SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1728

Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Ivan Grubišić

Zagreb, svibanj 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
$\mathsf{\boldsymbol{a}}$, $\boldsymbol{\theta}$	Slučajni vektor ili niz
$\mathsf{\boldsymbol{A}}$, $\boldsymbol{\Theta}$	Slučajna matrica ili višedimenzionalni niz
$\mathbb{A}$	Slučajni skup ili multiskup

Konstante

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

Skupovi i nizovi

$a..b$	Kraći zapis za $a, .., b$
$\{a..b\}$	Skup cijelih brojeva od a do b
$\{a_i \colon i = 1..n\}$, $\{a_1..a_n\}$, $\{a_i\}_{i=1..n}$	Skup s n elemenata

$\{f(a)\colon P(a)\}$, $\{f(a)\}_{P(a)}$, $\{f(a)\}_a$	Skup čiji su elementi definirani preko funkcije f i predikata P koji može biti implicitan ili neodređen
$(a_i)_i$, $\left(a_{i,j}\right)_{i,j}$, $\left(a_{i,j,k}\right)_{i,j,k}$	Višedimenzionalni niz s implicitnim ili neodređenim brojem elemenata
(a, b)	Otvoreni interval
$[a, b]$	Zatvoreni interval

Indeksiranje

Indeksi elemenata vektora ili višedimenzionalnih nizova se radi jednoznačnosti mogu pisati u 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 $\boldsymbol{a}_{[i]} = a_i$.

$\boldsymbol{a}_{[i]}$	i-ti element vektora $\boldsymbol{a}$
$\boldsymbol{a}_{[i_1:i_2]}$	Vektor kojeg čine elementi $\boldsymbol{a}_{[i_1]}, \boldsymbol{a}_{[i_1+1]}, .., \boldsymbol{a}_{[i_2]}$
$\boldsymbol{a}_{\left[(i_1 .. i_n)\right]}$	Vektor kojeg čine elementi $\boldsymbol{a}_{[i_1]}, \boldsymbol{a}_{[i_2]}, .., \boldsymbol{a}_{[i_n]}$
$\boldsymbol{A}_{[i,j]}$	Element i, j matrice $\boldsymbol{A}$
$\boldsymbol{A}_{[i,:]}$	i-ti redak matrice $\boldsymbol{A}$
$\boldsymbol{A}_{[:,i_1:i_2,j]}$	2-D odsječak 3-D niza $\boldsymbol{A}$

Operacije linearne algebre

$\langle \boldsymbol{a} \vert \boldsymbol{b} \rangle$	Skalarni produkt, može biti i $\boldsymbol{a}^\mathsf{T}\boldsymbol{b}$
$\boldsymbol{a} \odot \boldsymbol{b}$	Umnožak po elementima; Hadamardov produkt
$\boldsymbol{a} \oslash \boldsymbol{b}$	Dijeljenje po elementima
$\boldsymbol{A}\boldsymbol{B}$	Matrično množenje
$\boldsymbol{A}^{-1}$	Inverz matrice
$\boldsymbol{A}^\mathsf{T}$	Transponiranje
$\mathrm{diag}(\boldsymbol{a})$	Dijagonalna matrica kojoj dijagonalu čini vektor $\boldsymbol{a}$
$\det \boldsymbol{A}$	Determinanta matrice $\boldsymbol{A}$
$\left\Vert \boldsymbol{a} \right\Vert_p$	L^p-norma vektora $\boldsymbol{a}$
$\left\Vert \boldsymbol{A} \right\Vert_p$	Matrična L^p-norma matrice $\boldsymbol{A}$
$\left\Vert \boldsymbol{A} \right\Vert_\mathrm{F}$	Frobeniusova norma matrice $\boldsymbol{A}$

Diferencijalni račun

$\frac{\mathrm{d}y}{\mathrm{d}x}$	Derivacija y po x
$\frac{\partial y}{\partial x}$	Parcijalna derivacija y po x

$\nabla_{\boldsymbol{x}} y$	Gradijent y po $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Gradijent y po $\boldsymbol{X}$
$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$, $\boldsymbol{J}_{\boldsymbol{x}\mapsto\boldsymbol{y}}$, $\boldsymbol{J}$	Jakobijan iz $\mathbb{R}^{m\times n}$ za $\boldsymbol{y}\in\mathbb{R}^m$ i $\boldsymbol{x}\in\mathbb{R}^n$
$\int_{\mathbb{A}} f(x)\,\mathrm{d}x$, $\int_{x\in\mathbb{A}} f(x)$	Određeni integral funkcije $f(x)$ po $x\in\mathbb{A}$
$\int f(x)\,\mathrm{d}x$, $\int_x f(x)$	Određeni integral funkcije $f(x)$ po $x\in\mathbb{A}$, gdje je $\mathbb{A}$ poznat iz konteksta

