SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1728

Nadzirani pristupi za procjenu nesigurnosti predikcija dubokih modela

Ivan Grubišić

Zagreb, travanj 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 velikim podebljanim slovima, a skupovi slovima s udvostručenim linijama.

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

Konstante

$\mathbf{0}$	Nul-vektor
$\mathbf{I}$, $\mathbf{I}_n$	Matrica identiteta (s n redaka i stupaca)
$\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{C}$	Poznati skup

Skupovi i nizovi

$a \,..\, b$	Kraći zapis za $a, \ldots, b$
$\{a \,..\, b\}$	Skup cijelih brojeva od a do b
$\{a_i \mid i = 1 \,..\, n\}$, $\{a_1 \,..\, a_n\}$, $\{a_i\}_{i=1\,..\,n}$	Skup s n elemenata
$\{f(a) \mid P(a)\}$, $\{f(a)\}_{P(a)}$, $\{f(a)\}_a$	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 pišu u zagradama. Npr. ako je definiran vektor $\boldsymbol{a} = (a_1 \,..\, a_n)$, 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 $\boldsymbol{a}_{(i_1)}, \boldsymbol{a}_{(i_1+1)}, \ldots, \boldsymbol{a}_{(i_2)}$
$\boldsymbol{A}_{[i,j]}$	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 $\boldsymbol{a}$
$\vert \boldsymbol{A} \vert$	Determinanta matrice $\boldsymbol{A}$
$\Vert \boldsymbol{a} \Vert_p$	L^p-norma vektora $\boldsymbol{a}$
$\Vert \boldsymbol{A} \Vert_p$	Matrična L^p-norma matrice $\boldsymbol{A}$
$\Vert \boldsymbol{A} \Vert_\mathrm{F}$	Frobeniusova norma matrice $\boldsymbol{A}$

Diferencijalni račun

$\int_\mathbb{A} f(x)\,\mathrm{d}x$, $\int_{x\in\mathbb{A}} f(x)$	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 $p(a)$ i funkcija gustoće vjerojatnosti $p_a(a)$. 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. Diskretna razdioba s funkcijom vjerojatnosti $P(a)$ može se predstaviti kontinuiranom razdiobom s funkcijom gustoće vjerojatnosti $p(a) = \sum_{a'\in\mathbb{A}} P(a)\delta(a - a')$, gdje je $\mathbb{A}$ skup mogućih vrijednosti varijable a.

a	Slučajna varijabla
$(a \mid b{=}b)$, $(a \mid b)$	Uvjetna slučajna varijabla
(a, b)	Združena slučajna varijabla
$\mathcal{A}$	Razdioba
$\{R(a)\}$	Događaj koji uključuje slučajnu varijablu a, gdje je R neki predikat
$P(\{R(a)\})$, $P(R(a))$	Vjerojatnost događaja $\{R(a)\}$
$P(a)$	Razdioba diskretne slučajne varijable a
$P(a)$	Funkcija vjerojatnosti diskretne slučajne varijable a
$p(a)$	Razdioba slučajne varijable a
$p_a(a)$, $p(a{=}a)$, $p(a)$	Gustoća vjerojatnosti za događaj $\{a = a\}$ (vjerojatnost ako je a ima diskretnu razdiobu)
$p_{a\mid b}(a)$, $p(a{=}a \mid b{=}b)$, $p(a \mid b)$	Gustoća vjerojatnosti za događaj $\{a{=}a \mid b{=}b\}$
$p_{a,b}(a, b)$, $p(a{=}a, b{=}b)$, $p(a, b)$	Gustoća vjerojatnosti za događaj $\{a{=}a, b{=}b\}$
$a \sim \mathcal{A}$, $p(a) = \mathcal{A}$	*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 $\mathcal{A}$*
$a \sim a$, $a \sim p(a)$	*a se izvlači iz razidobe $p(a)$*
$\mathbf{E}_{a\sim a}\, f(a)$, $\mathbf{E}_a\, f(a)$	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
$\mathrm{H}(a)$	Shannonova entropija
$\mathrm{H}(a, b)$	Unakrsna entropija
$\mathrm{D}_{\mathrm{KL}}(a \parallel b)$	Kullback-Leiblerova divergencija

