SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Nesigurnost predikcije u modelima dubokog učenja

Ivan Grubišić
Voditelj: *prof.dr.sc. Bojana Dalbelo Bašić*
Mentori: *dr.sc. Tomislav Lipić i prof.dr.sc. Sonja Grgić*

Zagreb, prosinac 2021.

SADRŽAJ

1. Uvod

U poslijednjem desetljeću dogodio se nevjerojatni napredak u području strojnog učenja (eng. machine learning, ML), a posebno je napredovalo područje dubokog učenja (eng. deep learning, DL). Taj uspjeh omogućila su postignuća u raznim neovisnim područjima, kao što su: (1) računalne komponente - procesorska (CPU) i grafička (GPU) snaga, te dostupnost radne memorije (RAM); (2) programska podrška - programski okviri za razvoj DL modela, kao što su TensorFlow[2], PyTorch[47, 48], Caffe[32], MXNet[10]); 3) podatkovni skupovi - količina, kvaliteta (npr. rezolucija), razina detalja (anotacija); 4) metode - nove arhitekture modela, pristupi učenju. Jedna od glavnih prekretnica u modernom razvoju MLa dogodila se 2012. godine kada su Krizhevsky, Sutskever i Hinton osvojili ILSVRC 2012 [11] natjecanje korištenjem duboke neuronske mreže (eng. deep neural network, DNN), poznatom pod imenom AlexNet [35]. Prateći njihov uspjeh mnogi drugi znanstvenici su odlučili dalje razvijati DNN modele s ciljem rješavanja zadataka koji zahtjevaju čovjekovu razinu inteligencije, kao što su računalni vid (eng. computer vision, CV) i obrada prirodnog jezika (eng. natural language processing, NLP).

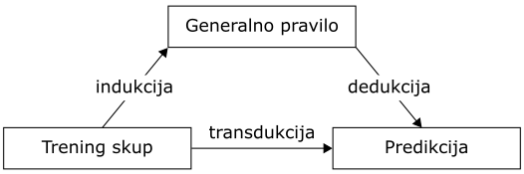
Danas su DNN modeli jako zastupljeni u automatskom donošenju odluka visokih uloga, gdje kriva odluka može imati teške posljedice. Dobar primjer je medicinska dijagnostika koja može dovesti i do smrtnog slučaja pacijenta. Upravo iz tog razloga neophodno je smanjiti pogrešku donošenja kritičnih odluka na minimum. Ovu pogrešku možemo značajno smanjiti ako uz predikciju (automatsku odluku) možemo dobiti i točnu razinu nesigurnosti uz tu predikciju. Na ovaj način možemo automatsko odlučivanje koristiti samo ako je sigurnost u odluku vrlo visoka, te proslijediti nesigurne slučajeve na stručnu procjenu (npr. doktoru specijalistu).

Nesigurnost predikcije modela možemo dobiti na razne načine. U konačnici sve te pristupe možemo podijeliti na dvije vrste: (1) metode temeljene na podacima; i (2) metode temeljene na modelu. U metode temeljene na podacima spadaju konformalni prediktori (eng. conformal predictors) u kojima prediktivni model procjenjuje svoju nesigurnost na osnovu prethodnog iskustva, tj. poznatih uzoraka [39]. S druge strane,

metode bazirane na modelu izračunavanju nesigurnost predikcije testnog uzorka bez usporedbe s poznatim uzorcima. Kod tih metoda nesigurnost se procjenjuje provedbom posebno dizajniranih testova na razini modela, kao što su Monte Carlo dropout i ansambli modela. Vrlo je važno da je procjena nesigurnosti što točnija. Zato je ključna kalibracija modela koja omogućuje da izlaz iz modela bolje predstavlja njegovu prediktivnu sigurnost[42].

Postoje razni izvori nesigurnosti [19], a u literaturi se najviše spominju podatkovna i modelska nesigurnost. Modelska ili epistemička nesigurnost predstavlja nesigurnost u procjenu parametara modela, te se smanjuje s povećanjem broja uzoraka za učenje [19]. Modelska nesigurnost pokazuje koliko je dobro model naučen na podatke [40]. Podatkovna ili aleatorna nesigurnost je nesigurnost koja potječe iz prirodne kompleksnosti samih podataka, kao što su preklapanje klasa, šum u klasama, homoskedastički i heterosedastički šum [40], te se ne može smanjiti učenjem modela [19]. Uz ova dva osnovna tipa nesigurnosti, čest je i pojam distribucijske nesigurnost koja nastaje zbog distribucijske (domenske) razlike između trening i test skupa podataka. Model nije upoznat s testnim podacima i zato ne može donositi sigurne predikcije [40].

Kod predikcije postoje dva ključna pristupa: induktivna i transduktivna predikcija, kao što je prikazano na slici 1.1 [63]. Induktivna predikcija sastoji se od dva koraka. Prvi korak nazivamo induktivni korak i kada iz poznatih primjera stvaramo općenito pravilo (eng. prediction or decision rule), model ili teoriju [63]. Drugi korak je dedukcijski, tada primjenjujemo dobiveno općenito pravilo kako bismo dobili predviđanje na novim primjerima [63]. S druge strane, kod transdukcijske predikcije idemo prečacem i dajemo predikciju za novi objekt direktno iz poznatih primjera [63].



Slika 1.1: Induktivna i transduktivna predikcija, slika preuzeta iz [63].

U ovom seminarskom radu dajemo pregled značajnijih metoda procjene prediktivne nesigurnost modela, sagledavamo primjenu na selektivnu predikciju i kalibraciju modela, te uspoređujemo metode za procjenu kvalitete kalibracije kod DNN modela. Uz to predstavljamo experimentalni postupak i prezentiramo rezultate odabranih metoda nad dubokim modelom EfficientNet-B0 arhitekture [58] prenaučenom na rješavanje klasifikacijskog i regresijskog zadatka pamtljivosti slike na LaMem skupu podataka

[34].

Ovaj seminarski rad podijeljen je na sljedeća poglavlja:

2. Konformalni prediktori

Konformalni prediktori (eng. Conformal Predictors, CP) su prediktivni modeli, klasifikacijski ili regresijski, koji uz predikcije daju i statistički valjanu mjeru pouzdanosti [39]. CPovi za svaki testni primjer daju višeznačnu predikciju (set ili interval vrijednosti) koja sadrži točnu predikciju s odabranom vjerojatnosti [39]. CPovi se mogu nadodati na bilo koji postojeći algoritam za klasifikaciju ili regresiju, a ključni uvjet za statističku valjanost konformalnih prediktora jest da su trening i test objekti zamjenjivi (eng. exchangeable) [39].

Zamjenjivost objekata zahtjeva da je svaka permutacija π nad rasporedom objekata u sekvenci $\{z_1, ..., z_n\}$ jednako vjerojatna, tj.:

$$P(z_1, ..., z_n) = P(z_{\pi(1)}, ..., z_{\pi(n)}) \tag{2.1}$$

Upravo zato se često koristi i izraz "torba uzoraka" (eng. bag of samples), ukazujući na to da raspored uzoraka nije bitan. Zamjenjivost objekata podrazumijeva da varijable imaju jednaku distribuciju, ali ne zahtjeva njihovu nezavisnost [68]. Zbog toga je zamjenjivost objekata slabiji uvjet od slučajnosti objekata (slučajno uzorkovanje, eng. random sampling) gdje je uvjet da su objekti uzorkovani nezavisno iz neke distribucije vjerojatnosti [68].

Ideja konformalnih prediktora je predvidjeti labelu/vrijednost testnom uzorku na osnovu prethodnog iskustva. Konformalni prediktori su bazirani na pretpostavci da je pouzdanost predikcije to veća što je testni uzorak sličniji poznatim uzorcima [39]. Testni uzorak (x_i, y_i) se sastoji od ulaznog objekta x_i i predikcije y_i. Različitost testnog uzorka od poznatih uzoraka može doći iz više smjerova: (1) razlike ulaznih objekta; (2) razlike predikcije; (3) kompleksnosti predikcije.

U prvom slučaju testni ulazni objekt je značajno različit od ulaznih objekata poznatih uzoraka, tj. postoji značajan pomak u domeni. Budući da temeljimo predikciju na osnovu poznatih primjera, veća je vjerojatnost da ćemo napraviti točnu predikciju ako su primjeri sličniji poznatim primjerima. Ova nesigurnost potječe od samih podataka, zbog pomaka u domeni nije moguće naučiti prediktivni model koji bi imao sigurnu

predikciju.

S druge strane, očekivano je da za sličan ulazni objekt imamo i sličnu predikciju. Nesigurnost u drugom slučaju potječe od samog modela. Ako se predikcija značajno razlikuje labela/vrijednosti sličnih ulaznih objekata poznatih primjera to znači da prediktivni model nije dobar.

U trećem slučaju, nesigurnost potječe iz nesigurnosti samih podataka. Sigurnost predikcije je velika ako bliski poznati podaci imaju istu labelu/vrijednost. S druge strane, nesigurnost će biti velika ako se testni ulazni objekt nalazi u dijelu ulaznog prostora gdje u blizini imamo veliku razliku labela/vrijednosti u poznatom skupu.

Iako se CPovi mogu koristiti i za modele klasifikacije i regresije, u nastavku ovog poglavlja glavni fokus će biti na primjeni CPova na klasifikacijskim problemima. Krenuti ćemo od opisa ključnih pojmova neophodnih za pojašnjenje konformalnih prediktora. Potom ćemo redom predstaviti: transduktivne, induktivne, kros, bootstrap i agregirane CPove, pri tom ćemo koristiti oznake, jednadžbe i definicije po uzoru na članak [39]. Na samom kraju poglavlja u kratko predstavljamo različite mjere nekonformnosti koje su upotrebljavane u CPovima.

2.1. Ključni pojmovi konformalnih prediktora

U ovom dijelu ćemo predstaviti pojmove ključne za opis konformalnih prediktora. Prije svega što je to konformnost? Konformnost predstavlja sličnost, slaganje ili vjerojatnost pojavljivanja određenog uzorka unutar nekog specifičnog prostora problema [39]. Suprotno od konformnosti je nekonformnost koja predstavlja razliku i neslaganje. Upotreba konformnosti i nekonformnost je jednako valjana ali se u kontekstu konformalnih prediktora češće upotrebljava nekonformnost kao mjera. U nastavku ovog dokumenta ćemo konformalne prediktore i njihove jednadžbe predstaviti iz smjera nekonformnost.

Razina značajnosti (eng. significance level) ili razina nepokrivenosti (eng. miscoverage level), $\epsilon \in [0, 1]$, je jedan od ključnih parametara konformalnih prediktora kojim definiramo potrebnu razinu pouzdanosti za predikciju [68]. ***Razina sigurnosti*** (eng. confidence level) pokazuje koliko smo sigurni u pojedinu predikciju, a ona iznosi $1 - \epsilon$.