Teorija vjerojatnosti i teorija informacije

Svakoj slučajnoj varijabli a jednoznačno je dodijeljena jedna razdioba $\mathrm{p}(a)$ (ili $\mathrm{P}(a)$) i funkcija gustoće vjerojatnosti (poopćena funkcija) $p_a(a) = \mathrm{p}(a = a)$. Funkcija gustoće vjerojatnosti se može napisati još na 2 načina. Najkraći zapis je $p(a)$, gdje se po slovu implicitno pretpostavlja slučajna varijabla označena istim slovom bez serifa. $\mathrm{P}(A)$ označava vjerojatnost događaja A, a $P_a(a)$ funkciju vjerojatnosti slučajne varijable a. Mogu se koristiti i druge oznake za funkciju vjerojatnosti ili funkciju gustoće vjerojatnosti.

a	Slučajna varijabla
$(a \mid b = b)$, $(a \mid b)$	Uvjetna slučajna varijabla
(a, b)	Združena slučajna varijabla
$\mathcal{A}$, p, q	Razdioba ili funkcija gustoće vjerojatnosti
$\mathbb{A}$	Događaj
$\{R(a)\}$	Događaj definiran predikatorm slučajne varijable a
$\mathrm{P}(\{R(a)\})$, $\mathrm{P}(R(a))$	Vjerojatnost događaja $\{R(a)\}$
$\mathrm{P}(a)$, $\mathrm{p}(a)$	Razdioba slučajne varijable a; P ako je a diskretna slučajna varijabla, a p ako nije ili ako se ne zna
$P_a(a)$, $P(a)$	Funkcija vjerojatnosti slučajne varijable a: $P_a(a) = \mathrm{P}(a = a)$
$\mathrm{p}(a = a)$	Gustoća vjerojatnosti događaja $a = a$
$p_a(a)$, $p(a)$	Funkcija gustoće vjerojatosti slučajne varijable a
$p_{a\mid b}(a)$, $p(a \mid b)$	Gustoća vjerojatnosti za događaj $\{a = a \mid b = b\}$; $p_{a\mid b}(a) = \mathrm{p}(a = a \mid b = b)$
$p_{a,b}(a, b)$, $p(a, b)$	Gustoća vjerojatnosti za događaj $\{a = a, b = b\}$; $p_{a,b}(a, b) = \mathrm{p}(a = a, b = b)$
$a \sim \mathcal{A}$, $\mathrm{p}(a) = \mathcal{A}$	*Slučajna varijabla a ima razdiobu $\mathcal{A}$*
$a \sim \mathbb{A}$	*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)*

$a \sim \mathcal{A}$	a *se izvlači iz razidobe* $\mathcal{A}$
$a \sim A$, $a \sim \mathrm{p}(A)$	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
$\mathrm{I}(A)$	Sadržaj informacije događaja A
$\mathrm{H}(A)$	Shannonova entropija
$\mathrm{H}_B(A)$	Unakrsna entropija
$\mathrm{h}(A)$	Diferencijalna entropija
$\mathrm{D}_{\mathrm{KL}}(A \parallel B)$	Kullback-Leiblerova divergencija

Ostale oznake

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

1. Uvod

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

duboko učenje

neizvjesnost modela

primjene procjene nesigurnosti

primjena na semantičkoj segmentaciji i procjeni dubine

struktura rada

2. Osnovni pojmovi

2.1. Teorija vjerojatnosti i teorija informacije

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

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

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

2.1.1. Slučajne varijable i razdiobe

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

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

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

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

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

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

za koju vrijedi

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

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

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

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

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

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

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

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

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

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

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.

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 $\{y = y\}$:

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

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

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

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

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

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

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

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

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

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

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

što možemo i ovako zapisati:

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

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**, znanje o ishodu jedne utječe na znanje o ishodu druge, tj. uvjetna razdioba $\mathrm{p}(x \mid y = y)$ ovisi o ishodu y. Slučajne varijable x i y su **nezavisne**, što se označava $x \perp y$, akko za svaki par mogućih

vrijednosti (x, y) vrijedi

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

ili, ekvivalentno, $p(x \mid y) = p(x)$ i $p(y|x) = p(y)$. 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 ostvarenje 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. To se preko (gustoće) vjerojatnosti može ovako izraziti:

$$p(x, y \mid z) = p(x \mid z)p(y \mid z), \tag{2.12}$$

Isto tako, slučajne varijable x i y koje su nezavisne mogu biti **uvjetno zavisne** uz ostvarenje 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.