Funkcije i operatori

$f\colon \mathbb{A} \to \mathbb{B}$	Funkcija s domenom $\mathbb{A}$ i kodomenom $\mathbb{B}$
$\delta(\cdot)$	Diracova delta razdioba; poopćena funkcija za koju vrijedi $\delta(x) = 0$ za $x \neq 0$ i $\int_x \delta(x)\,\mathrm{d}x = 1$
$[\![\cdot]\!]$	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 informaicije

Jako važan pojam u strojnom učenju je neizvjesnost. Ona dolazi od šuma u mjerenju i iz konačnosti skupa podataka (Bishop, 2006). Teorija vjerojatnosti nam omogućuje modelirati neizvjesnost pronalaziti optimalne zaključke korištenjem dostupnih informacija. Ovaj odjeljak daje kratak pregled nekih od osnovnih pravila vjerojatnosti.

2.1.1. Vjerojatnost, razdioba i slučajna varijabla

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 još se naziva i **prostor elementarnih događaja**. Elementarni događaj je ostvarenje neke vrijednosti iz prostora elementarnih događaja i, ako je a slučajna varijabla za koju se u nekom eksperimentu opaža vrijednost a, taj događaj ima zapis $\{a = a\}$, a njegova vjerojatnost $P(\{a = a\})$ ili, kraće, $P(a = a)$. Događaj može biti općenitiji, npr. $\{a < a\}$, i općenito se može izraziti predikatom: $\{R(a)\}$.

Razlikujemo diskretne i kontinuirane razdiobe.

Neka je a slučajna varijabla koja može poprimiti vrijednosti iz skupa $\mathbb{A}$. Ako je $\mathbb{A}$ konačan skup, a ima diskretnu razdiobu $P(a)$ s funkcijom vjerojatnosti $P_a(a)$. Ako je $\mathbb{A}$ beskonačan skup, a ima kontinuiranu razdiobu $p(a)$ s funkcijom gustoće vjerojatnosti $p_a(a)$. Za diskretnu razdiobu mora vrijediti $\sum_a P_a(a) = 1$, a za kontinuiranu $\int_a p_a(a) = 1$.

U popisu oznaka na početku rada prikazane su još neke oznake

Ovdje funkcije gustoće vjerojatnosti smatramo poopćenim funkcijama, što znači

da za neke vrijednosti gustoća može biti beskonačna (tj. vjerojatnost veća od 0), ali i dalje mora vrijediti $\int_a p(a) = 1$. Svaka razdioba definirana je svojom funkcijom (gustoće) vjerojatnosti.

događaj

Svakoj slučajnoj varijabli je jednoznačno dodijeljena jedna razdioba.

Dvije slučajne varijable koje imaju istu razdiobu ne moraju biti u istom odnosu prema drugim slučajnim varijablama. Npr. ako $a_1 \sim \mathcal{A}$, $a_2 \sim \mathcal{A}$ i $b \sim \mathcal{B}$, ne mora vrijediti $p(a_1, b) = p(a_2, b)$.