Mjera nekonformnosti (eng. nonconformity measure) je funkcija $f : Z^* \times Z \to \mathbb{R}$, gdje je $Z : X \times Y$, koja daje vrijednost nekonformnosti $\alpha_i = f(\zeta, z_i)$ za svaki uzorak u skupu podataka $z_i = (x_i, y_i) \in Z$ [39].

Vrijednost nekonformnosti (eng. nonconformity score) α pokazuje koliko je pojedini uzorak u podacima nekonforman (stran ili naočekivan) naprema ostalim podacima u skupu $\zeta \in Z^*$ [39].

Glavni indikatori koliko se konformalni prediktor dobro ponaša su ***valjanost*** (tj. razina pouzdanosti) i ***efikasnost*** (tj. informativna efikasnost) [68]. ***Valjanost*** konformalnog prediktora ukazuje na to koliko je njegova procjena nesigurnosti točna. Postoje različiti tipovi valjanosti prediktora, a nama su značajni: (1) ***točna valjanost*** (eng. exactly valid); (2) ***konzervativna valjanost*** (eng. conservatively valid); i (3) ***podjednaka valjanost*** (eng. respectively valid).

Konformalni prediktor je ***točno valjan*** za razinu značajnosti $\epsilon \in [0, 1]$ ako je vjerojatnost pogreške predikcije jednaka ϵ za dani testni uzorak [68]. S druge strane, ako vjerojatnost pogreške ne prelazi ϵ, u određenim uvjetima, tada kažemo da je konformalni prediktor ***konzervativno valjan*** [68]. U slučaju da valjanost prediktora vrijedi za sve moguće razine značajnosti $\epsilon \in [0, 1]$, tada je prediktor ***podjednaka valjan***, npr. prediktor moze biti podjednako konzervativno valjan [68].

Uz valjanost, jako je važna i ***informativna efikasnost*** prediktora. U pravilu želimo da nam prediktor bude što boljih performansa, tj. da nam daje predikcije s visokom sigurnosti. Ključno je naglasiti da je valjanost prioritet, jer se bez valjanosti gubi smisao prediktivnog područja, te je lako moguće ostvariti maksimalnu efikasnost [68]. Upravo zbog toga nas zanimaju najefikasniji konformalni prediktori od onih koji spadaju u klasu valjanih prediktora, tj. dobro-kalibrirani konformalni prediktori [68].

Kod konformalnih prediktora postoje dva različita aspekta između kojih balansiramo: (1) pouzdanosti i informativne efikasnosti; (2) informativne i računalne efikasnosti [39].

Kompromis između pouzdanosti i informativne efikasnosti (eng. confidence for informational efficiency trade-off) potječe od odabira udjela poznatih podataka koje ćemo iskoristiti za učenje modela (informativna efikasnost) naprema udjelu koji preostaje za kalibraciju (pouzdanost). Ovaj kompromis je najizraženiji kod indukcijskih konformalnih prediktora.

Kompromis između informativne i računalne efikasnosti, u kontekstu konformalnih prediktora temelji se na odabiru metode konformalnog prediktora. Transdukcijski konformalni prediktori imaju slabu računalnu efikasnost, ali veliku informativnu efikasnost, jer cijeli poznati skup podataka možemo koristiti za učenje modela. S druge strane indukcijski konformalni prediktori imaju visoku računalnu efikasnost, ali zato imaju nižu informativnu efikasnost, jer samo dio poznatog skupa podataka možemo koristiti za učenje, a ostatak je rezerviran za kalibraciju.

2.2. Transduktivni konformalni prediktori

Transduktivni konformalni prediktori (eng. Transductive Conformal Predictors, TCP) [21] se temelje na transduktivnoj predikciji. Za dani skup primjera za učenje $Z^n = \{z_1, ..., z_n\}$, testni objekt x_{n+1} i probne testne vrijednosti $\tilde{y} \in Y$, konstruiramo prošireni skup $Z^{n+1} = Z^n \bigcup (x_{n+1}, \tilde{y})$. Tada izračunavamo vrijednost nekonformnosti za svaki trening uzorak $z_i \in Z^n$ sljedećom jednadžbom:

$$\alpha_i^{\tilde{y}} = f(Z^{n+1} \setminus z_i, z_i); \tag{2.2}$$

u kojoj znak $\setminus$ označava operaciju ***izuzimanja*** uzorka z_i iz skupa Z^{n+1}. Dok vrijednost nekonformnosti za testni uzorak s probnom testnom labelom izračunavamo s:

$$\alpha_{n+1}^{\tilde{y}} = f(Z^{n+1} \setminus z_{n+1}, z_{n+1}) = f(Z^n, (x_{n+1}, \tilde{y})); \tag{2.3}$$

budući da je $Z^{n+1} \setminus z_{n+1} = Z^n$. Tada definiramo CP kao set prediktor:

$$\Gamma_{n+1}^{\epsilon} = \left\{ \tilde{y} \in Y : p_{n+1}^{\tilde{y}} > \epsilon \right\}; \tag{2.4}$$

gdje je:

$$p_{n+1}^{\tilde{y}} = \frac{\left| \left\{ z_i \in Z^{n+1} : \alpha_i^{\tilde{y}} \geq \alpha_{n+1}^{\tilde{y}} \right\} \right|}{n+1}. \tag{2.5}$$

Korištenjem ove jednadžbe za izračun p vrijednosti TCPa osigurava se njegova ***konzervativna valjanost***.

Za osiguravanje ***točne valjanosti*** potrebno je uvesti novi parametar Θ_{n+1} čiju vrijednost slučajno odabiremo unutar intervala $[0, 1]$ s uniformnom vjerojatnosti $\Theta_{n+1} \sim U[0, 1]$. Takav CP nazivamo ***izglađeni konformalni prediktor*** (eng. smoothed conformal predictor, SCP), a jednadžba za izračun njegove p vrijednosti je:

$$p_{n+1}^{\tilde{y}} = \frac{\left| \left\{ z_i \in Z^{n+1} : \alpha_i^{\tilde{y}} > \alpha_{n+1}^{\tilde{y}} \right\} \right| + \Theta_{n+1} * \left| \left\{ z_i \in Z^{n+1} : \alpha_i^{\tilde{y}} = \alpha_{n+1}^{\tilde{y}} \right\} \right|}{n+1}. \tag{2.6}$$

Budući da nas u konačnici zanima točna valjanost, u ostalim slučajevima ćemo koristiti isključivo jednadžbe za SCP.

Glavni problem kod TCPova je slaba računalna efikasnost. Za svaki testni uzorak moramo osnovni model h nanovo učiti $(n+1) * |Y|$ puta, gdje je n broj poznatih uzoraka, $|Y|$ ukupan broj mogućih izlaznih labela [39].

Moguće je smanjiti računalnu kompleksnost upotrebom sljedeće jednadžbe za izračun vrijednosti nekonformnosti:

$$\alpha_i^{\tilde{y}} = f(Z^{n+1}, z_i); \tag{2.7}$$

gdje je $z_i \in Z^{n+1}$. U ovom slučaju je potrebno model ponovo učiti $|Y|$ puta za svaki testni objekt, što je značajno poboljšanje, ali je to i dalje vrlo slaba računalna efikasnost [39].

2.3. Induktivni konformalni prediktori

Induktivni konformalni prediktori (eng. Inductive Conformal Predictors, ICP) [45] su osmišljeni s ciljem poboljšanja računalne efikasnosti transduktivnih konformalnih prediktora. ICPovi zahtjevaju samo jedno učenje osnovnog modela h, tj. osnovni model nije potrebno ponovo učiti [39].

To je postignuto tako što je poznati skup podataka Z^n dijunktivno podijeljen na dva dijela: (1) ***pravi skup za učenje*** Z^t (eng. proper training set); i (2) ***kalibracijski skup*** Z^c (eng. calibration set) [39]. Smanjenje skupa podataka za učenje direktno utječe na prediktivnu moć modela, što je i razlog slabije informativne efikasnosti induktivnih konformalnih prediktora naprema transduktivnih. U kratko, bazni model h učimo samo iz pravog skupa za učenje, a mjeru nekonformnosti računamo isključivo iz kalibracijskog skupa i testnog uzorka:

$$\alpha_i = f(Z^t, z_i); \tag{2.8}$$

gdje je $z_i \in Z^c$, dok je mjera nekonformnosti za testni uzorak:

$$\alpha_{n+1}^{\tilde{y}} = f(Z^t, (x_{n+1}, \tilde{y})). \tag{2.9}$$

Sada p vrijednost za testni objekt možemo izračunati iz sljedeće jednadžbe:

$$p_{n+1}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z^c : \alpha_i > \alpha_{n+1}^{\tilde{y}}\right\}\right| + \Theta_{n+1} * \left(\left|\left\{z_i \in Z^c : \alpha_i = \alpha_{n+1}^{\tilde{y}}\right\}\right| + 1\right)}{c+1}; \tag{2.10}$$

gdje je c broj elemenata u kalibracijskom skupu Z^c. U brojniku, unutar zagrade koju množi Θ_{n+1} nadodajemo 1. Razlog tome je što za testni uzorak, $n+1$, uvijek vrijedi jednakost $\alpha_{n+1}^{\tilde{y}} = \alpha_{n+1}^{\tilde{y}}$. Tako smo ostvarili da skup podataka kojim ispitujemo nekonformnost ne uključuje samo kalibracijski skup, već se uz njega nadodaje i testni uzorak, slično kao što je bilo i kod TCPova.

2.4. Unakrsno-konformalni prediktori

Unakrsno-konformalni prediktori (eng. Cross-Conformal Predictors, CCP) [61] su razvijeni s ciljem poboljšavanja informativne efikasnosti ICPova, s težnjom zadržavanja

visoke računalne efikasnosti [39]. Kod CCPova dijelimo skup podataka za učenje Z^n na k ne-praznih disjunktivnih podskupova podataka, $Z_1, ..., Z_k$ kalibracijskih skupova podataka, a svaki od baznih prediktivnih modela h_l učimo na odabranom pravom skupu podataka za učenje $Z_{-l} = \bigcup_{r=1,...,k} Z_r \setminus Z_l$, vrlo slično kao što je preklapanje (eng. folding) skupa podataka u kros-validacijskoj (eng. cross-validation) metodi [39].

