi). Prema Goodfellow et al. (2016), dovoljan uvjet za

konvergenciju gradijentnog spusta je

$$\sum_{k=1}^{\infty} \eta_k = \infty \wedge \sum_{k=1}^{\infty} \eta_k^2 < \infty, \tag{5.10}$$

gdje je k broj iteracije od početka učenja.

Kako bi se ublažio šum procjene gradijenta i ubrzalo učenje, obično se upotrebljava inercija, kao što je opisano u pododjeljku 2.3.1. Empirijski se pokazuje da stohastički gradijentni spust s momentom postiže dobru generalizaciju. U pododjeljku 2.3.1 su opisana i dva algoritma koja koriste pokretne prosjeke momenata gradijenta i prilagođeni su stohastičkom učenju s mini-grupama. RMSProp skalira gradijent po elementima korištenjem pokretnog prosjeka kvadrata gradijenta tako da norma elemenata pomaka ne ovisi jako o prosječnoj normi gradijenta u zadnjim iteracijama. Adam koristi inerciju i obavlja skaliranje slično kao RMSProp.

5.2.4. Inicijalizacija parametara

Kod učenja dubokih modela jako je bitna inicijalizacija parametara. Sve težine mreže (ili dijela nje), npr. kao na slici 5.2, se ne smiju se inicijalizirati konstantnom vrijednošću. Zamjenom dvaju jedinica istog sloja, npr. kao na slici 5.2, dobiva se ista mreža i gradijent je jednak za sve težine unutar istog sloja, osim za zadnji sloj. To se rješava inicijalizacijom težina nasumičnim vrijednostima. Ako su inicijalizirane težine manje, sporije će se *razbijati* simetrija, a ni prevelike težine nisu dobre. Ako se koriste prijenosne funkcije sa zasićenjem problem mogu biti težine s prevelikom apsolutnim vrijednostima jer mogu uzrokovati zasićenje i tako onemogućavati učenje. Taj problem rješava zglobnica, ali množenjem velikih težina kroz više slojeva daje sve veće izlaze, što kod linearnih slojeva daje prevelik gradijent, što se vidi u tablici 5.1.

Heuristike korištene za inicijalizaciju težina se temelje na aproksimiranju mreže nizom matričnih množenja i postizanju da varijance gradijenata (i izlaza) budu otprilike konstantne kroz mrežu (Goodfellow et al., 2016). Otprilike konstantna varijanca gradijenta može se ostvariti inicijalizacijom iu Guassove ili unifomne razdiobe s varijancom $\frac{1}{n}$, gdje je n broj ulaza. Glorot i Bengio (2010) kao kompromis između jednake varijance gradijenta i jednake varijance izlaza slojeva predlažu varijancu $\frac{1}{n+m}$, gdje je m broj izlaza (Goodfellow et al., 2016).

Pomaci se obično inicijaliziraju na neku konstantu.

5.2.5. Problem nekonveksnosti funkcije pogreške

Goodfellow et al. (2016) navode sljedeće probleme koji se javljaju kod nekonveksne optimizacije:

1. **Loše kondicioniranje Hesseove matrice**. Loše kondicioniranje Hesseove matrice može biti razlog da i s jako malim korakom učenje funkcija pogreške raste u smjeru gradijenta zato što kvadratni član u Taylorovom rezvoju u jednadžbi (2.55) bude pozitivan i prevlada linearni član.

2. **Lokalni minimumi**. I ako se zanemare ekvivalentni lokalni minimumi koji postoje zbog simetričnosti zamjenjivosti položaja neurona u istom sloju i drugih simetričnosti u neurnoskim mrežama, funkcija pogreške ima velik broj lokalnih minimuma. Empirijski se pokazuje da loši lokalni minimumi nisu problem i da nije potrebno pronaći globalni minimum kako bi se dobili dobri rezultati.

3. **Ostale stacionarne točke**. Kod visokodimenzionalnih optimizacijskih problema lokalni minimumi i maksimumi su obično rijetki zato što bi onda sve vlastite vrijednosti Hesseove matrice morale biti istog predznaka. Zato su češće sedlaste točke kod kojih se predznak barem jedne vlastite vrijednosti razlikuje od predznaka ostalih. Empirijski se pokazuje da sedlaste točke kod dubokih mreža nisu velik problem za optimizacijske postupke prvog reda koje ne privlače sedlaste točke. I ako se parametri nalaze točno u sedlastoj točki tako da je gradijent $\mathbf{0}$, stohastički gradijentni spust može imati drugačije gradijente.

4. **Litice i eksplodirajući gradijenti**. Kod nekih modela javlja se problem velikih vrijednosti gradijenta u nekim točkama. To se može riješiti ograničavanjem norme gradijenta.

5.3. Regularizacija i poboljšavanje učenja

U ovom pododjeljku opisani su neki od češćih postupaka koji se koriste za poboljšavanje učenja, tj. postizanja bolje generalizacije i bržeg učenja. Dijelovi ovog

pododjeljka temelje se na Goodfellow et al. (2016), gdje se može naći opširniji i dublji pregled.

5.3.1. Kažnjavanje norme težina

Najjednostavniji način regularizacije je kažnjavanje norme težina. Regularizacijski dio funkcije pogreške R_{R} se najčešće definira kao kvadrat L^2 norme, tj. koristi se L^2 **regularizacija** koja odgovara apriornoj pretpostavci Gaussove razdiobe težina s dijagonalnom kovarijacijskom matricom i očekivanjem $\mathbf{o}$. Može se koristiti i L^1 **regularizacija** koja potiče rijetkost težina, tj. postavlja minimum funkcije pogreške u ovisnosti o nekim težinama točno u 0. To je detaljnije objašnjeno npr. u Goodfellow et al. (2016). L^1 regularizacija odgovara Laplaceovoj apriornoj razdiobi. Općenito, gubitak L^p regularizacije ima oblik:

$$R_{\mathsf{R}}(\boldsymbol{\theta}) = \frac{\lambda}{p}\|\boldsymbol{\theta}\|_p^p = \frac{\lambda}{p}\sum_i \left|\boldsymbol{\theta}_{[i]}\right|^p, \tag{5.11}$$

gdje λ određuje jačinu regularizacije, tj. koncentraciju apriorne razdiobe. Gustoći apriorne razdiobe odgovara