2.1.4. Očekivanje, varijanca i kovarijanca

Očekivanje (prvi moment) slučajne varijable definirano je ovako:

$$\mathbf{E}\, x \coloneqq \int x p(x)\, \mathrm{d}x, \tag{2.13}$$

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

$$\mathop{\mathbf{E}}_{x \sim x} f(x) \coloneqq \mathbf{E}\, f(x) = \int f(x)p(x)\, \mathrm{d}x. \tag{2.14}$$

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

$$\mathbf{E}[\alpha f(x) + \beta g(x)] = \alpha\, \mathbf{E}\, f(x) + \beta\, \mathbf{E}\, g(x). \tag{2.15}$$

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

ovako:

$$\mathbf{D}\,x \coloneqq \mathbf{E}\Big[(x - \mathbf{E}\,x)^2\Big] = \int (x - \mathbf{E}\,x)^2 p(x)\,\mathrm{d}x. \tag{2.16}$$

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

$$\mathbf{D}\,x = \mathbf{E}\Big[(x - \mathbf{E}\,x)^2\Big] = \mathbf{E}\Big[\Big(x^2 - 2x\,\mathbf{E}\,x + (\mathbf{E}\,x)^2\Big)\Big] \tag{2.17}$$

$$= \mathbf{E}\,x^2 - 2(\mathbf{E}\,x)^2 + (\mathbf{E}\,x)^2 = \mathbf{E}\,x^2 - (\mathbf{E}\,x)^2. \tag{2.18}$$

Drugi korijen varijance je standardna devijacija σ_x.

Kovarijanca para slučajnih varijabli definirana je ovako:

$$\mathrm{Cov}(x, y) \coloneqq \mathbf{E}[(x - \mathbf{E}\,x)(y - \mathbf{E}\,y)] = \mathbf{E}\,xy - (\mathbf{E}\,x)(\mathbf{E}\,y). \tag{2.19}$$

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

$$\mathrm{Cov}(\boldsymbol{x})_{[i,j]} = \mathrm{Cov}\Big(\boldsymbol{x}_{[i]}, \boldsymbol{x}_{[j]}\Big). \tag{2.20}$$

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 deterministički funkcijom f: $y = f(x)$. Ako su x i y diskretne slučajne varijable, onda je razdioba slučajne varijable y definirana ovako:

$$P_y(y) = \sum_{x\colon\, f(x)=y} P_x(x). \tag{2.21}$$

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

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|. \tag{2.22}$$

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

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

$$p_z(z) = \int p_{x,y}(x, z - x)\,\mathrm{d}x. \tag{2.23}$$

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

$$p_z(z) = \int p_x(x) p_y(z - x)\,\mathrm{d}x =: (p_x * p_y)(z). \tag{2.24}$$

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:

$$P(x) = \mu\,[\![x = 1]\!] + (1 - \mu)\,[\![x = 0]\!] = \mu^x (1 - \mu)^{1-x}, \tag{2.25}$$

$$\mathbf{E}\,x = \mu, \tag{2.26}$$

$$\mathbf{D}\,x = \mu(1 - \mu). \tag{2.27}$$

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.

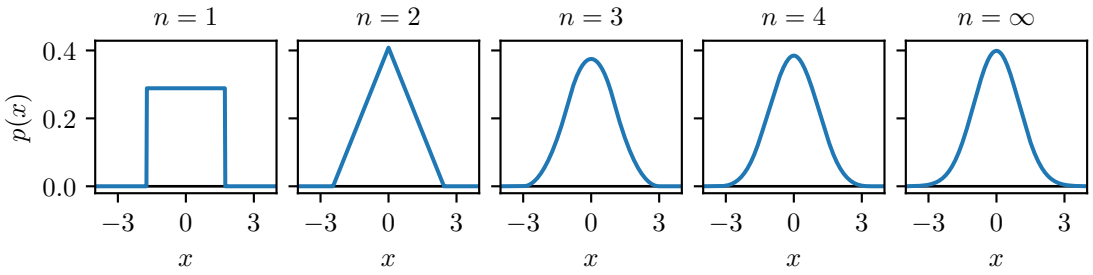
Eksponencijalna razdioba je kontinuirana razdioba s domenom $\mathbb{R}_{\geq 0}$. Ona je definirana parametrom $\lambda \in \mathbb{R}_{>0}$ ili $\beta = \lambda^{-1}$ i ima ova svojstva:

$$p(x) = \lambda \exp(-\lambda x) \tag{2.28}$$

$$\mathbf{E}\,x = \lambda^{-1}, \tag{2.29}$$

$$\mathbf{D}\,x = \lambda^{-2}. \tag{2.30}$$

Laplaceova razdioba je kontinuirana razdioba definirana parametrima $\beta \in \mathbb{R}_{>0}$ i



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.