U ovom slučaju vrijednost nekonformnosti za pojedini testni uzorak računamo iz odabranog kalibracijskog skupa za svaki od preklopa $l = 1, ..., k$ po jednadžbi:

$$\alpha_{i,l} = f(Z_{-l}, z_i); \tag{2.11}$$

gdje je $z_i \in Z_l$, dok je mjera nekonformnosti za testni uzorak:

$$\alpha^{\tilde{y}}_{n+1,l} = f(Z_{-l}, (x_{n+1}, \tilde{y})). \tag{2.12}$$

Sada ukupnu p vrijednost za testni objekt možemo izračunati iz sljedeće jednadžbe:

$$p^{\tilde{y}}_{n+1} = \frac{\sum_{l=1}^{k} \left[\left| \left\{ z_i \in Z_l : \alpha_{i,l} > \alpha^{\tilde{y}}_{n+1,l} \right\} \right| + \Theta_{n+1,l} * \left(\left| \left\{ z_i \in Z_l : \alpha_{i,l} = \alpha^{\tilde{y}}_{n+1,l} \right\} \right| \right) \right] + \Theta_{n+1}}{n+1}. \tag{2.13}$$

2.5. Bootstrap konformalni prediktori

Bootstrap konformalni prediktori (eng. Bootstrap Conformal Predictors, BCP) su CPovi bazirani na ansamblu modela, kao i CCPovi kojima su jako slični. Jedina značajna razlika je u tome da su podskupovi podataka $Z_{-1}, ..., Z_{-k}$ iz Z^n ovog puta odabrani izvlačenjem uzoraka ***sa zamjenama*** (eng. with replacement), poznatog pod nazivom bootstrap uzorkovanje. U ovom slučaju za svaki temeljni model h_l možemo izračunati vrijednost nekonformnosti na jednak način kao i kod CCPova:

$$\alpha_{i,l} = f(Z_{-l}, z_i); \tag{2.14}$$

gdje je $z_i \in Z_l$ i $Z_l = Z^n \setminus Z_{-l}$, dok je mjera nekonformnosti za testni uzorak:

$$\alpha^{\tilde{y}}_{n+1,l} = f(Z_{-l}, (x_{n+1}, \tilde{y})). \tag{2.15}$$

Tada p vrijednosti izračunavamo:

$$p^{\tilde{y}}_{n+1} = \frac{\sum_{l=1}^{k} \left[\left| \left\{ z_i \in Z_l : \alpha_{i,l} > \alpha^{\tilde{y}}_{n+1,l} \right\} \right| + \Theta_{n+1,l} * \left(\left| \left\{ z_i \in Z_l : \alpha_{i,l} = \alpha^{\tilde{y}}_{n+1,l} \right\} \right| \right) \right] + \frac{t}{n}\Theta_{n+1}}{t + \frac{t}{n}}; \tag{2.16}$$

gdje je t ukupni broj uzoraka u kalibracijskom setu, $t = \sum_{l=1}^{k} |Z_l|$.

2.6. Agregirani konformalni prediktori

Agregirani konformalni prediktori (eng. Aggregated Conformal Predictors, ACP) razvijeni su kao poopćenje CCPa i BCPa [8]. Kod njih se od trening skupa podataka bira k podskupova pravih skupova za učenje $Z_{-1}, ..., Z_{-k}$ i njihovih pripadajućih kalibracijskih skupova $Z_1, ..., Z_k$, na sličan način kao što je i kod CCPa i BCPa. Ukupnu p vrijednost AKPa izračunavamo sljedećom jednadžbom:

$$p_{n+1}^{\tilde{y}} = \frac{1}{k} \sum_{l=1}^{k} p_{n+1,l}^{\tilde{y}}, \tag{2.17}$$

a svaku pojedinačnu p vrijednosti za Z_l kalibracijski skup i Z_{-l} pravi skup za učenje, dobivamo jednadžbom:

$$p_{n+1,l}^{\tilde{y}} = \frac{\left|\left\{z_i \in Z_l : \alpha_{i,l} > \alpha_{n+1,l}^{\tilde{y}}\right\}\right| + \Theta_{n+1,l} * \left(\left|\left\{z_i \in Z_l : \alpha_{i,l} = \alpha_{n+1,l}^{\tilde{y}}\right\}\right| + 1\right)}{|Z_l| + 1}. \tag{2.18}$$

Dobro je naglasiti da CCP, BCP i ACP ne osiguravaju dobru kalibraciju kao što je to slučaj kod TCPa i ICPa, već je kod njih kvaliteta kalibracije jako ovisna o ***stabilnosti*** osnovnog modela. U pravilu, što je osnovni model nestabilniji to je veća šansa da je CCP, BCP i ACP slabo kalibriran, kao što se može vidjeti iz experimentalne analize provedene u članku [39].

2.7. Mjere nekonformnosti

Mjere nekonformnosti ne utječu na valjanost konformalnih prediktora, ali imaju veliki utjecaj na njihovu informativnu efikasnost [39]. Standardne metode za izračun nekonformnosti (tj. funkcije nekonformnosti) temelje se na tradicionalnim modelima strojnog učenja:

$$f(\zeta, (x_i, y_i)) = \Delta(h(x_i), y_i); \tag{2.19}$$

gdje je h osnovni prediktivni model konformalnog prediktora treniran na setu podataka ζ, a Δ je funkcija pogreške prediktivnog modela h [39].

U cilju poboljšanja informativne efikasnosti CPova korišteni su različiti algoritmi strojnog učenja za mjeru nekonformnosti [39], kao što su sljedeće metode: najmanji kvadrati (eng. least squares), regresija grebenom (eng. ridge regression), logistička regresija (eng. logistic regression), k-najbližih susjeda (eng. k-nearest neighbors), SVM (eng. support vector machines), stabla odluke (eng. decision trees), šume odluke (eng. decision trees), boosting, bootstrap, neuronske mreže (eng. neural networks).

Primjeri primjene svih ovih metoda u svrhu mjerenja nekonformnosti dani su i opisani u knjizi [63].

2.8. Test zamjenjivosti

Test zamjenjivosti (eng. exchangeability testing) je pristup kojim provjeravamo izglednost pretpostavke da su podaci zamjenjivi, tj. da je svaka permutacija njihova rasporeda jednako vjerojatna. Zamjenjivost podataka standardna je pretpostavka u strojnom učenju, te je neophodan uvjet i za CPove. U nastavku ćemo predstaviti metode za provjeru zamjenjivosti za on-line slučaj. U tom slučaju primjeri stižu jedan po jedan, a nakon svakog primljenog primjera želimo dati valjanu mjeru stupnja do kojeg je pretpostavka o zamjenjivosti lažirana [13]. U skladu s de Finettijevim teoremom, svaki test zamjenjivosti ekvivalentan je provjeri jesu li uzorci neovisni s jednakom distribucijom [13].

Prema [60], tradicionalni statistički pristupi za testiranje nisu pogodni za visokodimenzionalne podatke [13]. Rješenje za ovaj problem predloženo je u članku [62], a temelji se na teoriji konformalnih prediktora i izračuna ***martingala zamjenjivosti*** (eng. exchangeability martingales). Test se provodi u dva koraka. Prvi korak je izvedba CPa koji daje sekvencu p-vrijednosti, gdje se svaka p-vrijednost izračunava iz trenutnog primjera i svih koji mu prethode [13]. U drugom koraku se koriste martingali zamjenjivosti, funkcije koje primaju p-vrijednosti i prate devijaciju od pretpostavke [13]. U slučaju da vrijednosti martingala postanu velike tada možemo odbaciti pretpostavku zamjenjivosti podataka [13].

U prvom koraku, kod CPa, potrebno je izračunati vrijednost nekonformnosti za $i \in \{1, ..., n+1\}$:

$$\alpha_i = A\big(z_i, \{z_1, ..., z_n, z_{n+1}\}\big); \tag{2.20}$$

gdje je A funkcija nekonformnosti, te u slučaju 1-najbližih susjeda (1-NN) jednadžba postaje:

$$\alpha_i = \frac{\min_{j \neq i : y_i = y_j} d(x_i, x_j)}{\min_{j \neq i : y_i \neq y_j} d(x_i, x_j)}; \tag{2.21}$$

gdje je $d(x_i, x_j)$ funkcija Euklidske udaljenosti [13]. Odabrana mjera nekonformnosti za α_i daje visoke vrijednosti ako je primjer i blizak drugom primjeru s različitom labelom, a udaljen od svih primjera s istom labelom. Sada, izglađenim CPom, iz dobivenih vrijednosti nekonformnosti možemo dobiti p-vrijednost novog uzorka z_{n+1}

(kao u jednadžbi 2.6):

$$p_{n+1} = \frac{|\{i : \alpha_i > \alpha_{n+1}\}| + \theta_{n+1}|\{i : \alpha_i = \alpha_{n+1}\}|}{n+1} \tag{2.22}$$

Martingali zamjenjivosti su sekvence nenegativnih slučajnih varijabli $S_0, ..., S_n$ za koje vrijedi uvjetna vjerojatnost:

$$S_n = E(S_{n+1}|S_0, ..., S_n) \geq 0; \tag{2.23}$$

gdje E predstavlja očekivanu vrijednost s obzirom na bilo koju distribuciju izmjenjenih primjera, pri čemu je $S_0 = 1$ [13]. U teoriji vjerojatnosti, ***martingale*** nazivamo sekvencama slučajnih varijabli za koje je uvjetno očekivanje sljedeće vrijednosti u sekvenci jednako sadašnjoj vrijednosti, neovisno o prethodnim vrijednostima. U ovom specijalnom slučaju, vrijednosti martingala zamjenjivosti reflektiraju snagu dokaza protiv pretpostavke zamjenjivosti podataka [13].

Za svaki $i \in \{1, ..., n+1\}$, neka je $f_i : [0,1]^i \to [o, \infty)$, a $(p_1, p_2, ..., p_{n+1})$ sekvenca p-vrijednosti dobivenih u prvom koraku, jednadžbom 2.22, tada vrijednost martingala S_j za svaki uzorak $j \in \{1, ..., n+1\}$ možemo dobiti iz:

$$S_j = \prod_{i=1}^{j} f_i(p_i) \tag{2.24}$$

gdje funkciju $f_i(p_i)$ nazivamo ***funkcijom klađenja***, a izračunava se iz p-vrijednosti trenutnog i svih prethodnih uzoraka: $f_i(p_i) = f_i'(p_1, ..., p_{i-1}, p_i)$.