$$\mathrm{p}(\boldsymbol{\theta}) = \frac{1}{Z}\exp(-R_{\mathsf{R}}(\boldsymbol{\theta})) \tag{5.12}$$

$$= \frac{1}{Z}\prod_i \exp\left(-\frac{\lambda}{p}\left|\boldsymbol{\theta}_{[i]}\right|^p\right), \tag{5.13}$$

gdje je Z normalizacijska konstanta. Gradijent regularizacijskog gubitka s obzirom na $\boldsymbol{\theta}_{[i]}$ je $\lambda\,\mathrm{sgn}(\boldsymbol{\theta}_{[i]})\left|\boldsymbol{\theta}_{[i]}\right|^{p-1}$. Posebno, to je $\lambda\boldsymbol{\theta}_{[i]}$ za $p=2$ i $\lambda\,\mathrm{sgn}(\boldsymbol{\theta}_{[i]})$ za $p=1$.

5.3.2. Rano zaustavljanje učenja

Regularizacijski učinak koji se može usporediti s L^p regularizacijom ima **rano zaustavljanje** učenja zato što ograničava koliko se parametri mogu udaljiti od početne vrijednosti. Ako se model predugo uči, može se dogoditi da se težine modela s velikim kapacitetom previše prilagode skupu za učenje i zato dođe do loše generalizacije.

5.3.3. Generiranje podataka

Postupci koji značajno mogu utjecati na generalizaciju su postupci **proširivanja skupa podataka**, što znači da se tijekom učenja obično provode jednostavne slučajne transformacije nad primjerima prije nego što se daju kao ulaz modelu. Primjeri transformacija koje se mogu koristiti u računalnom vidu, ovisno o zadatku, su reflektiranje, translacija i rotacija. Generiranje podataka kod zadataka koji imaju veze sa zvukom isto može biti korisno Goodfellow et al. (2016).

Dodavanje šuma ulazu isto može biti korisno (Goodfellow et al., 2016). To odgovara pretpostavci da primjeri koji su slični u ulaznom prostoru trebaju biti slični i u izlaznim prostoru.

Pokazalo se da je moguće pronaći ulazne primjere koji su u ljudskoj percepciji slični prirodnim primjerima, ali i modeli koji ostvaruju rezultate usporedive s rezultatima ljudi daju krive predikcije (Szegedy et al., 2013; Goodfellow et al., 2014). Takvi ulazni primjeri nazivaju se **neprijateljski primjeri**. Oni se mogu dobiti i pomoću jednog koraka gradijentnog spusta pomicanjem ulaznog primjera u smjeru povećavanja gubitka. Jedan od uspješnijih načina postizanja otpornosti na neprijateljske primjere je proširivanje skupa za učenje neprijateljskim primjerima (Madry et al., 2017).

5.3.4. Isključivanje neurona - dropout

Dropout (Hinton et al., 2012; Srivastava et al., 2014) je postupak regularizacije koji unosi šum u izlaze skrivenih slojeva tijekom učenja. Obično se ostvaruje tako da se tijekom učenja za svaki primjer svaka jedinica mreže isključi s vjerojatnošću p, koja je hiperparametar. Tijekom testiranja, tj. zaključivanja, sve se jedinice skaliraju s $1 - p$, tj. očekivanjem skaliranja koje je tijekom učenja 0 s vjerojatnošću p, a 1 s vjerojatnošću $1 - p$. Dropout se obično primjenjuje nakon afine transformacije (ne nakon aktivacije).

Učenje s *dropoutom* se može interpretirati kao učenje eksponencijalnog broja modela koji dijele parametre, a zaključivanje kao aproksimacija usrednjavanja modela ili aproksimacija bayesovskog zaključivanja. Vremenski zahtjevniji, ali ispravniji postupak usrednjavanj modela bio bi uzorkovanje (Srivastava et al., 2014), tj. *Monte Carlo* aproksimacija izlaza. Gal i Ghahramani (2016) su dali bayesovku interpretaciju takvog usrednjavanja.

Umjesto Bernoullijeve razdiobe, skaliranje ili izlazi jedinica mogu imati npr. Gaussovu razdiobu ili neku drugu.

5.3.5. Normalizacija po grupama

Normalizacija po grupama (engl. *batch normalization*) (Ioffe i Szegedy, 2015) je postupak koji ublažava probleme pri učenji i omogućuje učenja jako dubokih modela. Prema Goodfellow et al. (2016), problem kod učenja jako dubokih modela je što se se sastoje od kompozicije velikog broja funkcija, zbog čega je velika međuzavisnost između parametara različitih slojeva, a u koraku gradijentnog spusta parametri svih funkcija ažuriraju se istovremeno. Gradijenti spust pretpostavlja da je utjecaj svakog parametra lokalno nezavisan, tj. svaka se funkcija (sloj afine transformacije) prilagođava ostatku mreže kakav je u trenutnom koraku, tj. očekuje da se prethodni slojevi neće promijeniti.

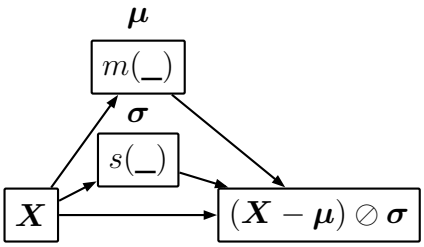
U Goodfellow et al. (2016) je to objašnjeno na jednostavnom primjeru $\hat{y} = xw_1w_2, .., w_j$, gdje su elementi gradijenta $g_i = \nabla_{\boldsymbol{w}_i}\hat{y} = x\prod_{j\neq i} w_j$. Novi izlaz nakon koraka gradijentnog spusta je $x\prod_i(w_i - \epsilon g_i)$ gdje članovi uz više potencije ϵ mogu imati utjecaj koji raste eksponencijalno s dubinom l.