$\mu \in \mathbb{R}$ i ima ova svojstva:

$$p(x) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta}\right) \tag{2.31}$$

$$\mathbf{E}\, x = \mu, \tag{2.32}$$

$$\mathbf{D}\, x = \beta^2. \tag{2.33}$$

Gaussova razdioba ili **normalna razdioba** je kontinuirana razdioba definirana parametrima $\mu \in \mathbb{R}$ i $\sigma \in \mathbb{R}_{>0}$ i ima ova svojstva:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.34}$$

$$\mathbf{E}\, x = \mu, \tag{2.35}$$

$$\mathbf{D}\, x = \sigma^2. \tag{2.36}$$

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.

$$\lim_{n\to\infty} \mathrm{P}(z_n < z) = \int_{-\infty}^{z} p_{\mathcal{N}(0,1)}(z')\, \mathrm{d}z'. \tag{2.37}$$

To je detaljnije objašnjeno i dokazano npr. u (Elezović, 2007). Centralni granični teorem je ilustriran na slici 2.1.

2.1.7. Teorija informacije

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

$$\mathrm{I}(x \in A) \coloneqq \log_b \frac{1}{\mathrm{P}(x \in A)} = -\log_b \mathrm{P}(x \in A). \tag{2.38}$$

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

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

$$\mathrm{H}(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(x = x) = -\mathop{\mathbf{E}}_{x} \log_b P(x). \tag{2.39}$$

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 (2.39) 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.

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

$$\mathrm{D}_{\mathrm{KL}}(x \parallel y) \coloneqq \mathop{\mathbf{E}}_{x}[\mathrm{I}(y = x) - \mathrm{I}(x = x)] \tag{2.40}$$

$$= \mathop{\mathbf{E}}_{x} \log_b \frac{P_x(x)}{P_y(x)} \tag{2.41}$$

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 $\mathrm{P}(y)$, što se

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

bolje vidi ako se ovako izrazi:

$$\mathrm{D}_{\mathrm{KL}}(x \parallel y) = \mathrm{H}_y(x) - \mathrm{H}(x), \tag{2.42}$$

gdje prvi član na desnoj strani jednadžbe označava **unakrsnu entropiju**:

$$\mathrm{H}_y(x) \coloneqq \mathop{\mathbf{E}}_{x} \mathrm{I}(y = x) = -\mathop{\mathbf{E}}_{x} \log_b P_y(x). \tag{2.43}$$

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

KL-divergencija će biti 0 akko x i y imaju iste razdiobe. Ona, kao ni unakrsna entropija, nije simetrična (slika 2.2), 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 (2.41) definirana i za kontinuirane slučajne varijable ako se funkcije vjerojatnosti zamijene funkcijama gustoće vjerojatnosti. Ona divergira kada postoji x za koji $P_x(x) > 0$ i $P_y(x) = 0$ ili, u slučaju kontinuiranih razdioba, $p_x(x) > 0$ i $p_y(x) = 0$.

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

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

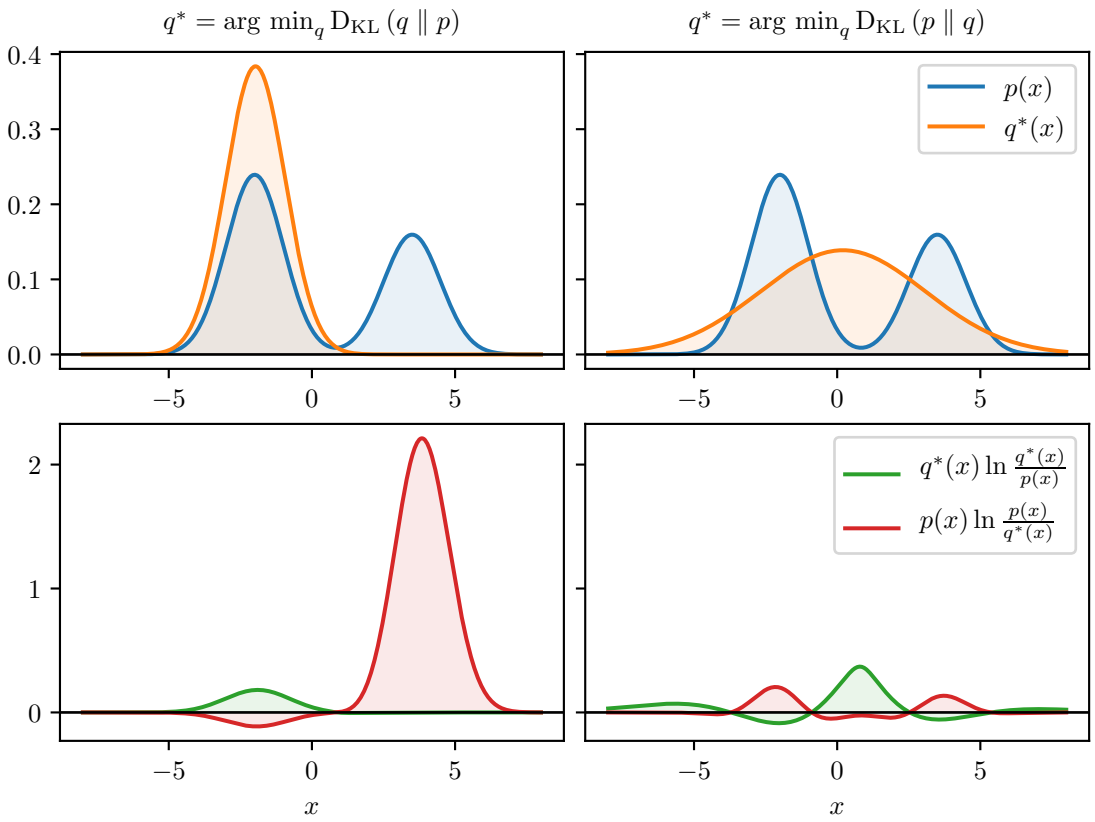
a može se i na ove načine izraziti:

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

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

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

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 $\mathrm{I}(x; y) = \mathrm{H}(x) = \mathrm{H}(y)$.



Slika 2.2: 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).

2.2. Nadzirano strojno učenje

Zadatak algoritama nadziranog strojnog učenja je preslikavanje ulaznih primjera $\boldsymbol{x} \in \mathbb{X}$ 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 što bolje **generalizira**, tj. osim primjera iz skupa za učenje, dobro preslikava i neviđene ulazne primjere u izlaze.

Neka je $\mathbb{D} = \{\boldsymbol{d}_i\}_i$ skup nezavisnih primjera izvučenih iz neke razdiobe $\mathcal{D}$. Možemo definirati **probabilistički model** $\mathbb{H}$ s nepoznatim parametrima $\boldsymbol{\theta}$ kojemu je cilj što bolje modelirati tu razdiobu pronalaskom najbolje hipoteze na temelju podataka: $p(\boldsymbol{d} \mid \mathbb{D}, \mathbb{H})$. 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 $p(\boldsymbol{y} \mid \boldsymbol{x})$ hipotezom koja ulazni primjer $\boldsymbol{x}$ preslikava u razdiobu $p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D})$. Neprobabilistički diskriminativni modeli modeliraju funkciju dodjeljivanja oznaka $f : \mathbb{X} \to \mathbb{Y}$ hipotezom $h(\boldsymbol{x})$. Modeliranje zajedničke razdiobe $p(\boldsymbol{x}, \boldsymbol{y})$ obično zahtijeva više računalnih resursa i podataka (Bishop, 2006).

2.2.1. Induktivna pristranost i komponente algoritma strojnog učenja

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 (Šnajder i Dalbelo Bašić, 2014).

Kod većine algoritama strojnog učenja možemo razlikovati 3 osnovne komponente (Šnajder i Dalbelo Bašić, 2014), 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 **funkcija 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:

$$E(\boldsymbol{\theta}, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{D}} L(h(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{y}). \tag{2.48}$$

   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 2.4.

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

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}, \mathbb{D}). \tag{2.49}$$

   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 često obuhvaća

i induktivnu pristranost ili dio nje. Riječ *model* može imati i značenje statističkog modela.

2.2.2. Kapacitet modela, prenaučenost i podnaučenost

TODO

2.2.3. Odabir modela

TODO

Murray i Ghahramani (2005)

2.2.4. Evaluacijske mjere

klasifikacija, odnos mF1 i mIoU

2.3. Procjena parametara i zaključivanje kod probabilističkih modela

2.3.1. Procjenitelji i točkaste procjene parametara

Ovaj pododjeljak se temelji na Elezović (2007).

Neka je x slučajna varijabla koju promatramo i $\mathcal{D}$ 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

$$\hat{\theta} = f(x_1, .., x_N). \tag{2.50}$$

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

$$\hat{\theta} = f(\mathbb{D}). \tag{2.51}$$

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

2.3.2. 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

$$\mathbf{E}\,\hat{\theta} = \theta. \tag{2.52}$$

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

$$\lim_{N\to\infty} \mathbf{D}\left[\hat{\theta}(\mathbb{D}_N)\right] = 0. \tag{2.53}$$

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

$$\mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbf{D}\,\hat{\theta} + \left(\mathbf{E}\,\hat{\theta}\right)^2. \tag{2.54}$$

2.3.3. Procjenitelj maksimalne izglednosti

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

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbb{D} \mid \boldsymbol{\theta}). \tag{2.55}$$

Zbog pretpostavke međusobne nezavisnosti primjera vrijedi

$$p(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{\boldsymbol{d}\in\mathbb{D}} p(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{2.56}$$

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

$$p(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(\boldsymbol{x}) \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{2.57}$$

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

2.3.4. Procjenitelj maksimalne aposteriorne vjerojatnosti

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

$$p(\boldsymbol{\theta} \mid \mathbb{D}) = \frac{p(\mathbb{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbb{D})} = \frac{p(\mathbb{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbb{D} \mid \boldsymbol{\theta}') p(\boldsymbol{\theta}') \,\mathrm{d}\boldsymbol{\theta}'}. \tag{2.58}$$

Maksimizacijom aposteriorne vjerojatnosti dobivaju se parametri

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbb{D}) = \arg\max_{\boldsymbol{\theta}} p(\mathbb{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{2.59}$$

Ovdje nije potrebno normalizirati aposteriornu vjerojatnost izračunavanjem **marginalne izglednosti** (engl. *marginal likelihood*, *evidence*) $p(\mathbb{D})$ u nazivniku na desnoj strani jednadžbe (2.58) jer ona ne ovisi $\boldsymbol{\theta}$, nego samo o modelu $\mathbb{H}$. Odabirom uniformne apriorne razdiobe MAP-procjenitelj postaje ekvivalentan ML-procjenitelju.