Da bi osigurali da je S_j (jednadžba 2.24) martingale trebamo uvesti sljedeće ograničenje na funkciju klađenja f_i:

$$\int_0^1 f_i(p)dp \tag{2.25}$$

za svaki $i \in \{1, ..., n+1\}$. Tada možemo provjeriti sljedeći izraz:

$$\begin{aligned} E(S_{n+1}|S_0, ..., Sn) &= \int_0^1 \prod_{i=1}^{n} (f_i(p_i)) f_{n+1}(p)dp \\ &= \prod_{i=1}^{n} (f_i(p_i)) \int_0^1 f_{n+1}(p)dp = \prod_{i=1}^{n} (f_i(p_i)) = S_n. \end{aligned} \tag{2.26}$$

Koristeći jednadžbu 2.24 možemo izračunati novu vrijednost maringala iz stare vrijednosti i p-vrijednosti novog primjera $S_i = S_{i-1} * f_i(p_i)$ [13]. Da bi potpuno opisali martingale potrebno je definirati funkciju klađenja.

Funkcija klađenja (eng. betting function) određuje koliko pojedina p-vrijednost doprinosi martingalu [13]. Funkcija klađenja može biti fiksna (ostaje ista, bez obzira na prethodne uzorke) ili promjenjiva (funkcija se mijenja, ovisno o prethodnim uzorcima).

U kontekstu testa zamjenjivosti, u članku [62] predložena je fiksna funkcija klađenja s formom:

$$f_i = \mathcal{E} p^{\mathcal{E}-1}; \tag{2.27}$$

za svaki $i \in \{1, .., n+1\}$, gdje je $\mathcal{E} \in [0, 1]$. Ova funkcija klađenja korištena je kod različitih martingala, kao što su ***martingal snage*** (eng. power martingale):

$$M_j^{\mathcal{E}} = \prod_{i=1}^{j} \mathcal{E} p_i^{\mathcal{E}-1}; \tag{2.28}$$

i ***martingal jednostavnog miješanja*** (eng. simple mixture martingale):

$$M_j = \int_0^1 M_j^{\mathcal{E}} d\mathcal{E}; \tag{2.29}$$

oba predložena u članku [62], a gdje je $j \in \{1, .., n+1\}$. Ovi martingali će rasti samo onda ako je velik broj uzoraka s malim p-vrijednostima u sekvenci [13].

U članku [13] predložena je nova metoda testa zamjenjivosti nazvana ***martingali uključivanja*** (eng. plug-in martingales), a temelji se na promjenjivoj funkciji klađenja $f_i(p_i)$. Za funkciju klađenja koristi se funkcija gustoće vjerojatnosti estimirana iz p-vrijednosti svih prethodnih uzoraka $p_1, ..., p_{i-1}$:

$$f_i(p_i) = \rho_i(p_i) = \hat{\rho}(p_1, ..., p_{i-1}, p_i). \tag{2.30}$$

Martingali uključivanja izbjegava klađenje ako su p-vrijednosti uniformno distribuirane, ali ako ako postoji neki šiljak u distribuciji, on će biti upotrijebljen za klađenje [13].

3. Dubokih k-najbližih susjeda (DkNN)

Dubokih k-najbližih susjeda (eng. Deep k-Nearest Neighbors, DkNN) je prediktivna metoda predložena u članku [46]. Cilj metode je iskoristiti strukturu duboke neuronske mreže, koja omogućava rješavanje zadataka visoke razine apstrakcije, i pritom poboljšati njenu sigurnost. DkNN metoda posebno adresira sljedeće tri komponente sigurnosti: procjenu prediktivne sigurnosti, interpretabilnost modela i robusnosti na napade [46].

DkNN je hibridni klasifikator koji kombinira metodu ***k-najbližih susjeda*** (eng. k-nearest neighbors, k-NN) s reprezentacijama ulaznih podataka dobivenih na svakom od slojeve naučenog DNN modela. Za svaki sloj u DNNu, DkNN izvodi pretragu najbližih susjeda da pronađe uzorke skupa za učenje za koje je izlaz sloja najbliži izlazu istog tog sloja za odabrani testni uzorak. Dakle, uspoređujemo reprezentacije uzoraka i pronalazimo najbliže reprezentacije. Da bi to ostvarili, potrebno je pamtiti uzorke skupa za učenje ili njihove reprezentacije sa svakog sloja DNN mreže.

U sljedećem koraku potrebno je analizirati labele susjednih uzoraka skupa za učenje te provjeriti jesu li međurezultati svakog sloja ostali konformalni s krajnjom predikcijom modela. U članku su pokazali da labele susjeda omogućuju estimaciju prediktivne sigurnosti za ulazne objekte koji se nalaze izvan naučenog manifolda modela, tj. primjere izvan poznate domene, a to uključuje i napadačke primjere [46]. Najbliži susjedi nam omogućuju procjenu nekonformnosti, tj. koliku podršku za predikciju testnog primjera imamo u podacima za učenje.

Estimacija nekonformnosti kod DkNN temelji se na ICPu (potpoglavlje 2.3), te je jednadžba za mjeru nekonformnosti DkNNa:

$$\alpha(x_i, y_i) = \sum_{l=1}^{L} |y_j \in \Omega_l : y_j \neq y_i| \tag{3.1}$$

gdje je x_i testni ulazni objekt, y_i njegova predikcija, $l \in \{1, ..., L\}$ je sloj DNN modela,

a Ω_l je skup labela od skupa podataka za učenje koji su najbliži testnom uzorku x_i za sloj l.

Za valjani izračun ICPa potrebno je osigurati kalibracijski skup podataka (X^c, Y^c) kojeg u članku [46] izdvajaju iz testnog skupa podataka. Potom računaju vrijednosti nekonformnosti za sve uzorke kalibracijskog skupa: $A = \{\alpha(x, y) : (x, y) \in (X^c, Y^c)\}$. Sada p-vrijednosti kandidata za test labelu $\tilde{y} \in Y$ možemo dobiti iz jednadžbe:

$$p_{\tilde{y}_i}(x_i) = \frac{|a \in A : a > \alpha(x_i, \tilde{y}_i)|}{|A|} \tag{3.2}$$

.

Uz procjenu prediktivne nesigurnosti, susjedi također omogućuju i objašnjivost predikcija modela na čovjeku shvatljiv način, a posebno su korisni u objašnjavanju pogreške predikcije [46]. Uz sve prednosti DkNN metode, pozitivno je i to što se može primijeniti na već naučenom DNN modelu, kao post-hoc metoda. Za to je jedino potrebno imati pristup skupu za učenje ili reprezentacijama, kao i kalibracijski skup odvojen od ostalih skupova.

4. Jackknife

Začetak jackknife metode pronalazimo u članku [50] gdje je predstavljena tehnika za smanjenje pristranosti kod procjenitelja serijske korelacije (eng. serial correlation estimator), a temelji se na tome da se skup dijeli na dva podskupa [41]. Generalizacija te metode objavljena je u članku [51] gdje se skup dijeli na g grupa veličine h, tako da je ukupan broj uzoraka $n = gh$ [41].

Ako uzmemo skup neovisnih slučajnih varijabli s jednakom distribucijom $Z_1, ..., Z_n$, te ako je $\hat{\theta}$ procjenitelj parametra θ temeljen na uzorku veličine n. Ako je $\hat{\theta}_{-i}$ odgovarajući procjenitelj temeljen na skupu veličine $(g-1)h$, gdje je i-ta grupa veličine h izuzeta iz ukupnog skupa. Tada definiramo pojedinačni procjenitelj:

$$\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i} = \hat{\theta} + (g-1)(\hat{\theta} - \hat{\theta}_{-i}); \tag{4.1}$$

i ukupni procjenitelj:

$$\tilde{\theta} = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_i = g\hat{\theta} - (g-1)\frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}; \tag{4.2}$$

koji posjeduje karakteristiku odstranjivanja reda $1/n$ iz sljedeće forme:

$$E(\hat{\theta}) = \theta + a_1/n + O(1/n^2). \tag{4.3}$$

U [59] je predloženo da se u velikom broju slučajeva prema $\tilde{\theta}_i$ vrijednostima (4.1) može pristupiti kao približno neovisnim slučajnim varijablama jednakih distribucija. U tom slučaju bi sljedeća statistika:

$$t = \frac{\tilde{\theta} - \theta}{\sigma/\sqrt{g}} = \frac{\sqrt{g}(\tilde{\theta} - \theta)}{\left\{\frac{1}{g-1}\sum_{i=1}^{g}(\tilde{\theta}_i - \tilde{\theta})^2\right\}^{\frac{1}{2}}} \tag{4.4}$$

trebala imati približno t distribuciju s $g-1$ stupnjeva slobode i predstavljati ključnu statistiku za robusnu procjenu intervala [41]. Prema [41], Tukey je u svom sljedećem neobjavljenom radu $\tilde{\theta}_i$ vrijednosti iz 4.1 nazvao ***pseudo-vrijednostima*** (eng. pseudo-values) i dao naziv ***jackknife procjenitelj*** (eng. jackknife estimator) za procjenitelja iz jednadžbe 4.2.

Uz smanjenje pristranost (eng. bias reduction) jackknife metoda se primjenjuje i za procjenu intervala vrijednosti (eng. interval estimation), tj. intervala povjerenja (eng. confidence intervals). U sklopu ovog dokumenta nam je ova druga primjena značajnija, te ćemo se u nastavku ovog poglavlja pretežito usmjeriti na nju.

4.1. Jackknife izbaci jednog

Iako generalni pristup jackknife metode s različitim veličinama g i h ima svoj značaj, najviše je istraživan specijalni slučaj kada je $g = n$ i $h = 1$ [41]. Taj slučaj ima prednost nad drugima jer se kod njega uklanja proizvoljnost u formiranju grupa [41]. Ovaj ***"izbaci jednog"*** (eng. leave-one-out) pristup se u prošlosti često upotrebljavao pod pojmom ***unakrsne validacije*** (eng. cross-validation), ali se u modernoj literaturi za njega uvriježio termin ***"jackknife"*** [3]. Ovo miješanje pojmova nije slučajno jer su unakrsna validacija i jackknife po mnogo čemu slični. Obje metode se temelje na izuzimanju jedne ili više uzoraka iz skupa podataka za učenje, a ključna razlika je što jackknifea metoda koristi pseudo-vrijednosti. [56].

U ovom specijalnom slučaju imamo n pseudo-vrijednosti dobivenih jednadžbom 4.1, gdje je procjenitelj $\hat{\theta}_{-i}$ temeljen na setu s izbačenim i-tim uzorkom. Tada estimacija pseudo-vrijednosti odgovara jackknife procjenitelju, kao u 4.2:

$$E(\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i; \tag{4.5}$$

dok je varijanca:

$$Var(\tilde{\theta}) = \sigma(\tilde{\theta})^2 = n(\hat{SE}(\tilde{\theta}))^2 = \frac{1}{n+1}\sum_{i=1}^{n}(\tilde{\theta} - \tilde{\theta}_i)^2; \tag{4.6}$$

U ovom slučaju varijancu računamo dijeljenjem s faktorom $(n-1)$, a ne (n). Razlog tome je što se ovdje radi o uzorku iz populacije, a ne o cijeloj populaciji. Sada možemo procijeniti interval vrijednosti s razinom značaja ϵ prema:

$$\hat{C}^{jacknife}_{n,\epsilon} = \tilde{\theta} \pm t_{1-\epsilon/2,n-1}\hat{SE}(\tilde{\theta}) = \tilde{\theta} \pm t_{1-\epsilon/2,n-1}\sqrt{\frac{1}{n}Var(\tilde{\theta})} \tag{4.7}$$

gdje je $\hat{SE}$ procjena standardne pogreške, a $t_{1-\epsilon/2,n-1}$ kvantil Studentove t distribucije s $n-1$ stupnjeva slobode i razine značaja ϵ.