Sloj normalizacije po grupama se dodaje nakon sloje linearne transformacije (prije prijenosne funkcije). On kod učenja obavlja ovakvu operaciju:

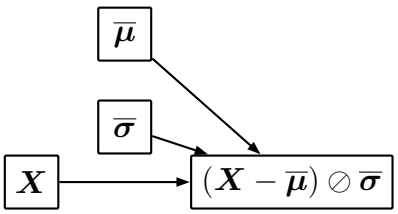
$$\boldsymbol{Y} = (\boldsymbol{X} - m(\boldsymbol{X})) \oslash s(\boldsymbol{X}), \tag{5.14}$$

gdje je $\boldsymbol{X} = [\boldsymbol{x}_i]^{\mathsf{T}}_{i=1..N} \in \mathbb{R}^{N\times n}$ matrica kojoj su reci vektori značajki pojedinih primjera, $\boldsymbol{Y} = [\boldsymbol{y}_i]^{\mathsf{T}}_{i=1..N} \in \mathbb{R}^{N\times n}$ matrica kojoj su reci značajke normalizirane ulazne značajki, $m(\boldsymbol{X}) \coloneqq \frac{1}{N}\sum_i \boldsymbol{X}_{[i,:]} \in \mathbb{R}^{1\times n}$ srednja vrijednost vektora značajki, $s(\boldsymbol{X}) \coloneqq \left(\frac{1}{N}\sum_i\left(\boldsymbol{X}_{[i,:]} - m(\boldsymbol{X})\right)^{\odot 2}\right)^{\odot\frac{1}{2}} \in \mathbb{R}^{1\times n}$ standardna devijacija vektora značajki po elementima. Oduzimanje u jednadžbi (5.14) je definirano tako da se od svakog retka $\boldsymbol{X}$ oduzima $m(\boldsymbol{X})$. Takvo značenje ima i dijeljenje. Izlaz sloja normalizacije po grupama tijekom učenja je invarijantan na skaliranje i pomak ulaza $\boldsymbol{X}$.

Statistike grupe, tj. srednje vrijednosti i standardne devijacije od grupe za koju se provodi zaključivanje, se koriste samo kod učenja. Inače se koriste statistike skupa za učenje koje se mogu procijenjivati pokretnim prosjekom tijekom učenja. Računski graf normalizacije po grupama kod učenja i kod ispitivanja je prikazan na slici 5.5.



(a) Graf normalizacije po grupama kod učenja.



(b) Graf normalizacije po grupama kod ispitivanja.

Slika 5.5: Grafovi normalizacije po grupama kod učenja i kod ispitivanja. m i s su funkcije koje računaju srednju vrijednost i standardnu devijaciju grupe. $\overline{\boldsymbol{\mu}}$ je srednja vrijednost, a $\overline{\boldsymbol{\sigma}}$ standardna devijacija ulaza kod skupa za učenje.

Kako se ne bi izgubila ekspresivnost, nakon sloja normalizacije po grupama obično se dodaje pomak $\boldsymbol{\beta} \in \mathbb{R}^{1\times n}$ i skaliranje $\boldsymbol{\gamma} \in \mathbb{R}^{1\times n}$ svake značajke. $\boldsymbol{\beta}$ i $\boldsymbol{\gamma}$ su parametri koji se uče. Kod ispitivanja je normalizacija po grupama uz skaliranje i pomak samo drugačija parametrizacija koja je uz prethodni sloj linearne transformacije s težinama $\boldsymbol{W}$ može svesti na sloj afine transformacije s težinama $\boldsymbol{W} \oslash \boldsymbol{\sigma}^{\mathsf{T}} \odot \boldsymbol{\gamma}^{\mathsf{T}}$ i pomacima $-\boldsymbol{\mu} \oslash \sigma \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$.

Kod konvolucijskih mreža se normalizacija po grupama ne provodi samo po dimenziji grupe, nego i po dimenzijama konvolucije, po kojima treba vrijediti translacijska ekvivarijantnost. Npr. ako je ulazni niz dimenzija $N \times H \times W \times C$, gdje je N veličina grupe, H visina slike, W širina slike, a C broj značajki, tj. broj filtara zadnje konvolucije, vektor srednjih vrijednosti i vektor srandardnih odstupanja je dimenzije C, tj. $1 \times 1 \times 1 \times C$.

5.4. Konvolucijske mreže

Konvolucijske mreže su mreže koje, prema definiciji u Goodfellow et al. (2016), na barem jednom mjestu, umjesto općenite linearne transformacije, koriste **konvoluciju**. Konvolucijske mreže koriste pretpostavku **translacijske ekvivarijantnosti** po nekim dimenzijama ulaza i posebno se uspješno primjenjuju na zadacima u vezi slika. Pojedini elementi izlaza **konvolucijskog sloja** računaju se množenjem manjeg **filtra** s elementima ulaza koje on prekriva na svakom položaju ulaza. Element i izlaza ovise o malom broju elemenata ulaza oko odgovarajućih položaja, tj. **povezanost** je **lokalna**. To omogućuje da se broj parametara konvolucijskog sloja značajno smanji u odnosu na **potpuno-povezani sloj**, tj. sloj

linearne transformacije. Pojedine težine uče se na različitim dijelovima ulaza i to sve omogućuje veću efikasnost i bolju generalizaciju.

5.4.1. Konvolucija

Konvolucija funkcija iz $\mathbb{Z} \to \mathbb{R}$ definirana je ovim izrazom:

$$(f * g)(t) \coloneqq \sum_{\tau} f(\tau) g(t - \tau). \tag{5.15}$$

Jednako tako, definirana je konvolucija funkcija iz $\mathbb{Z}^n \to \mathbb{R}$:

$$(f * g)(\boldsymbol{t}) \coloneqq \sum_{\boldsymbol{\tau}} f(\boldsymbol{\tau}) g(\boldsymbol{t} - \boldsymbol{\tau}). \tag{5.16}$$

Na isti način, s integralom umjesto zbroja, definirana je i konvolucija funkcija s kontinuiranom domenom. Neka od svojstava konvolucije su:

1. Komutativnost: $f * g = g * f$.
2. Distributivnost zbrajanja. Vrijedi $(f + g) * h = f * h + g * h$.
3. Translacijska ekvivarijantnost. Ako $f'(t) \coloneqq f(t + d)$, onda $(f' * g)(t) = (f * g)(t + d)$.
4. Konvolucija u vremenskoj domeni odgovara umnošku u Fourierovoj domeni, tj. $\mathrm{F}[f * g] = \mathrm{F}[f]\mathrm{F}[g]$, gdje F označava odgovarajuću Fourierovu transformaciju (Jeren, 2015).