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

2.3.5. 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 $p(\boldsymbol{\theta} \mid \mathbb{D})$ za koju je integriranjem po svim mogućim parametrima potrebno izračunati marginalnu izglednost $p(\mathbb{D}) = \int p(\mathbb{D} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}') \,\mathrm{d}\boldsymbol{\theta}'$ iz nazivnika na desnoj strani jednadžbe (2.72).

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 $p(\mathbb{D})$ ne može ni analitički ni numerički traktabilno računati.

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

$$p(\boldsymbol{d} \mid \mathbb{D}) = \int p(\boldsymbol{d} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbb{D}) \,\mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta}\mid\mathbb{D}} p(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{2.60}$$

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

$$p(\boldsymbol{d} \mid \mathbb{D}) \approx \int p(\boldsymbol{d} \mid \boldsymbol{\theta})\delta(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} = p(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}). \tag{2.61}$$

Za diskriminativne modele se bayesovsko zaključivanje može izraziti ovako:

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

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

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbb{D}) \,\mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta}\mid\mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{2.62}$$

Kod regresije je često, ako pretpostavljamo da pogreška izlaza ima Gaussovu

razdiobu, najbolja procjena hipoteze očekivanje po naučenoj razdiobi parametara (Neal, 1995):

$$h(\boldsymbol{x}) = \mathop{\mathbf{E}}_{\boldsymbol{\theta}|\mathbb{D}} h(\boldsymbol{x};\boldsymbol{\theta}) = \int h(\boldsymbol{x};\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbb{D}) \,\mathrm{d}\boldsymbol{\theta}. \tag{2.63}$$

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

2.4. Minimizacija rizika

2.4.1. 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. Očekivanje funkcije gubitka je (frekventistički) **rizik** (Murphy, 2012):

$$R(\boldsymbol{\theta}, \mathcal{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}). \tag{2.64}$$

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

$$R_{\mathrm{E}}(\boldsymbol{\theta}; \mathbb{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}). \tag{2.65}$$

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 $h(\boldsymbol{x};\theta) = p(\boldsymbol{x} \mid \theta)$, gubitak mjeri **pogrešku rekonstrukcije** i izraz za rizik je (Murphy, 2012):

$$R(\boldsymbol{\theta}; \mathcal{D}) = \mathop{\mathbf{E}}_{\boldsymbol{d}\sim\mathcal{D}} L(h(\boldsymbol{d};\boldsymbol{\theta}), \boldsymbol{d}). \tag{2.66}$$

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

$$R_{\mathrm{E}}(\boldsymbol{\theta}; \mathbb{D}) = -\ln p(\mathbb{D} \mid \boldsymbol{\theta}) = -\sum_{\boldsymbol{d}\in\mathbb{D}} \ln p(\boldsymbol{d} \mid \boldsymbol{\theta}), \tag{2.67}$$

tj. gubitak je onda $L(h(\boldsymbol{d};\boldsymbol{\theta}), \boldsymbol{d}) = -\ln p(\boldsymbol{d} \mid \boldsymbol{\theta})$. U slučaju diskriminativnog modela, uz zanemarivanja faktora izglednosti koji ne ovisi o $\boldsymbol{\theta}$ (jednadžba (2.57)),

vrijedi $L(h(\boldsymbol{x};\boldsymbol{\theta}),\boldsymbol{y}) = -\ln p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{\theta})$.

2.4.2. Strukturni rizik

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 $\lambda R_{\mathrm{R}}(\boldsymbol{\theta})$, $\lambda \geq 0$, koji predstavlja **strukturni rizik** koji daje prednost jednostavnijim hipotezama:

$$E(\boldsymbol{\theta};\mathbb{D}) = R_{\mathrm{E}}(\boldsymbol{\theta};\mathbb{D}) + \lambda R_{\mathrm{R}}(\boldsymbol{\theta}). \tag{2.68}$$

Kod opisanih modela koji uključuju apriorno znanje, regularizacijskom članu uz $\lambda = 1$ odgovara negativni logaritam apriorne vjerojatnosti parametara: $R_{\mathrm{R}}(\boldsymbol{\theta}) = -\frac{1}{|\mathbb{D}|}\ln p(\boldsymbol{\theta})$. λ različit od 1 odgovara izmjeni entropije apriorne razdiobe $\mathrm{H}\Big(p(\boldsymbol{\theta})^{\lambda}/Z\Big)$, tj. s većim λ apriorna razdioba postaje koncentriranija i regularizacija jača. Jačom regularizacijom se povećava pristranost i smanjuje varijanca procjenitelja.