Procjenu intervala vrijednosti primjenom jackknife metode možemo dobiti pristupom predstavljenim u članku [3]. Ovdje sagledavamo specifičan slučaj regresijskog problema s poznatim podacima za učenje $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, ..., n$,

koji su slučajno izvučeni iz iste distribucije vjerojatnosti. Cilj je ugoditi funkciju $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}$ tako da predviđa Y_{n+1} vrijednost iz testnog ulaznog objekta X_{n+1}, te dati predikcijski interval oko $\hat{\mu}(X_{n+1})$ unutar kojeg se nalazi prava vrijednost Y_{n+1} s vjerojatnosti većom od $1 - \epsilon$ [3]. Pri tom $\hat{\mu}$ predstavlja funkciju ugođenu na cijelom skupu$\{(X_1, Y_1), ..., (X_n, Y_n)\}$, dok je $\hat{\mu}_{-i}$ ugođena na istom skupu iz kojeg je izbačen i-ti element.

Za svaku vrijednost v_i s indeksima $i = 1, ..., n$ i razine značaja ϵ, definiramo $\hat{q}^{+}_{n,\epsilon}\{v_i\}$ kojoj dodjeljujemo vrijednost $\lceil (1-\epsilon)(n+1) \rceil$-tog po redu najmanjeg elementa iz skupa vrijednosti $v_1, ..., v_n$. S druge strane definiramo $\hat{q}^{-}_{n,\epsilon}\{v_i\}$ kojoj dodjeljujemo vrijednost $\lfloor (\epsilon)(n+1) \rfloor$-tog po redu najmanjeg elementa iz skupa vrijednosti $v_1, ..., v_n$, te vrijedi:

$$\hat{q}^{-}_{n,\epsilon}\{v_i\} = -\hat{q}^{+}_{n,\epsilon}\{-v_i\}; \tag{4.8}$$

Tada definiramo jackknife predikcijski interval:

$$\hat{C}^{jacknife}_{n,\epsilon}(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm \hat{q}^{+}_{n,\epsilon}\{R^{LOO}_i\}; \tag{4.9}$$

$$\hat{C}^{jacknife}_{n,\epsilon}(X_{n+1}) = \left[\hat{q}^{-}_{n,\epsilon}\{\hat{\mu}(X_{n+1}) - R^{LOO}_i\}, \hat{q}^{+}_{n,\epsilon}\{\hat{\mu}(X_{n+1}) + R^{LOO}_i\}\right]; \tag{4.10}$$

gdje je $R^{LOO}_i = |Y_i - \hat{\mu}_{-i}(X_{n+1})|$, koji opisuje rezidual s izbačenim i-tim elementom skupa podataka [3].

4.2. Jackknife+

Jackknife daje valjanu prediktivnu pokrivenost pod pretpostavkama algoritamske stabilnosti [55], tj. ako osnovni model naučen na svim podacima $\hat{\theta}$ ima sličnu predikciju na testnom uzorku kao i model naučen na podacima s izuzetim jednim uzorkom $\hat{\theta}_{-i}$ [3]. S druge strane, u slučaju nestabilnosti osnovnog modela jackknife metoda može izgubiti svoju prediktivnu pokrivenost [3]. U članku [3] autori predlažu poboljšanje jackknife metode pod nazivom ***"jackknife+"***, u kojoj je predikcijski interval definiran jednadžbom:

$$\hat{C}^{jacknife}_{n,\epsilon}(X_{n+1}) = \left[\hat{q}^{-}_{n,\epsilon}\{\hat{\mu}_{-i}(X_{n+1}) - R^{LOO}_i\}, \hat{q}^{+}_{n,\epsilon}\{\hat{\mu}_{-i}(X_{n+1}) + R^{LOO}_i\}\right]; \tag{4.11}$$

Razlika je u tome što jackknife daje intervale centrirane oko predikcije testnog objekta koristeći $\hat{\mu}$ (4.10), dok jackknife+ za predikcije koristi $\hat{\mu}_{-i}$ (4.11), a time uzima u obzir varijabilnost regresijske funkcije. U većini slučajeva jackknife+ metoda funkcionira kao i jackknife, dok prednost jackknife+ dolazi do izražaja u slučajevima kada je predikcijski algoritam jako nestablian [55].

5. Bayesovski modeli

Bayesovski modeli (eng. Bayesian models, BM), su prediktivni modeli koji se temelje na Bayesovskom (probabilističkom) pristupu [4]. Za razliku od većine pristupa u strojnom učenju temeljenih na frekvencionističkom pristupu ('točkastom' predikcijom), Bayesianski modeli se temelje na probabilističkom pristupu (predikcijom 'distribucije'). Glavna razlika između ova dva pristupa je u tome što je Bayesovski pristup temeljen na marginalizaciji a ne na optimizaciji [64].

5.1. Osnove teorije vjerojatnosti

Slučajne varijable su varijable čije vrijednosti ovise o rezultatu nekog slučajnog fenomena, a mogu biti diskretne i kontinuirane. ***Diskretne slučajne varijable*** mogu poprimiti ograničen set vrijednosti $X \in x_1, ..., x_n$, a u slučaju da je broj vrijednosti jednak ***dva*** tada ih zovemo ***binarne slučajne varijable***. Diskretne slučajne varijable opisujemo na način da svakoj mogućoj vrijednosti pridružujemo vjerojatnost njenog pojavljivanja $P(x_i)$, za koje vrijedi: $P(x_i) \geq 0$ i $\sum_{i=1}^{n} P(x_i) = 1$.

S druge strane, ***kontinuirane*** slučajne varijable mogu poprimiti beskonačan broj vrijednosti unutar nekog intervala $X \in [x_{min}, x_{max}]$ te ih opisujemo funkcijom gustoće vjerojatnosti $P(x)$, za koju vrijedi $P(x) \geq 0$ i $\int_{x_{min}}^{x_{max}} P(x)dx = 1$. Površina ispod krivulje reprezentira vjerojatnost pojavljivanje vrijednosti unutar nekog intervala $[x_l, x_h]$, te se računa integralom: $P(x_l \geq x \geq x_h) = \int_{x_l}^{x_h} P(x)dx$.

Uvjetna vjerojatnost je vjerojatnost pojavljivanja neke vrijednosti na jednoj slučajnoj varijabli $X = x$, uz uvjet da je na drugoj slučajnoj varijabli bila točno određena vrijednost $Y = y$.

$$P(x|y) = P(X = x|Y = y). \tag{5.1}$$

Skupna vjerojatnost je vjerojatnost istovremenog pojavljivanja određenih vrijednosti na dvije (ili više) slučajnih varijabli, a jednadžba za dvije varijable je:

$$P(x, y) = P(X = x \wedge Y = y). \tag{5.2}$$

Uvjetnu i skupnu vjerojatnost povezujemo s ***produktnim pravilom***:

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X); \tag{5.3}$$

iz kojeg proizlazi ***Bayesov teorem***:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \tag{5.4}$$

Uz ova pravila jako je važna i ***marginalizacija***, koje za diskretne slučajne varijable ima jednadžbu:

$$P(Y) = \sum_{i=1}^{n} P(Y|x_i); \tag{5.5}$$

dok je za kontinuirane sljedeća jednadžba:

$$P(Y) = \int_{x_{min}}^{x_{max}} P(Y|x)dx. \tag{5.6}$$

5.2. Bayesovske neuronske mreže

Bayesovske neuronske mreže (eng. Bayesian Neural Networks, BNN) su neuronske mreže s Bayesovskim zaključivanjem. Kod BNN modela, parametrima i izlazima pristupamo kao slučajnim varijablama te pronalazimo marginalnu distribuciju koja najbolje odgovara podacima. Na ovaj način omogućuje se procjena prediktivne nesigurnosti modela. Problem u korištenju BNNa je u tome što zahtjevaju značajne promjene u proceduri učenja i računarski su skuplje od standardnih (ne-Bayesovskih neuronskih mreža). Budući da direktno Bayesovsko zaključivnje nije izvedivo za NN modele, razvijene su različite metode njegove aproksimacije [37], kao što su Laplaceova aproksimacija, metoda Monte Carlo Markovljevog lanca (eng. Markov chain Monte Carlo, MCMC), varijacijske Bayesovske metode, pretpostavljeno filtriranje gustoće (eng. assumed density filtering), propagacija očekivanja, kao i MCMC varijante slučajnog gradijenta, npr. Langevinove metode difuzije i Hamiltoninske metode. Kvaliteta prediktivne nesigurnosti BNN modela najviše ovisi o: (1) stupnju aproksimacije, s obzirom na računalne resurse; (2) odabiru apriorne distribucije [37].

6. Monte Carlo dropout

Dropout je regularizacijska tehnika koja ograničava prenaučenost modela tako što nasumično odbacuje pojedine jedinice DNN modela za vrijeme treninga [54]. Autori iz [20] predlažu upotrebu Monte-Carlo (MC) dropout tehnike za vrijeme testiranja modela kako bi se mogla procijeniti prediktivna nesigurnost DNN modela. Također su pokazali da se upotrebom MC dropout tehnike DNN mreža ponaša kako Bayesianski aproksimator. Na osnovu MC dropout tehnike, autori iz [40] su predložili mjere za procjenu tri različita tipa prediktivne nesigurnosti:

1. podatkovna ili aleatorna nesigurnost;
2. modelska ili epistemička nesigurnost;
3. distribucijska nesigurnost;

U tom slučaju aleatorna i epistemička nesigurnost mogu se izračunati iz sljedećih jednadžbi:

$$U_{model} = \mathcal{I}[y, \Theta | x^*, \mathcal{D}]; \tag{6.1}$$

$$U_{ukupno} = \mathcal{H}[\mathbb{E}_{p(\Theta|\mathcal{D})}[P(y|x^*, \Theta)]]; \tag{6.2}$$

$$U_{podatak} = \mathbb{E}_{p(\Theta|\mathcal{D})}[\mathcal{H}[P(y|x^*, \Theta)]]; \tag{6.3}$$

$$U_{model} = U_{ukupno} - U_{podatak}; \tag{6.4}$$

a distribucijsku nesigurnost možemo izračunati iz:

$$U_{distribucija} = \mathcal{I}[y, \mu | x^*, \mathcal{D}]; \tag{6.5}$$

$$U_{ukupno} = \mathcal{H}[\mathbb{E}_{p(\mu|x^*;\mathcal{D})}[P(y|\mu)]]; \tag{6.6}$$

$$U_{podatak} = \mathbb{E}_{p(\mu|x^*;\mathcal{D})}[\mathcal{H}[P(y|\mu)]]; \tag{6.7}$$

$$U_{distribucija} = U_{ukupno} - U_{podatak}; \tag{6.8}$$

gdje je $\mathcal{D}$ skup podataka koji se sastoji od N uzoraka $\mathcal{D} = x_n, y_n{}_{n=1}^{N}$, u kojem je $x \in \mathbb{R}^D$ ulazni podatak u D-dimenzionalnom prostoru, dok je izlazni podatak $y \in 1, ..., K$ u slučaju klasifikacijskog problema ili $y \in \mathbb{R}$ u slučaju regresijskog problema.