Konvolucija se može poopćiti na funkcije s kodomenom koja može općeniti vektorski prostor, tj. funkcije iz $\mathbb{Z}^m \to \mathbb{R}^n$. Jedan način je ovaj, gdje se po svakoj komponenti paralelno obavlja konvolucija:

$$(f *_{\mathsf{p}} g)(\boldsymbol{t}) \coloneqq \sum_{\boldsymbol{\tau}} f(\boldsymbol{\tau}) \odot g(\boldsymbol{t} - \boldsymbol{\tau}). \tag{5.17}$$

Drugi način je ovaj, gdje se izlazni vektori funkcija skalarno množe:

$$(f *_{\mathsf{s}} g)(\boldsymbol{t}) \coloneqq \sum_{\boldsymbol{\tau}} \langle f(\boldsymbol{\tau}) | g(\boldsymbol{t} - \boldsymbol{\tau}) \rangle. \tag{5.18}$$

U ovom slučaju, kodomena funkcije $f *_{\mathsf{s}} g$ je $\mathbb{R}$. Isti izraz vrijedi i ako je kodomena funkcija f i g neki skup n-dimenzionalnih nizova, tj. $\mathbb{R}^{d_1 \times \cdots \times d_n}$, gdje su d_i pojedine dimenzije niza. Zato se za skalarni produkt ovdje koristi oznaka skalarnog produkta.

5.4.2. Konvolucijski sloj

Jednom umjetnom neuronu kod konvolucijskih mreža, ako se zanemari pomak, obično odgovara operacija u jednadžbi (5.18), samo što funkcijama f i g odgovaraju konačni $(m+1)$-dimenzionalni (ili m-dimenzionalni ako $n=1$) nizovi pa treba prilagoditi definiciju konvolucije na nizove. Jednoj funkciji odgovara ulazni niz, a drugoj **konvolucijska jezgra** (**filtar**) koja je obično manja i neovisna o veličini ulaza. Izlaz konvolucije je onda m-dimenzionalni niz kojem su dimenzije obično iste kao privih m dimenzija ulaznog niza, ovisno o prilagodbi definicije konvolucije na nizove. Ovakvu konvoluciju ćemo nazivati m-**dimenzionalna konvolucija**. Ovdje se neće razmatrati m-dimenzionalna konvolucija $(m+n)$-dimenzionalnih nizova kod kojih $n>1$, tj. $\boldsymbol{A}_{[i_1,..,i_m,:]}$ su vektori ako je $\boldsymbol{A}$ $(m+1)$-dimenzionalan.

Slojevi koji obavljaju konvoluciju nazivaju se **konvolucijski slojevi**. Izlaz jedne jedinice (dobiven jednim filtrom) u konvolucijskom sloju naziva se **mapa značajki**. Izlaz konvolucijskog sloja sastoji se od više mapa značajki i čini $(m+1)$-dimenzionalni izlaz kojem je zadnja dimenzija jednaka broju mapa značajki. m-dimenzionalnu konvoluciju s k jezgri nazivat ćemo k-**struka** m-**dimenzionalna konvolucija**.

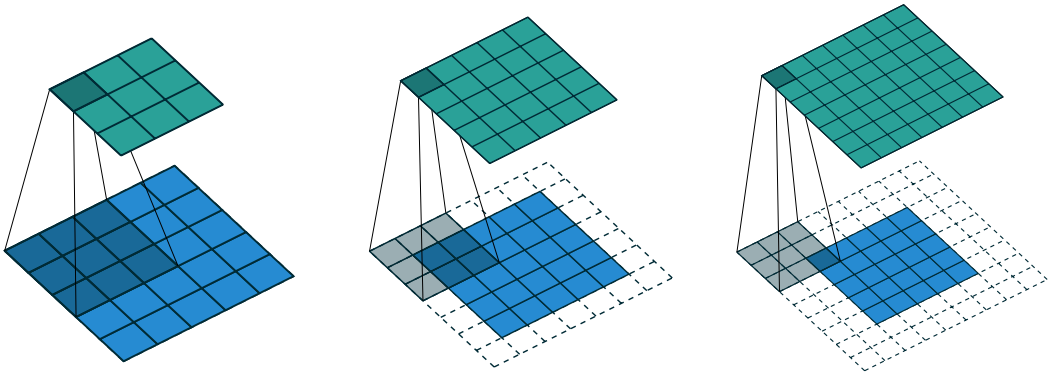
Osnovni način definiranja m-dimenzionalne konvolucije (unakrsne korelacije ako ne reflektiramo jezgru) $(m+1)$-dimenzionalnog ulaza $\boldsymbol{X}$ s $(m+1)$-dimenzionalnom jezgrom $\boldsymbol{W}$, što daje m-dimenzionalni niz $\boldsymbol{X} *_{\mathsf{s}} \boldsymbol{W}$, može se ovako izraziti:

$$(\boldsymbol{X} *_{\mathsf{s}}^{\mathsf{v}} \boldsymbol{W})_{[\boldsymbol{t}]} \coloneqq \left\langle \boldsymbol{X}_{[\boldsymbol{t}:\boldsymbol{t}+\boldsymbol{d_W}+\boldsymbol{1},:]} \middle| \boldsymbol{W} \right\rangle, \tag{5.19}$$

gdje je $\boldsymbol{d_W} = \dim(\boldsymbol{W})_{[1:m]}$ vektor dimenzija jezgre po kojima se obavlja konvolucija. Skalarni produkt na desnoj je definiran ako $\forall i \in \{1..m\}\ \boldsymbol{t}_{[i]} \in \left\{1, .., \boldsymbol{d_X}_{[i]} - \boldsymbol{d_W}_{[i]} - 1\right\}$, gdje je $\boldsymbol{d_X} = \dim(\boldsymbol{X})_{[1:m]}$ vektor dimenzija ulaza. Izlaz takve operacije je dimenzija $\left(\boldsymbol{d_X}_{[i]} - \boldsymbol{d_W}_{[i]} - \boldsymbol{1}\right)$. Kod obrade slike obično želimo da izlaz konvolucije bude jednakih dimenzija kao ulaz. To se može ostvariti dopunjavanjem ulaza nulama po rubu dimenzija po kojima treba obavljati konvoluciju tako da sredina jezgre, za koju pretpostavljamo da ima neparne dimenzije, može doći do ruba originalnog ulaza. Neka $\mathrm{pad}\left(\boldsymbol{X}, \frac{1}{2}(\boldsymbol{d_W} - \boldsymbol{1})\right)$ označava takvu operaciju dopunjavanja. Definiramo novu operaciju:

$$\boldsymbol{X} *_{\mathsf{s}}^{\mathsf{s}} \boldsymbol{W} \coloneqq \mathrm{pad}\left(\boldsymbol{X}, \tfrac{1}{2}(\boldsymbol{d_W} - \boldsymbol{1})\right) *_{\mathsf{s}}^{\mathsf{v}} \boldsymbol{W}. \tag{5.20}$$

U gornjem indeksu operatora "v" dolazi od riječi *valid* zato što se filtar pomiče samo



(a) Bez dopunjavanja.