2.1.2. Teorija informacije

***

$$\mathrm{D}_{\mathrm{KL}}(a \parallel b) = \mathop{\mathbf{E}}_{a \sim a} \ln \frac{p_a(a)}{p_b(a)} = \int_a p(a) \ln \frac{p_a(a)}{p_b(a)} \tag{2.1}$$

$$\lim_{\varepsilon \to 0} \int_{-\varepsilon}^{\varepsilon} \delta(x_0 + \varepsilon') \, \mathrm{d}\varepsilon' = 1 \tag{2.2}$$

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 $\mathbb{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i,j}$. 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 koji izravno modeliraju uvjetne razdiobe $p(\boldsymbol{y} \mid \boldsymbol{x})$ hipotezom oblika $p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D})$. Neprobabilistički diskriminativni modeli modeliraju funkciju dodjeljivanja oznaka $f \colon \mathbb{X} \to \mathbb{Y}$ 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, učenjem se traže parametri koji do kraja definiraju traženu 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 izlaz hipoteze, a drugi ciljna oznaka koja odgovara ulaznom primjeru. Funkciju pogreške možemo definirati kao prosječni gubitak na skupu za učenje:

$$E(\boldsymbol{\theta}, \mathbb{D}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}). \tag{2.3}$$

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:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} E(\boldsymbol{\theta}, \mathbb{D}). \tag{2.4}$$

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.

2.2.2. Kapacitet modela, prenaučenost i podnaučenost

TODO

2.2.3. Odabir modela

TODO Murray i Ghahramani (2005)

2.3. Procjena parametara i zaključivanje kod probabilističkih modela

2.3.1. Procjenitelj

TODO: definicija, poželjna svojstva, SU 3.3 TODO: varijanca i pristranost

2.3.2. Procjenitelj maksimalne izglednosti

Procjenitelj maksimalne izglednosti (**ML-procjenitelj**, engl. *maximum likelihood*) je statistika koja uzorku dodjeljuje parametre koji imaju najveću **izglednost**:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbb{D} \mid \boldsymbol{\theta}). \tag{2.5}$$

Zbog pretpostavke međusobne nezavisnosti primjera vrijedi

$$p(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{\boldsymbol{d} \in \mathbb{D}} p(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{2.6}$$

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

$$p(\mathbb{D} \mid \boldsymbol{\theta}) = \prod_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{x} \mid \boldsymbol{\theta}) = p(\boldsymbol{x}) \prod_{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{2.7}$$

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

2.3.3. 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

$$p(\boldsymbol{\theta} \mid \mathbb{D}) = \frac{p(\mathbb{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbb{D})} = \frac{p(\mathbb{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbb{D} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'}. \tag{2.8}$$

Maksimizacijom aposteriorne vjerojatnosti dobivaju se parametri

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbb{D}) = \arg\max_{\boldsymbol{\theta}} p(\mathbb{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.9}$$

Ovdje nije potrebno normalizirati aposteriornu vjerojatnost izračunavanjem marginalne izglednosti $p(\mathbb{D})$ u nazivniku na desnoj strani jednadžbe (2.8) 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 **konjugatnim razdiobama** (Šnajder i Dalbelo Bašić, 2014).

2.3.4. 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 potrebno izračunati integral $p(\mathbb{D}) = \int p(\mathbb{D} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}'$ iz nazivnika na desnoj strani jednadžbe (2.19).

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

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

$$p(\boldsymbol{d} \mid \mathbb{D}) = \int p(\boldsymbol{d} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbb{D})\,\mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta}|\mathbb{D}} p(\boldsymbol{d} \mid \boldsymbol{\theta}). \tag{2.10}$$

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

$$p(\boldsymbol{d} \mid \mathbb{D}) \approx \int p(\boldsymbol{d} \mid \boldsymbol{\theta}) \delta(\hat{\boldsymbol{\theta}} - \theta) \, \mathrm{d}\boldsymbol{\theta} = p(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}). \tag{2.11}$$

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

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathbb{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbb{D}) \, \mathrm{d}\boldsymbol{\theta} = \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}). \tag{2.12}$$

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

$$h(\boldsymbol{x}) = \mathop{\mathbf{E}}_{\boldsymbol{\theta} \mid \mathbb{D}} h(\boldsymbol{x}; \boldsymbol{\theta}) = \int h(\boldsymbol{x}; \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbb{D}) \, \mathrm{d}\boldsymbol{\theta}. \tag{2.13}$$