2.5. Probabilistički grafički modeli

2.5.1. Bayesovski modeli

DL 3.14

2.6. Monte Carlo aproksimacija

Ovaj odjeljak temelji se 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

zbroj ili integral. Ako se on može ovako izraziti:

$$s = \sum_x p(x) f(x) = \mathbf{E}\, f(x) \tag{2.69}$$

ili

$$s = \int p(x) f(x)\, \mathrm{d}x = \mathbf{E}\, f(x), \tag{2.70}$$

on se može procijeniti uzorkovanjem:

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{2.71}$$

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 $\mathbf{D}\, \hat{s}_n = \frac{1}{n} \mathbf{D}\, f(x)$.

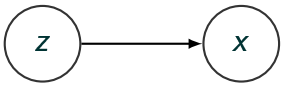
U širem smislu, postupci *Monte Carlo* obuhvaćaju i generiranje uzoraka slučajne varijable čije se očekivanje procjenjuje.

2.7. Aproksimacija razdioba i aproksimacijsko zaključivanje

Ovaj odjeljak uglavnom se 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[3] aposteriorna razdioba se ne može lako izračunati i treba koristiti aproksimacijske postupke od kojih su glavni **varijacijski** postupci (Jordan et al., 1999) i **Monte Carlo** postupci koji se temelje na uzorkovanju pomoću **Markovljevog lanca** (MCMC, engl. *Markov chain Monte Carlo*) i Hamiltonovski (ili hibridni) MC-postupci (HMC). MCMC i HMC temelje se na definiranju stohastičkog procesa koji ima stacionarnu razdiobu jednaku razdiobi koja se aproksimira, omogućuju asimptotski egzaktno uzorkovanje i bolje su istraženi. Varijacijski postupci temelje se na aproksimaciji razdiobe nekom jednostavnijom koja se pronalazi rješavanjem optimizacijskog problema, brži su i jednostavniji za ostvariti za složenije modele.

[3] Bayesovski model (ili Bayesova mreža) je probabilistički grafički model sa strukturom usmjerenog acikličkog grafa.



Slika 2.3: Prikaz grafičkog modela čija je združena razdioba $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z})$.

Razmatramo bayesovski model koji ima jednu latentnu varijablu z i jednu vidljivu varijablu x. Model je prikazan an slici 2.3 i opisan je ovom jednadžbom združene vjerojatnosti:

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

Zaključivanjem se određuje aposteriorna razdioba latentne varijable:

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\int p(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}. \tag{2.72}$$

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

2.7.1. Varijacijsko zaključivanje

Za razliku od uzorkovanja kod MCMC-postupaka, osnovna ideja kod varijacijskog zaključivanja je optimizacija. Prvo se odabire familija razdioba $\mathbb{Q} = \{p(\tilde{\boldsymbol{z}})\}_{\tilde{\boldsymbol{z}}} = \left\{\mathcal{Q}_{\boldsymbol{\phi}}\right\}_{\boldsymbol{\phi}}$ koje su lakše za računanje. Razdiobe iz $\mathbb{Q}$ su parametrizirane tzv. **varijacijskim parametrima** $\boldsymbol{\phi}$. Cilj je na temelju podataka kao zamjenu za aposteriornu razdiobu $p(\boldsymbol{z} \mid \boldsymbol{x})$ 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:

$$\mathcal{Q}^* = \underset{p(\tilde{\boldsymbol{z}}) \in \mathbb{Q}}{\arg\min} \, \mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\boldsymbol{z} \mid \boldsymbol{x})). \tag{2.73}$$

Ovakva ciljna funkcija obično se ne može lako izračunati jer zahtijeva računanje

marginalne izglednosti $p(\boldsymbol{x})$ iz nazivnika u jednadžbi (2.72) marginalizacijom po z:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\mathsf{z} \mid \boldsymbol{x})) &= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln \frac{p(\tilde{\boldsymbol{z}})}{p(\mathsf{z}{=}\tilde{\boldsymbol{z}} \mid \boldsymbol{x})} \\ &= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\mathsf{z}{=}\tilde{\boldsymbol{z}} \mid \boldsymbol{x}) \\ &= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\mathsf{z}{=}\tilde{\boldsymbol{z}}, \boldsymbol{x}) + \ln p(\boldsymbol{x}). \end{aligned} \tag{2.74}$$