(b) Dopunjavanje takvo da izlaz ima iste dimenzije kao ulaz.

(c) Potpuno dopunjavanje.

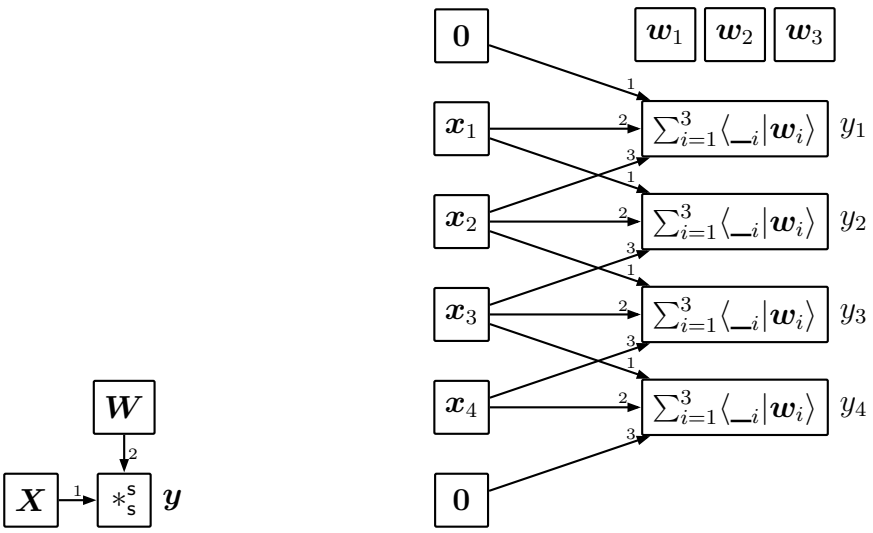
Slika 5.6: Ilustracija dopunjavanja kod dvodimenzionalne konvolucije. Slika 5.6b i slika 5.6c su preuzete, a slika 5.6a je napravljena na temelju slika iz Dumoulin i Visin (2016).

unutar granica ulaza, a "s" od riječi *same* zato što je izlaz istih dimenzija kao ulaz (osim zadnje). Na slici 5.6 ilustrirani su najčešći načina dopunjavanja na primjeru jednostruke dvodimenzionalne konvolucije dvodimenzionalnih nizova. Na slici 5.7 prikazana je jednostruka jednodimenzionalna konvolucija (unakrsna korelacija) dvodimenzionalnih nizova s dopunjavanjem kao u jednadžbi (5.20).

Izlazni korak konvolucije i dilatacija jezgre

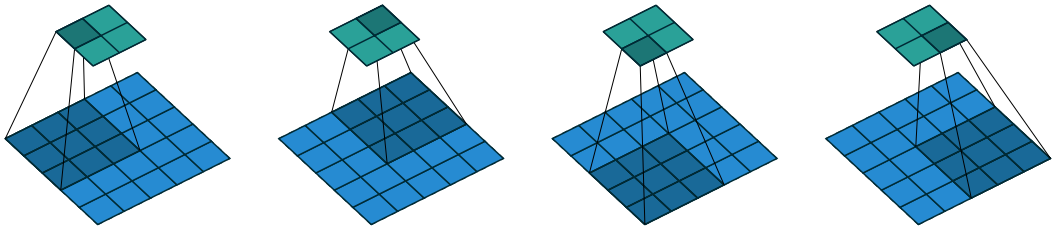
Kod konvolucijski mreža još se koriste neke izmjene konvolucije kako bi se postigla veća računalna efikasnost. Jedna je korištenje **izlaznog koraka** (ili **korak**). Izlazni korak veći od 1 da jezgra po toj dimenziji preskače neke položaje. Na taj način se postiže da dimenzije izlaza budu manje za otprilike za faktor veličine izlaznog koraka. Konvolucija s Izlazniim korakom 2 po svim dimenzijama konvolucije ilustrirana ja na slici 5.8.

Kako bi se povećalo **receptivno polje** jedinice konvolucijskog sloja bez povećavanja dimenzija jezgre, koristi se konvolucija s **dilatacijom** (ili **dilacijom**), tj. **širenjem jezgre**. Na slici 5.9 ilustrirana konvolucija s dilacijom 1. Takva konvolucija je ekvivalentan konvoluciji kod koje se koristi veća jezgra kod koje se svaki drugi redak ili stupac sastoji od nula.

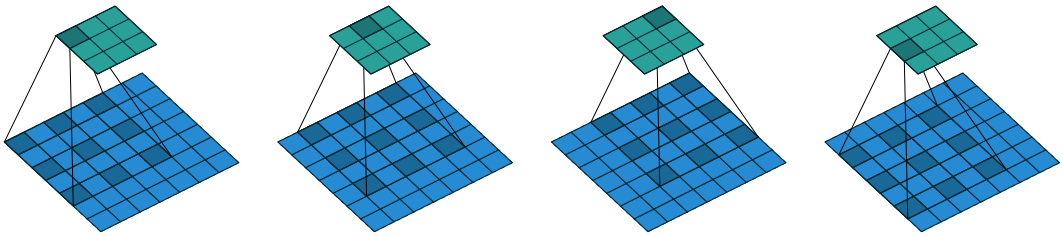


(a) Apstraktni prikaz konvolucije. **(b)** Detaljniji prikaz konvolucije.