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

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):

$$R(\boldsymbol{\theta}, \mathcal{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} L(h(\boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{y}). \tag{2.14}$$

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

$$R_{\mathrm{E}}(\boldsymbol{\theta}; \mathbb{D}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}) = \frac{1}{|\mathbb{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}} L(h(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}). \tag{2.15}$$

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):

$$R(\boldsymbol{\theta}; \mathcal{D}) = \mathop{\mathbf{E}}_{\boldsymbol{d}\sim\mathcal{D}} L(h(\boldsymbol{d};\boldsymbol{\theta}), \boldsymbol{d}). \tag{2.16}$$

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

$$R_{\mathrm{E}}(\boldsymbol{\theta}; \mathbb{D}) = -\ln p(\mathbb{D} \mid \boldsymbol{\theta}) = -\sum_{\boldsymbol{d}\in\mathbb{D}} \ln p(\boldsymbol{d} \mid \boldsymbol{\theta}), \tag{2.17}$$

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 (2.7)), 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:

$$E(\boldsymbol{\theta}; \mathbb{D}) = R_{\mathrm{E}}(\boldsymbol{\theta}; \mathbb{D}) + \lambda R_{\mathrm{R}}(\boldsymbol{\theta}). \tag{2.18}$$

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

2.6. Bayesovski modeli

DL 3.14

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[1] 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.

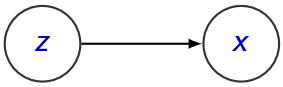
Razmatramo bayesovski model koji ima jednu latentnu varijablu z i jednu vidljivu varijablu x. Model je prikazan an slici 2.1 i opisan je ovom jednadžbom združene vjerojatnosti:

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

Zaključivanjem se određuje aposteriorna razdioba latentne varijable

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{\int p(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}. \tag{2.19}$$

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



Slika 2.1: Prikaz grafičkog modela čija je združena razdioba $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x} \mid \boldsymbol{z})$.

na temelju opažanih vrijednosti 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 $\mathbb{Q} = \{p(\tilde{\boldsymbol{z}})\}_{\tilde{\boldsymbol{z}}} = \left\{p(\tilde{\boldsymbol{z}}_\phi)\right\}_\phi$ koje su lakše za računanje. Razdiobe iz $\mathbb{Q}$ su parametrizirane tzv. **varijacijskim parametrima** ϕ. 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:

$$p(\tilde{\boldsymbol{z}}^*) = \arg\min_{p(\tilde{\boldsymbol{z}}) \in \mathbb{Q}} \mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\boldsymbol{z} \mid \boldsymbol{x})). \tag{2.20}$$

Ovakva ciljna funkcija obično se ne može lako izračunati jer zahtijeva računanje marginalne izglednosti $p(\boldsymbol{x})$ iz nazivnika u jednadžbi (2.19) marginalizacijom po z, što može biti netraktabilno:

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\tilde{\boldsymbol{z}} \parallel (\boldsymbol{z} \mid \boldsymbol{x})) &= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln \frac{p(\tilde{\boldsymbol{z}})}{p(\boldsymbol{z}{=}\tilde{\boldsymbol{z}} \mid \boldsymbol{x})} \\
&= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\boldsymbol{z}{=}\tilde{\boldsymbol{z}} \mid \boldsymbol{x}) \\
&= \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\boldsymbol{z}{=}\tilde{\boldsymbol{z}}, \boldsymbol{x}) + \ln p(\boldsymbol{x}).
\end{aligned} \tag{2.21}$$