7. Ansambli modela

Osnovna ideja ansambla modela je kombiniranje predikcije više modela s ciljem poboljšanja prediktivne točnosti u odnosu na pojedinačne modele. Na takav način od prediktivno "slabih model" (eng. "weak learners") u konačnici dobivamo "jaki model" (eng. "strong learners"). Slabe modele još nazivamo i baznim modelima (eng. base learners), jer ih generiramo baznim algoritmima za strojno učenje [69]. Za uspješnost ansambla je važno da bazni modeli budu međusobno različiti i komplementarni.

7.1. Metode ansambla

Postoje različiti pristupi učenju i povezivanju "slabih" modela u ansambl, kao što su: ***boosting*** [52], ***bagging*** [5] i ***stacking*** [65]. Ove pristupe možemo podijeliti na slijedne i paralelene. Kod ***slijednih*** metoda učenje i izvođenje svakog baznog modela ovisi o modelu koji ga prethodi (npr. boosting). Suprotno tome kod ***paralelnih*** pristupa osnovni modeli se uče i izvode neovisno, što omogućuje paralelno izvođenje. Njihovu raznolikost možemo postići na različite načine, kao što su: učenje na različitim podskupovima podataka (npr. bagging [5]) ili upotrebom različitih algoritama za učenje (npr. stacking [65]).

Osnovna ideja ***boosting*** [52] pristupa je slijedno povezivanje baznih modela, tako da izlaz iz jednog modela ulazi u sljedeći model. Na takav način, slijednim prolaženjem kroz bazne modele, predikcija postaje sve točnija, tj. možemo reći da se finalni model poboljšava (eng. boosting). Popularniji predstavnici boosting pristupa su: ***Adaboost*** [15, 16] i ***Gradient-Boosting*** [17] (npr. XGBoost [9]).

Bagging [5] (skraćeno od eng. "Bootstrap Aggregating") pristupi kombiniraju izlaze različitih baznih modela zasebno naučenih na rješavanje traženog prediktivnog zadatka. Osnovni modeli se uče nad podskupovima podataka dobivenima ponovnim uzorkovanjem skupa za treniranje sa zamjenama, tj. ***bootstrap*** metodom ponovnog uzorkovanja. Bagging metoda smanjuje varijancu osnovnih modela, zato je posebno pogodna za modele visoke varijance, kao što su stabla odluke. ***Stabla odluke*** (eng. de-

cision trees) su jako osjetljiva na podatke na kojima su naučena, te su međusobno jako različita u slučaju da su učena na drugom podskupu podataka. Upravo zato je bagging vrlo često korištena metoda u građenju ansambla s velikim brojem stabala odluke, tj. ***šuma odluke***.

Problem korištenja standardnih pristupa za građenje stabla odluke (npr. CART) je u tome što oni pri svakom grananju na pohlepan (eng. greedy) način minimiziraju pogrešku, provjeravajući sve ulazne varijable za optimalno rješenje. Zbog toga, usprkos upotrebi bagging pristupa, većina stabala sadržava veliku međusobnu strukturalnu sličnost. U članku [6], Breiman je pokazao da se generalizacijska pogreška može smanjiti ako se smanji sličnost između stabala u šumi odluke, te predlaže randomizacijsku shemu za dekoreliranje predikcija stabala, bez značajnog smanjivanja njihove prediktivne moći. Ovaj pristup je nazvan ***slučajna šuma*** (eng. random forest), te predstavlja poboljšanje bagging metode za građenje šuma odluke smanjenjem korelacije između stabala. Kod slučajnih šuma, stabla odluke se više ne granaju pohlepno (optimirajući između svih varijabli), nego su ograničena na pretragu na vrlo malom podskupu slučajno odabranih varijabli.

Stacking pristup predložen je u članku [65] kao generalni pristup za ostvarivanje bolje generalizacijske točnosti, a metoda je nazvana ***stack generalizacija*** (eng. stacked generalization). Cilj metode je pronaći najbolji generalizator, estimator roditeljske funkcije. Za svaki skup podataka postoji velik broj mogućih generalizatora koji mogu ekstrapolirati izvan skupa podataka za učenje. Kod neparametrijskih statističkih pristupa, kao što je unakrsna validacija i bootstrapping bira se jedan (najbolji) generalizator. S druge strane, stack generalizacija je parametrijski pristup kod kojeg se kombiniraju generalizatori, te se može promatrati kao napredna verzija neparametrijskih statističkih metoda [65]. Ključna razlika stacking i bagging metoda za ansamble modela je u tome što stacking nije ograničen na isti tip osnovnog modela kojeg učimo na različitim podskupovima (kao bagging), već je općenit, tj. odnosi se na različite osnovne modele (generalizatore), neovisno o tome koji je izvor njihove različitosti.

7.2. Ansambli dubokih modela

Stabla odluke nisu jedini model za koji koristimo metode ansambla, za bazni model možemo koristiti i druge modele. ***Ansambli dubokih modela***, poznati i pod nazivom ***duboki ansambli*** (eng. deep ensembles), pokazali su se kao uspješna metoda za predikciju, te se često koriste na različitim natjecanjima strojnog učenja. Uz poboljšanje prediktivne točnosti ansambli dubokih modela mogu se koristiti za procjenu predik-

tivne nesigurnosti [37].

U članku [37] autori predlažu recept kako ostvariti dobru prediktivnu točnost i estimaciju prediktivne nesigurnosti primjenom ansambla DNN modela. Recept se sastoji od tri stavke: (1) koristiti odgovarajuća pravila bodovanja (eng. proper scoring rules) za kriterij učenja; (2) koristiti učenje s napadačkim primjerima (eng. adversarial training); te (3) učiti ansambl modela.

Kod učenja DNN modela ne vrijede jednaka pravila kao što su bila kod stabala odluke. DNN modeli puno su stabilniji od stabala odluke, te se njihova prediktivna točnost poboljšava s brojem uzoraka u skupu za učenje. U članku [23], autori su pokazali da je bagging nepotreban ako se u proces odabira podskupa može ugraditi dodatni izvor slučajnosti. U skladu s tim autori članka [37] za učenje baznih DNN modela koriste cijeli skup za učenje. Također su došli do zaključka da je dovoljno upotrijebiti slučajnu inicijalizaciju vrijednosti parametara DNNa i slučajno permutiranje rasporeda uzoraka za učenje kako bi se ostvarili dobri rezultati u praksi. S druge strane, upotreba bagging pristupa samo je narušila prediktivnu točnost krajnjeg modela [37], što je u skladu s rezultatima dobivenima u članku [38].

Autori također predlažu upotrebu neprijateljskih primjera u učenju pojedinačnih DNN modela [37]. ***Neprijateljski primjer*** (eng. adversarial examples) [57] su ulazni objekti jako slični originalnim primjerima za učenje (naizgled isti), ali su pogrešno klasificirani od strane DNN modela. Autori u [27] predlažu jednostavnu i brzu metodu za generiranje napadačkih primjera baziranu na predznaku gradijenta. Na ovaj način možemo generirati nove primjere za učenje koji vode u smjeru u kojem je vjerojatno da će se povećati pogreška modela. Korištenje napadačkih primjera u učenju modela naziva se napadačko učenje (eng. adversarial training), a pokazalo se da pospješuje robusnost klasifikatora [27]. Napadačko učenje se može upotrijebiti kao efikasna metoda za izglađivanje prediktivne distribucije [37].

U članku [37] koriste uniformne težine miješanja osnovnih modela u ansambl, tako da je predikcija: $p(y|x) = \frac{1}{M}\sum_{m=1}^{M} p_{\theta_m}(y|x, \theta_m)$, gdje su θ_m parametri m-tog modela, gdje je $m \in \{1, ...M\}$. Kod klasifikacije to odgovara usrednjavanju prediktivne vjerojatnosti, a kod regresije miješanju Gaussovskih distribucija [37].

7.3. Bayesovsko usrednjavanje modela

Bayesovsko usrednjavanje modela (eng. Bayesian Model Averaging, BMA) je primjena Bayesovskog zaključivanja na problemu selekcije i zajedničke estimacije modela [14]. Bayesovsko zaključivanje dolazi do posebnog značaja kod DNN modela.

Podaci za učenje ne specificiraju u potpunosti DNN modele, ovisno o postavkama možemo dobiti veliki broj različitih DNN modela visoke prediktivne točnosti [64]. Upravo u ovom slučaju marginalizacija može napraviti najveću razliku u točnosti i kalibraciji modela, a duboke ansamble možemo promatrati kao aproksimaciju Bayesovske marginalizacije [64].

U mnogim slučajevima prediktivnu distribuciju možemo dobiti iz:

$$p(y|x, \mathcal{D}) = \int p(y|x, \omega)p(\omega|\mathcal{D})d\omega; \tag{7.1}$$

gdje su y izlazne vrijednosti, x ulazni objekti, ω parametri modela $f(x; \omega)$, a $\mathcal{D}$ skup podataka [64]. Jednadžba 7.1 reprezentira BMA. Umjesto da sav ulog stavljamo na jednu hipotezu, tj. jedinstvene postavke parametara ω, mi želimo uzeti u obzir sve moguće postavke parametara, s težinama njihovih aposteriornih vjerojatnost [64]. BNA reprezentira epistemičku nesigurnost, tj. nesigurnost dobrog odabira parametara (hipoteze) modela.

8. Selektivna predikcija

Vrlo važna primjena procjene nesigurnosti modela je selektivna predikcija. Generalna ideja selektivnih prediktora je da na osnovu procijenjene sigurnosti model daje predikciju samo za podskup uzoraka na kojem je prediktivna sigurnost dovoljno visoka (iznad definiranog praga).