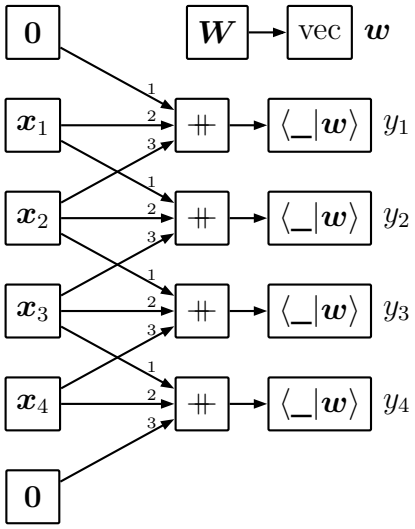
Slika 5.7: Grafički prikaz jednodimenzionalne konvolucije s dopunjavanjem. Na slici b detaljnije su prikazani dvodimenzionalni nizovi $\boldsymbol{X} \in \mathbb{R}^{4\times n}$ i $\boldsymbol{W} \in \mathbb{R}^{3\times n}$ iz slike a rastavljeni na vektore, dopunjavanje i konvolucija na razini vektora $\boldsymbol{x}_i = \boldsymbol{X}_{[i,:]}$ i $\boldsymbol{w}_i = \boldsymbol{W}_{[i,:]}$. Rezultat konvolucije je $\boldsymbol{y} = [y_1, .., y_4] \in \mathbb{R}^4$. $\_i$ označava i-ti ulaz čvora u smjeru obrnutom od kazaljke na satu od desne strane.



Slika 5.8: Ilustracija konvolucije s korakom 2. Slike su preuzete iz Dumoulin i Visin (2016).



Slika 5.9: Ilustracija konvolucije s dilacijom 1. Slike su preuzete iz Dumoulin i Visin (2016).



Slika 5.10: Alternativni prikaz konvolucije ekivalentan onom na slici 5.7. $+\!\!+$ ovdje označava združivanje vektora $\boldsymbol{x}_i \in \mathbb{R}^n$ u vektor iz $\mathbb{R}^{3n}$, vec funkciju koja $\boldsymbol{W} \in \mathbb{R}^{3\times n}$ preslikava u $\boldsymbol{w} \in \mathbb{R}^{3n}$.

Konvolucija kao matrično množenje

Konvolucija je linearna operacija. Na slici 5.10 je konvolucija sa slike 5.7 prikazana malo drugačije. Jezgra je pretvorena u vektor, a ulaz je pretvoren u vektore koji se skalarno množe s vektorom koji predstavlja jezgru. Možemo ulaz $\boldsymbol{X}$ pretvoriti u matricu $\boldsymbol{X}_{\mathsf{M}} \in \mathbb{R}^{4\times 3n}$, a jezgru $\boldsymbol{W}$ u vektor $w \in \mathbb{R}^{3n}$ tako da njihov matrični umnožak daje izlaz konvolucije:

$$\underbrace{\begin{bmatrix} \mathbf{0}_{1\times n} & \boldsymbol{x}_1^\mathsf{T} & \boldsymbol{x}_2^\mathsf{T} \\ \boldsymbol{x}_1^\mathsf{T} & \boldsymbol{x}_2^\mathsf{T} & \boldsymbol{x}_3^\mathsf{T} \\ \boldsymbol{x}_2^\mathsf{T} & \boldsymbol{x}_3^\mathsf{T} & \boldsymbol{x}_4^\mathsf{T} \\ \boldsymbol{x}_3^\mathsf{T} & \boldsymbol{x}_4^\mathsf{T} & \mathbf{0}_{1\times n} \end{bmatrix}}_{\boldsymbol{X}_{\mathsf{M}}} \underbrace{\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \boldsymbol{w}_3 \end{bmatrix}}_{\mathrm{vec}(\boldsymbol{W})} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}}_{\boldsymbol{y}}. \tag{5.21}$$

Konvolucijski sloj obično ima više jezgri $\boldsymbol{W}_i$. Sada se lako vidi da vrijedi $\frac{\partial \boldsymbol{y}}{\partial \mathrm{vec}(\boldsymbol{W})} = \boldsymbol{X}_{\mathsf{M}}$. To možemo poopćiti na k-struku konvoluciju:

$$\boldsymbol{X}_{\mathsf{M}} \begin{bmatrix} \mathrm{vec}(\boldsymbol{W}_1) & \mathrm{vec}(\boldsymbol{W}_2) & \cdots & \mathrm{vec}(\boldsymbol{W}_k) \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_k \end{bmatrix}. \tag{5.22}$$

To se može poopćiti i na višedimenzionalnu konvoluciju (Chetlur et al., 2014). Onda su reci matrice $\boldsymbol{X}_{\mathsf{M}}$ vektori $\mathrm{vec}\Big(\mathrm{pad}\Big(\boldsymbol{X}, \frac{1}{2}(\boldsymbol{d}_{\boldsymbol{W}} - \mathbf{1})\Big)_{[\boldsymbol{t}:\boldsymbol{t}+\boldsymbol{d}^{\boldsymbol{W}}_{[1:m]}+\mathbf{1},:]}\Big)$ redom

po $\boldsymbol{t}$, uz oznake iz jednadžbe (5.19), tj. reci su vektori koji sadrže elemente ulaza koje *pokriva* jezgra za svaki položaj. Jezgra je opet vektor, a kao izlaz se dobije vektor koji treba preoblikovati tako da mu prvih m dimenzija bude jednako prvih m dimenzija ulaza.