Umjesto minimizacije KL-divergencije, možemo maksimizirati funkciju koja se naziva **varijacijska donja granica** (engl. *variational lower bound*) ili **donja granica (logaritma) marginalne izglednosti** (ELBO, engl. *(log) evidence lower bound*):

$$\mathrm{ELBO}(\tilde{\boldsymbol{z}}) \coloneqq \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\boldsymbol{z}{=}\tilde{\boldsymbol{z}}, \boldsymbol{x}) - \mathop{\mathbf{E}}_{\tilde{\boldsymbol{z}}} \ln p(\tilde{\boldsymbol{z}}). \tag{2.22}$$

To se i ovako može zapisati:

$$\mathrm{ELBO}(\tilde{z}) = \mathop{\mathbf{E}}_{\tilde{z}} \ln p(\boldsymbol{x} \mid z = \tilde{\boldsymbol{z}}) - \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel z). \tag{2.23}$$

Maksimiziranje takve ciljne funkcije daje razdiobu $p(\tilde{z})$ koja dobro obješnjava podatke (maksimizacija očekivanje logaritma izglednosti) i nije previše daleko od apriorne razdiobe (minimizacija KL-divergencije između varijacijske razdiobe i apriorne razdiobe) (Gal i Ghahramani, 2015). Zamjenom prva dva člana u jednadžbi (2.21), KL-divergencija se može ovako zapisati:

$$\mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel (z \mid \boldsymbol{x})) = -\mathrm{ELBO}(\tilde{z}) + \ln p(\boldsymbol{x}). \tag{2.24}$$

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

$$\ln p(\boldsymbol{x}) = \mathrm{ELBO}(\tilde{z}) + \mathrm{D}_{\mathrm{KL}}(\tilde{z} \parallel (z \mid x)) \geq \mathrm{ELBO}(\tilde{z}). \tag{2.25}$$

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.

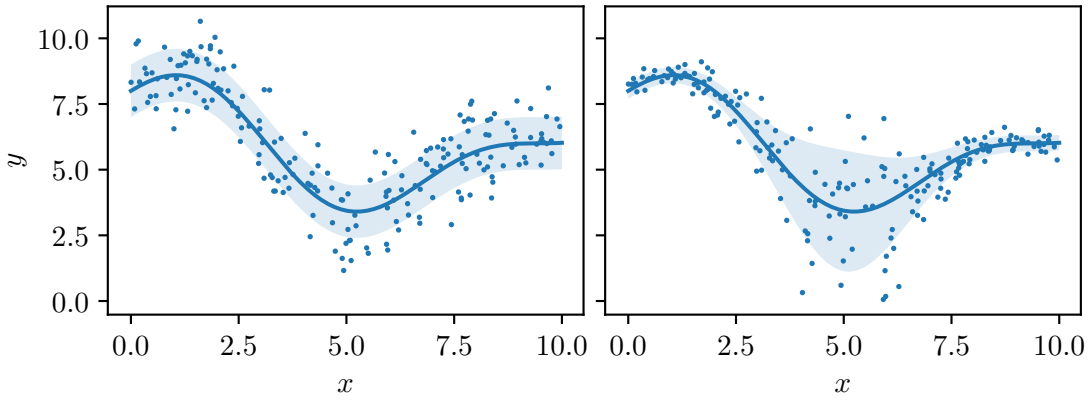
Postoje različiti izvori nesigurnosti (C. Kennedy i O'Hagan, 2002), ali nesigurnost općenito možemo podijeliti na dvije vrste (Kiureghian i Ditlevsen, 2009): **aleatornu nesigurnost** i **epistemičku nesigurnost**. Riječ *aleatorna* izvedena je od latinske riječi *alea* koja znači *kocka* i asocira na nasumičnost bacanja kocke, 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 parametre modela. 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

5.33331pt
11.74983pt

412.56497pt



Slika 4.1: Homoskedaktički (lijevo) i heteroskedaktički (desno) Gaussov šum. Crta prikazuje očekivanje $f(x)$, svijetloplava 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 $s(x) = 1$.

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.

Yarin Gal i Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2015. URL http://arxiv.org/abs/1506.02142.

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.

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...