Za selektivnu predikciju postoje razne mjere za procjenu nesigurnosti, kao što su: (1) Područje ispod krivulje odnosa specifičnosti i osjetljivosti (AUROC); (2) Područje ispod krivulje odnosa preciznosti i osjetljivosti (AUPR); i (3) Područje ispod krivulje odnosa rizika i pokrivenosti (AURC).

Područje ispod krivulje odnosa specifičnosti i osjetljivosti (eng. Area Under Receiver Operating Characteristic curve, AUROC) [30] se izračunava iz krivulje specifičnosti i osjetljivosti. Specifičnost predstavlja udio pravih negativnih (eng. truth negative, TN) u ukupnom broju negativnih primjera, a osjetljivost je udio pravih pozitivnih (eng. truth positive, TP) u ukupnom broju pozitivnih primjera:

$$specificnost = \frac{TN}{TN + FP}; \tag{8.1}$$

$$osjetljivost = \frac{TP}{TP + FN}; \tag{8.2}$$

Područje ispod krivulje odnosa preciznosti i osjetljivosti (eng. Area Under Precision-Recall curve, AUPR) se izračunava iz krivulje preciznosti i osjetljivosti. Preciznost predstavlja udio pravih pozitivnih uzoraka od svih uzoraka s pozitivnom predikcijom:

$$preciznost = \frac{TP}{TP + FP}; \tag{8.3}$$

Područje ispod krivulje odnosa rizika i pokrivenosti (eng. Area Under Risk-Coverage, AURC) [12] se izračunava iz krivulje rizika i pokrivenosti. Pokrivenost predstavlja udio ulaznih podataka koje model procjenjuje sam, a rizik predstavlja razinu rizika, tj. pogrešku predikcije modela za odabrani podskup ulaznih podataka [12]:

$$pokrivenost = \frac{|X_h|}{|X|}; \tag{8.4}$$

$$rizik = \mathcal{L}(\hat{Y_h}); \tag{8.5}$$

gdje je X_h skup selektiranih ulaznih podataka, X skup svih ulaznih podataka, $\hat{Y_h}$ skup predikcija dobivenih nad X_h, a $\mathcal{L}$ je odabrana funkcija pogreške za procjenu kvalitete predikcije.

Autori u [12] su napravili usporedbu ove tri metrike (AUROC, AUPR i AURC), te su pokazali da je između njih jedino AURC pouzdana u slučaju promjene predikcijskog modela [12].

9. Kalibracija nesigurnosti modela

Kalibracija modela je proces prilagodbe parametara kako bi rezultati iz modela bolje opisivali promatrane podatke [33]. U kontekstu nesigurnosti, kalibracija modela omogućuje da izlaz iz modela bolje predstavlja njegovu prediktivnu sigurnost [42].

9.1. Procjena kvalitete kalibracije

Postoje razne metrike za procjenu kvalitete kalibracije modela. Autori u [37] navode da se za procjenu kvalitete kalibracije mogu koristiti ***strogo odgovarajuća pravila bodovanja*** (eng. strictly proper scoring rules) [26], kao što su Brierovo bodovanje i negativna log-izglednost. Uz njih, kod kalibracije DNN modela često se koriste očekivana kalibracijska pogreška (ECE) i maksimalna kalibracijska pogreška (MCE).

Brierovo bodovanje (eng. Brier score) [7], tj. ***kvadratno bodovanje*** (eng. quadratic score), jedna je od često korištenih mjera kvalitete. Za klase $k \in \{1, ..., K\}$, gdje je $\hat{p}_i^{(k)}$ procjena vjerojatnosti klase k za uzorak $i \in \{1, ..., n\}$, gdje je y_i labela uzorka i, tada je Brierovo bodovanje za za jedan uzorak :

$$B_i^{(k)} = -\sum_{j=1}^{K}(\delta_i^{(jk)} - p_j^{(k)})^2 = 2p_i^{(k)} - \sum_{j=1}^{K}(p_i^{(j)})^2 - 1; \tag{9.1}$$

gdje je $\delta_i^{(jk)} = 1$ ako $j = k$, a inače $\delta_i^{(jk)} = 0$ [26], dok je bodovanje za cijeli skup podataka:

$$B = \sum_{i=1}^{n} B_i^{(y_i)} = \sum_{i=1}^{n}(2p_i^{(y_i)} - \sum_{j=1}^{K}(p_i^{(j)})^2 - 1); \tag{9.2}$$

$$B = 2\sum_{i=1}^{n} p_i^{(y_i)} - \sum_{i=1}^{n}\sum_{j=1}^{K}(p_i^{(j)})^2 - n; \tag{9.3}$$

Negativna log-izglednost (eng. negative log likelihood, NLL) je standardna mjera za procjenu kvalitete probabilističkih modela [18], a u kontekstu dubokog učenja također se naziva i ***gubitak unakrsne entropije*** (eng. cross entropy loss). Za dani probabi-

listički model $\hat{\pi}(Y|X)$ i n uzoraka, NLL definiramo jednadžbom:

$$\mathcal{L} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i|x_i)). \tag{9.4}$$

Procjenu točnosti predikcije nije moguće napraviti nad jednim uzorkom. Upravo iz tog razloga koristimo ***pretince*** (eng. bin). Ideja ***strategije pretinaca*** je da odabrani skup ulaznih podataka raspodijelimo na podskupove na osnovu dobivene prediktivne sigurnosti. Pretinci su definirani s intervalom prediktivne sigurnosti, a u praksi se najčešće koriste ravnomjerno raspoređeni intervali. U tom slučaju za odabrani broj pretinaca M i pojedini pretinac $m \in \{1, ..., M\}$, neka je B_m skup svih indeksa ulaznog skupa na intervalu procijenjene prediktivne vjerojatnosti $\mathcal{I}_m = (\frac{m-1}{M}, \frac{m}{M}]$. Tada točnost pretinca B_m možemo definirati kao:

$$tocnost(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i); \tag{9.5}$$

gdje je $\mathbf{1}$ indikatorska funkcija, a $\hat{y}_i$ i y_i su estimirana i točna labela klase za uzorak i [29]. Srednju sigurnost predikcije u pretincu B_m definiramo kao:

$$sigurnost(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i; \tag{9.6}$$

gdje je $\hat{p}_i$ procijenjena prediktivna sigurnost za uzorak i.

Dijagram pouzdanosti (eng. reliability diagram) prikazuje prediktivnu točnost u odnosu na sigurnost, a temelji se na pristupu s pretincima. U njemu su uzorci raspodijeljni u pretince odabranih intervala sigurnosti, a sigurnost (9.6) je iskazana na x-osi, a točnost (9.5) na y-osi dijagrama. Dobro kalibriran model bi trebao imati vrijednosti na pravcu $sigurnost = tocnosti$ na dijagramu pouzdanosti, tj. trebalo bi vrijediti $sigurnost(B_m) = tocnost(B_m)$ za svaki pretinac $m \in \{1, ..., M\}$. Odstupanje dobivenih vrijednosti od tog pravca predstavlja pogrešku u kalibraciji, tako da je dijagram pouzdanosti jednostavan način vizualne provjere kvalitete kalibriranosti modela.

Očekivana kalibracijska pogreška (eng. Expected Calibration Error, ECE) pokazuje koliko je očekivanje razlike između sigurnosti predikcije i točnosti:

$$ECE = \mathbb{E}_{\hat{P}}\Big[\Big|\mathbb{P}(\hat{Y} = y|\hat{P} = p) - p\Big|\Big]. \tag{9.7}$$

Aproksimaciju ECE vrijednosti možemo dobiti strategijom pretinaca jednadžbom:

$$\hat{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big|tocnost(B_m) - sigurnost(B_m)\Big|. \tag{9.8}$$

Maksimalna kalibracijska pogreška (eng. Maximum Calibration Error, MCE) je mjera koju koristimo kada želimo minimizirati najveću razliku između procijenjene sigurnosti i točnosti:

$$MCE = \max_{\hat{p}\in[0,1]} \Big|\mathbb{P}(\hat{Y} = y|\hat{P} = p) - p\Big|. \tag{9.9}$$

Aproksimaciju MCE vrijednosti možemo dobiti strategijom pretinaca jednadžbom:

$$\hat{MCE} = \max_{m\in\{1,\dots,M\}} \big|tocnost(B_m) - sigurnost(B_m)\big|. \tag{9.10}$$

Sve prethodno spomenute mjere za provjeru kalibriranosti modela jako ovise o strategiji odabiranja pretinaca. Autori u [12] navode tri potencijalna nedostatka strategije pretinaca:

1. ***Nevidljiva pogreška*** (eng. undetectable error) - greška koja nastaje na podskupu (manjem intervalu) nekog pretinca. Ova greška nastaje zbog velikih intervala pretinaca i jake neuniformnosti distribucije podataka;

2. ***Interna kompenzacija*** (eng. internal compensation) - unutar pretinca moguće je da dio uzoraka ima pozitivnu, a dio negativnu kalibracijsku pogrešku. To u konačnici može smanjiti dobivenu vrijednost kalibracijske pogreške;

3. ***Pogreška u procjeni točnosti*** (eng. inaccurate accuracy estimation) - ona nastaje u slučaju malog broja uzoraka u pretincu.

Postoje različite ***strategije pretinaca***. Uz strategiju pretinaca s ***ravnomjernim intervalima***, koriste se i pretinci s ***ravnomjernim brojem uzoraka***, koji povećavaju rezoluciju za intervale s gustom distribucijom uzoraka (što je poželjno), ali zato smanjuju rezoluciju u području s rijetkom distribucijom uzoraka (što nije poželjno).

Autori u [12] predlažu ***adaptivne pretince*** (eng. adaptive bining) s ciljem zadržavanja prednosti, a umanjivanja nedostatka prethodno spomenutih strategija pretinaca. Adaptivnim pretincima možemo ostvariti dinamički balans između rezolucije i kvalitete procjene točnosti, koja je ovisna o broju uzoraka u pretincu. To ostvarujemo tako što uske intervale velike gustoće uzoraka proširujemo da imamo bolju procjenu točnosti, dok široke intervale s malim brojem uzoraka sužavamo da imamo bolju rezoluciju. Autori u [12] koriste sljedeću jednadžbu za procjenu broja uzoraka u pojedinom pretincu potrebnih za procjenu točnosti:

$$n = 0,25(\frac{Z_{\alpha/2}}{\epsilon})^2 \tag{9.11}$$

gdje je ϵ marginalna pogreška sigurnosti, $Z_{\alpha/2}$ je Z-vrijednost standardne normalne distribucije (SND), a $1-\alpha$ interval sigurnosti. Z-vrijednost je standardizirana vrijednost, a imožemo izračunati iz sljedeće formule:

$$Z = (x - \mu)/\sigma \tag{9.12}$$

gdje je μ srednja vrijednost, a σ standardna devijacija distribucije uzoraka.