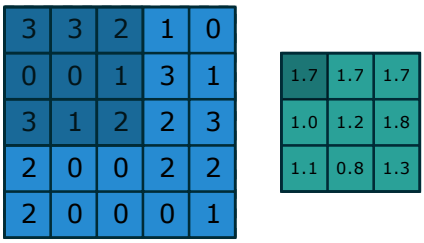
Drugi način pretvaranja konvolucije u matrično množenje je ovakav:

$$\underbrace{\begin{bmatrix} \boldsymbol{w}_2^\mathsf{T} & \boldsymbol{w}_3^\mathsf{T} & \mathbf{0}_{1\times n} & \mathbf{0}_{1\times n} \\ \boldsymbol{w}_1^\mathsf{T} & \boldsymbol{w}_2^\mathsf{T} & \boldsymbol{w}_3^\mathsf{T} & \mathbf{0}_{1\times n} \\ \mathbf{0}_{1\times n} & \boldsymbol{w}_1^\mathsf{T} & \boldsymbol{w}_2^\mathsf{T} & \boldsymbol{w}_3^\mathsf{T} \\ \mathbf{0}_{1\times n} & \mathbf{0}_{1\times n} & \boldsymbol{w}_1^\mathsf{T} & \boldsymbol{w}_2^\mathsf{T} \end{bmatrix}}_{\boldsymbol{W}_\mathsf{M}} \underbrace{\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \end{bmatrix}}_{\mathrm{vec}(\boldsymbol{X})} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}}_{\boldsymbol{y}}. \tag{5.23}$$

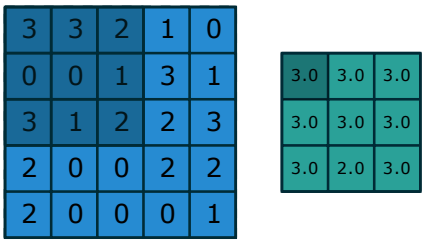
Ovdje se vidi da $\frac{\partial \boldsymbol{y}}{\partial \mathrm{vec}(\boldsymbol{X})} = \boldsymbol{W}_\mathsf{M}$. Gradijent gubitka L po ulazu je $\left(\frac{\partial L}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \mathrm{vec}(\boldsymbol{X})}\right)^\mathsf{T} = \boldsymbol{W}_\mathsf{M}^\mathsf{T} \nabla_{\boldsymbol{y}} L$. To isto odgovara jednoj vrsti konvolucije koja se naziva **transponirana konvolucija** (Šegvić, 2018).

5.4.3. Slojevi sažimanja

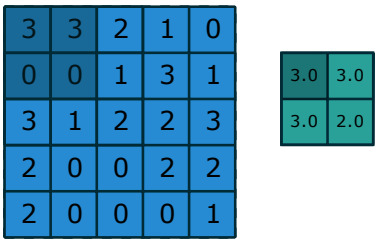
U konvolucijskim mrežama se, uglavnom radi smanjivanja dimenzija, mogu koristiti **slojevi sažimanja**. Operacije sažimanja, slično konvolucijskim slojevima, primjenjuju neku funkciju pomicanjem okna po dimenzijama konvolucije, obično s korakom većim od 1. Za razliku od konvolucijskih slojeva, oni obično djeluju na svakoj mapi značajki posebno i izlazi sažimanja su invarijantni na zamjenu elemenata unutar okna. To svojstvo se naziva **lokalna invarijantnost**. Najčešće se kao funkcija koja preslikava skup elementa okna u izlaz koristi $\max$ ili prosjek. Veličina okna je često jednaka veličini koraka tako da se susjedna okna ne preklapaju. Na slici 5.11 ilustrirani su primjeri sažimanja.



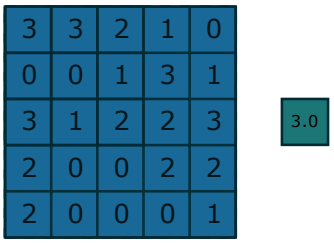
(a) Sažimanje prosječnom vrijednošću s oknom dimenzija 3×3 i korakom 1.



(b) Sažimanje maksimalnom vrijednošću s oknom dimenzija 3×3 i korakom 1.



(c) Sažimanje maksimalnom vrijednošću s oknom dimenzija 2×2 i korakom 2.



(d) Globalno sažimanje maksimalnom vrijednošću.

Slika 5.11: Ilustracije primjera dvodimenzionalnog različitih sažimanja. Slike su preuzete iz Dumoulin i Visin (2016) i prilagođene.

6. Procjenjivanje nesigurnosti kod dubokih mreža

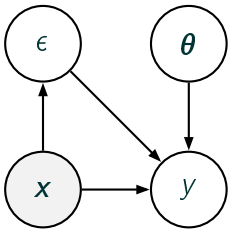
6.1. Aleatorna i epistemička nesigurnost

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 modela** i **parametre modela** (Gal, 2016). Ona se zato još nazive **nesigurnost modela**. Ona dolazi od neznanja i može se smanjiti uz više podataka.

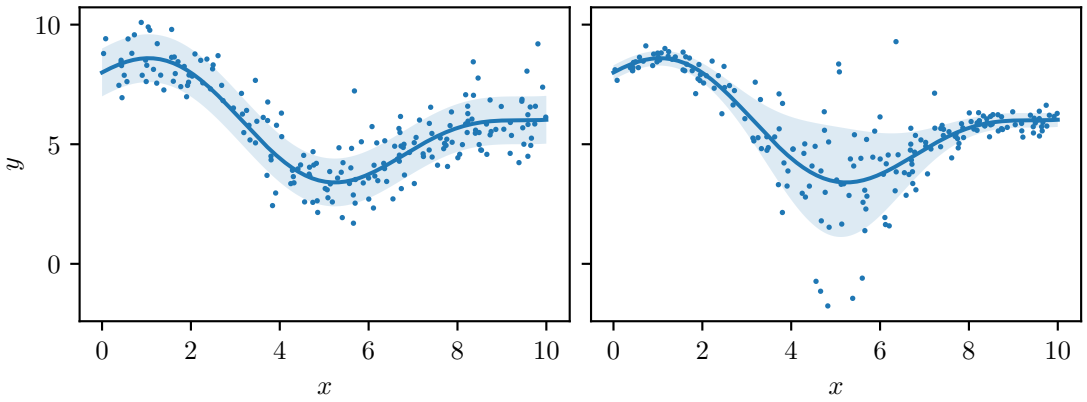
Razlikovanje aleatorne i epistemičke nesigurnosti ovisi o modelu. Nešto što je kod jednostavnijeg modela aleatorna nesigurnost, kod složenijeg modela može biti epistemičkog karaktera. Ako su neke pojave po prirodi nasumične ili se ne mogu ili ne žele modelu dati informacije koje bi ih mogle objasniti,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

Na temelju aleatorne i epistemičke nesigurnosti može se procijeniti **nesigurnost predikcije**. Kod bayesovskih modela nesigurnost predikcije izražava se razdiobom po vrijednostima varijable čija vrijednost se procjenjuje, a može se izraziti i nekom mjerom kao što je entropija ili varijanca, ovisno o tome što je prikladno.