Marginalna izglednost se može zanemariti jer marginalna izgledost ne ovisi o parametrima po kojima se optimizira pa umjesto minimizacije KL-divergencije maksimiziramo funkciju koja se naziva **varijacijska donja granica** (engl. *variational lower bound*) ili **donja granica (logaritma) marginalne izglednosti** (engl. *(log) evidence lower bound*, *ELBO*):

$$L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}}) \coloneqq \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\mathsf{z}{=}\tilde{\boldsymbol{z}}, \boldsymbol{x}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}). \tag{2.75}$$

Vrijedi $\mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\mathsf{z} \mid \boldsymbol{x})) = -L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}}) + \ln p(\boldsymbol{x})$. Varijacijska donja granica se može i ovako izraziti:

$$L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}}) = \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\boldsymbol{x} \mid \mathsf{z}{=}\tilde{\boldsymbol{z}}) - \mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel \mathsf{z}). \tag{2.76}$$

Maksimiziranje takve ciljne funkcije daje razdiobu $p(\tilde{\boldsymbol{z}})$ koja dobro objašnjava podatke jer se potiče veće očekivanje logaritma izglednosti, i 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 $\ln p(\boldsymbol{x}) \geq L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}})$ preko Jensenove nejednakosti. Ta nejednakost slijedi i iz prethodne jednadžbe i nenegativnosti KL-divergencije:

$$\ln p(\boldsymbol{x}) = L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}}) + \mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\mathsf{z} \mid \mathsf{x})) \geq L_{\boldsymbol{x}}(\tilde{\boldsymbol{z}}). \tag{2.77}$$

TODO mean-field, calculus of variations

2.7.2. Monte Carlo aproksimacija

1.3.1 Neal (1995).

3. Duboko učenje i konvolucijske mreže

3.1. Duboke neuronske mreže

3.2. Konvolucijske mreže

3.3. Optimizacija

3.3.1. Propagacija pogreške unatrag

3.3.2. Isključivanje neurona - dropout

3.3.3. Normalizacija po grupama

4. Procjenjivanje nesigurnosti

4.1. Aleatorna i epistemička nesigurnost

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

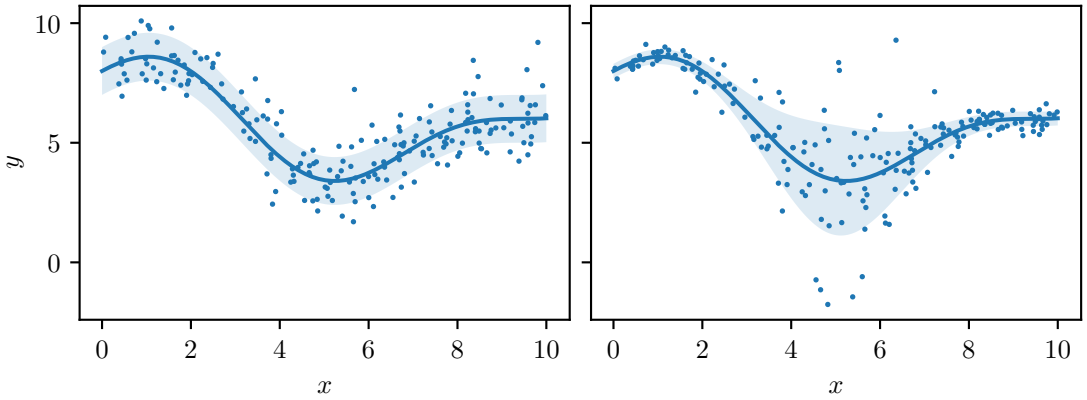
Postoje različiti izvori nesigurnosti (C. Kennedy i O'Hagan, 2002), ali nesigurnost općenito možemo podijeliti na dvije vrste (Kiureghian i Ditlevsen, 2009): **aleatornu nesigurnost** i **epistemičku nesigurnost**. Riječ *aleatorna* izvedena je vjerojatno od latinske riječi *aleator* (Gal, 2016) koja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

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

TODO: homoskedastička, heteroskedastička nesigurnost

TODO: model uncertainty Gal-thesis 1.2

5.33331pt
11.74983pt
small 10.95pt
footnotesize 10.0pt



Slika 4.1: Homoskedastički (lijevo) i heteroskedastički (desno) Gaussov šum. Crta prikazuje očekivanje $f(x)$, svjetloplava površina standardnu devijaciju šuma $s(x)$, a točke slučajne uzorke. Točke su generirane prema $(y \mid x) \sim \mathcal{N}(f(x), s(x)^2)$. Na lijevoj slici je $s(x) = 1$.

412.56497pt

5. Bayesovske neuronske mreže

6. Procjenjivanje nesigurnosti kod konvolucijskih mreža

7. Eksperimentalni rezultati

7.1. Skupovi podataka

8. Zaključak

Zaključak.

Literatura

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.

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

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

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

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

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

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

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

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

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

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

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

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

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.

A. Izvod donje varijacijske granice

The contents...