Aleatorna nesigurnost može biti **homoskedastička** ili **heteroskedastička**. Homoskedastička nesigurnost znači da je aleatorna nesigurnost (šum) neovisna o primjeru, a heteroskedastička da ovisi o primjeru. Na slici 6.1 je prikazan primjer



Slika 6.1: Model regresije kod kojeg su θ nepoznati parametri, x opažani ulaz, y nepoznati izlaz, a ϵ heteroskedastički šum koji ovisi o ulazu x. Čvorovi koji predstavljaju podatke za učenje nisu prikazani.



Slika 6.2: 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 $s(x) = 1$.

grafičkog modela koji pretpostavlja aleatorni šum, a na slici 6.2 je ilustrirana usporedba regresijskih zadataka bez i sa šumom koji ovisi u ulaznom primjeru.

Jedan poseban slučaj nesigurnosti modela je nesigurnost u to **pripada li primjer razdiobi skupa za učenje**. Modelu se može kao ulazni primjer dati nešto za što ne bi trebala biti definirana hipoteza. Takav primjer može biti npr. slika koja pripada nekoj klasi koja nije među onima koje model treba raspoznavati ili samo slučajni šum.

Problem prepoznavanja primjera koji su izvan razdiobe skupa za učenje prirodno rješavaju generativni probabilistički modeli, ali kod složenih visokodimenzionalnih podataka to postaje problem zbog prokletstva dimenzionalnosti i složenosti modeliranja i statističkog zaključivanja. Kod diskriminativnih modela je problem to što oni ne modeliraju razdiobu ulaznih primjera $\mathrm{p}(\boldsymbol{x})$, nego samo uvjetnu razdiobu izlaza uz dani ulaz $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$.

6.2. Bayesovske neuronske mreže

Poželjno svojstvo modela strojnog učenja je mogućnost je da može prepozna

Iako duboki

Kod dubokih modela

problem visoka dimenzionalnost

7. Eksperimentalni rezultati

7.1. Programska izvedba

7.2. Skupovi podataka

8. Zaključak

Zaključak.

Literatura

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

Yoshua Bengio, Olivier Delalleau, i Nicolas Le Roux. The curse of dimensionality for local kernel machines. Technical report, 2005.

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.

Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, i Evan Shelhamer. cudnn: Efficient primitives for deep learning. **CoRR**, abs/1410.0759, 2014. URL http://arxiv.org/abs/1410.0759.

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.

Vincent Dumoulin i Francesco Visin. A guide to convolution arithmetic for deep learning, 2016. URL http://arxiv.org/abs/1603.07285.

Siniša Šegvić. Duboko učenje: Unatražno učenje konvolucijskih slojeva, 2018. URL http://www.zemris.fer.hr/~ssegvic/du/du2convnet_bp.pdf.

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.

Yarin Gal i Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. U **Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48**, ICML'16, stranice 1050–1059. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045502.

Xavier Glorot i Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. U Yee Whye Teh i Mike Titterington, urednici, **Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics**, svezak 9 od **Proceedings of Machine Learning Research**, stranice 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/glorot10a.html.

Xavier Glorot, Antoine Bordes, i Yoshua Bengio. Deep sparse rectifier neural networks. U Geoffrey J. Gordon, David B. Dunson, i Miroslav Dudík, urednici, **AISTATS**, svezak 15 od **JMLR Proceedings**, stranice 315–323. JMLR.org, 2011. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlrp15.html#GlorotBB11.

Gabriel Goh. Why momentum really works. **Distill**, 2017. doi: 10.23915/distill.00006. URL http://distill.pub/2017/momentum.

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

Ian J. Goodfellow, Jonathon Shlens, i Christian Szegedy. Explaining and harnessing adversarial examples. **CoRR**, abs/1412.6572, 2014. URL http://arxiv.org/abs/1412.6572.

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

Geoffrey Hinton. Neural networks for machine learning, lecture 6a: Overview of mini-batch gradient descent. 2012. URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, i Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. **CoRR**, abs/1207.0580, 2012. URL http://arxiv.org/abs/1207.0580.

Sergey Ioffe i Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. **CoRR**, abs/1502.03167, 2015. URL http://arxiv.org/abs/1502.03167.

Branko Jeren. Signali i sustavi: Cjelina 5, 2015. URL http://www.fer.unizg.hr/predmet/sis2.

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

Diederik P. Kingma i Jimmy Ba. Adam: A method for stochastic optimization. **CoRR**, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

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: $10.1038/\mathrm{nature14539}$. 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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, i Adrian Vladu. Towards deep learning models resistant to adversarial attacks. **CoRR**, abs/1706.06083, 2017. URL http://arxiv.org/abs/1706.06083.

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

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.

Yurii Nesterov. **Introductory Lectures on Convex Optimization: A Basic Course**. 2014.

Christopher Olah. Calculus on computational graphs: Backpropagation, 2015a. URL http://colah.github.io/posts/2015-08-Backprop/.

Christopher Olah. Visual information theory, 2015b. 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.

Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016. URL http://arxiv.org/abs/1609.04747. cite arxiv:1609.04747Comment: 12 pages, 6 figures.

D. E. Rumelhart, G. E. Hinton, i R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. poglavlje Learning Internal Representations by Error Propagation, stranice 318–362. MIT Press, Cambridge, MA, USA, 1986. ISBN 0-262-68053-X. URL http://dl.acm.org/citation.cfm?id=104279.104293.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, i Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, 15:1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Ilya Sutskever. Training recurrent neural networks. **University of Toronto, Toronto, Ont., Canada**, 2013. URL http://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, i Rob Fergus. Intriguing properties of neural networks. **CoRR**, abs/1312.6199, 2013. URL http://arxiv.org/abs/1312.6199.

D. Randall Wilson i Tony R. Martinez. The general inefficiency of batch training for gradient descent learning. **Neural Netw.**, 16(10):1429–1451, Prosinac 2003. ISSN 0893-6080. doi: 10.1016/S0893-6080(03)00138-2. URL http://dx.doi.org/10.1016/S0893-6080(03)00138-2.

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.