9.2. Metode kalibracije

Postoje razne metode kalibracije (tj. rekalibracije) prediktivne nesigurnosti modela, ovdje ćemo istaknuti neke od značajnijih metoda temeljene na pristupu skaliranjem i pristupu histograma pretinaca (eng. histogram binning).

Binarni klasifikator daje predikciju vjerojatnosti $\hat{p}_i$ da ulazni uzorak x_i pripada nekoj klasi $y_i = 1$, $y \in \{0, 1\}$, gdje je $z_i \in \mathbb{R}$ neprobabilistički izlaz, tj. ***"logit"*** [29]. Predikciju vjerojatnosti $\hat{p}_i$ izračunavamo iz z_i sigmoidnom funkcijom σ, tako da je $\hat{p}_i = \sigma(z_i)$ [29]. Cilj kalibracije binarnih klasifikatora je postići kalibrirane vjerojatnosti $\hat{q}_i$ iz y_i, $\hat{p}_i$ i z_i.

Višeklasni klasifikator daje predikciju klase $\hat{y}_i \in \{1, ..., K\}, K > 2$ i njene vjerojatnosti $\hat{p}_i$ za svaki ulazni uzorak x_i. Logiti modela su vektori z_i, gdje je $\hat{y}_i = \arg\max_k z_i^{(k)}$, a $\hat{p}_i$ se najčešće računa softmax funkcijom σ_{SM} [29]:

$$\sigma_{SM}(z_i)^{(k)} = \frac{\exp\left(z_i^{(k)}\right)}{\sum_{j=1}^{K} \exp\left(z_i^{(j)}\right)}; \tag{9.13}$$

$$\hat{p}_i = \max_k \sigma_{SM}(z_i)^{(k)}. \tag{9.14}$$

Cilj kalibracije višeklasnih klasifikatora je dati kalibrirane vjerojatnosti $\hat{q}_i$ za novu predikciju klase $\hat{y}_i'$ iz y_i, $\hat{y}_i$, $\hat{p}_i$ i z_i [29].

U nastavku ćemo neke od metoda predstaviti u kontekstu binarne klasifikacije: histogram pretinaca; izotoničnu regresiju; Bayesovski pretinci u kvantilima; i Platt skaliranje. Sve se te metode mogu proširiti na slučaj višeklasne predikcije tako da svedemo problem na K jedan-protiv-svih (tj. binarnih) problema [67]. Za svaku od $k = 1, ..., K$ klasa definiramo binarni kalibracijski problem gdje je labela $I(y_i = k)$, a predviđena vjerojatnost $\sigma_{SM}(z_i)^{(k)}$ [29]. U konačnici imamo K kalibriranih modela, po jedan za svaku klasu. Za vrijeme testa dobivamo vektor ne-normaliziranih vjerojatnosti za klase $[\hat{q}_i^{(1)}, ..., \hat{q}_i^{(K)}]$, gdje je $\hat{q}_i^{(k)}$ kalibrirana vjerojatnost za klasu k [29]. Nova predikcija klase $\hat{y}_i'$ je indeks najvećeg elementa tog vektora, a ukupna estimirana vjerojatnost $\hat{q}_i$ postaje maksimalna vrijednost tog vektora normalizacirana faktorom $\sum_{k=1}^{K} \hat{q}_i^{(k)}$.

Histogram pretinaca (eng. histogram binning) [66] je jednostavna neparametrijska kalibracijska metoda, gdje sve nekalibrirane predikcije $\hat{p}_i$ raspodjeljujemo u pretince $B_1, ..., B_M$. Svakom pretincu se pridjeljuje kalibracijska vrijednost θ_m, tako da ako $\hat{p}_i$ upada u taj pretinac, tada postavljamo $\hat{q}_i = \theta_m$. Ovdje se primjenjuju pretinici s ravnomjernim intervalima ili s ravnomjernim brojem uzoraka, s granicama $0 = a_1 \leq a_2 \leq ... \leq a_{M+1} = 1$, gdje je svaki pretinac B_m definiran intervalom $(a_m, a_{m+1}]$ [29]. Predikcije θ_i su odabrane tako da minimiziraju kvadratnu pogrešku pretinca, a u slučaju binarne klasifikacije jednadžba je:

$$\min_{\theta_1,...,\theta_M} \sum_{m=1}^{M} \sum_{i=1}^{n} \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1})(\theta_m - y_i)^2 \tag{9.15}$$

gdje je $\mathbf{1}$ indikatorska funkcija. U slučaju pretinaca s ravnomjernim intervalima θ_m rezultira srednjoj vrijednosti točnih predikcija od uzoraka u pretincu B_m [29].

Izotonična regresija (eng. isotonic regression) [67] je neparametrijska kalibracijska metoda. Ona uči ***komadno konstantnu funkciju*** (eng. piecewise constant function) f da transformira nekalibrirane izlaze, $\hat{q}_i = f(\hat{p}_i)$, na način da minimizira kvadratnu pogrešku $\sum_{i=1}^{n}(f(\hat{p})_i - y_i)^2$ [29]. Budući da je f ograničena da bude komadno konstantna funkcija, tada za binarnu klasifikaciju vrijedi:

$$\min_{\substack{M;\\ \theta_1,...,\theta_M;\\ a_1,...,a_{M+1}}} \sum_{m=1}^{M} \sum_{i=1}^{n} \mathbf{1}(a_m \leq \hat{p}_i < a_{m+1})(\theta_m - y_i)^2 \tag{9.16}$$

gdje je M broj intervala, granice $0 = a_1 \leq a_2 \leq ... \leq a_{M+1} = 1$, a vrijednosti funkcije $\theta_1, ..., \theta_n$ [29]. Ovom parametrizacijom izotonična regresija predstavlja generalizaciju histograma pretinca u kojem su granice i vrijednosti pretinca združeno optimirani [29].

Bayesovski pretinci u kvantilima (eng. Bayesian Binning into Quantiles, BBQ) [43] je proširenje histograma pretinca upotrebom ***Bayesovskog usrednjavanja modela*** (eng. Bayesian model averaging, BMA) [29]. BBQ marginalizira sve moguće sheme pretinaca i daje $\hat{q}_i$. Svaku shemu pretinaca možemo opisati s parom $(M, \mathcal{I})$, gdje M predstavlja broj pretinaca, a $\mathcal{I}$ cijepanje intervala $[0, 1]$ na disjunktivne intervale $(0 = a_1 \leq a_2 \leq ... \leq a_{M+1})$. Parametri sheme pretinaca su $\theta_1, ..., \theta_M$. Dok histogram pretinaca i izotonična regresija daju samo jednu shemu pretinaca, BBQ uzima u obzir prostor S mogućih shema pretinaca za dani validacijski skup podataka D, a budući da je validacijski skup konačan, onda je i S konačan [29]. BBQ izvodi Bayesovsko usrednjavnje od vjerojatnosti dobivenih za svaku pojedinačnu shemu, a za binarnu klasifikaciju jednadžba je:

$$\mathbb{P}(\hat{q}_{te}|\hat{p}_{te}, D) = \sum_{s \in S} \mathbb{P}(\hat{q}_{te}, S = s|\hat{p}_{te}, D); \tag{9.17}$$

$$\mathbb{P}(\hat{q}_{te}|\hat{p}_{te}, D) = \sum_{s \in S} \mathbb{P}(\hat{q}_{te}|\hat{p}_{te}, S = s, D)\mathbb{P}(S = s|D); \tag{9.18}$$

gdje je $\mathbb{P}(\hat{q}_{te}|\hat{p}_{te}, S = s, D)$ kalibrirana vjerojatnost za odabir sheme pretinca s. Uz uvjet uniformnog priora, izraz $\mathbb{P}(S = s|D)$ se može dobiti iz Bayesovog pravila:

$$\mathbb{P}(S = s|D) = \frac{\mathbb{P}(D|S = s)}{\sum_{s' \in S} \mathbb{P}(D|S = s')}. \tag{9.19}$$

Parametri $\theta_1, ..., \theta_M$ se mogu opisati kao parametri od M nezavisnih binomialnih distribucija, i u slučaju Beta priora na $\theta_1, ..., \theta_M$ možemo dobiti zatvoreni izraz za marginalnu vjerojatnost $P(D|S = s)$, te možemo izračunati $P(\hat{q}_{te}|\hat{p}_{te}, D)$ za bilo koji testni uzorak [29].

Platt skaliranje (eng. Platt scaling, PS) [49] je parametrijski pristup kalibraciji modela. Ne-probabilističke predikcije klasifikatora koriste se kao ulazne značajke za model logističke regresije, koji se uči na validacijskom skupu podataka da vraća vjerojatnosti [29]. U kontekstu neuronskih mreža, predstavljenim u [44], Platt skaliranje uči skalirajuće parametre $a, b \in \mathbb{R}$ da daju $\hat{q}_i = \sigma(az_i + b)$ kao kalibrirane vjerojatnosti. Parametri a i b mogu se optimirati korištenjem NLL mjere pogreške na validacijskom skupu, pri tom je važno da se parametri osnovne neuronske mreže zaključaju [29].

Matrično i vektorsko skaliranje (eng. matrix and vector scaling) su dva proširenja Platt skaliranja za klasifikatore za više klasa [29]. Ako za ulazne objekte x_i imamo vektor logita z_i prije prolaska kroz softmax sloj σ_{SM}. Tada matrično skaliranje primjenjuje linearnu transformaciju $Wz_i + b$ na logite, a jednadžbe su:

$$\hat{q}_i = \max_k \sigma_{SM}(Wz_i + b)^{(k)} \tag{9.20}$$

$$\hat{y}_i' = \arg\max_k \sigma_{SM}(Wz_i + b)^{(k)} \tag{9.21}$$

gdje su parametri W i b optimirani s obzirom na NLL i to na validacijskom skupu podataka [29]. Budući da broj parametara u matričnom skaliranju raste s kvadratom broja klasa, u [29] predlažu vektorsko skaliranje kao varijantu matričnog skaliranja u slučaju velikog broja klasa. U vektorskom skaliranju je matrica W dijagonalna matrica.

Temperaturno skaliranje (eng. temperature scaling, TS) [29] je vrlo jednostavno proširenje Platt skaliranja. Ideja je da se pronađe jedan skalarni parametar $T > 0$ koji skalira logit vektor z_i podjednako za sve klase modela, tako da je kalibrirana sigurnost predikcije:

$$\hat{q}_i = \max_k \sigma_{SM}(z_i/T)^{(k)}. \tag{9.22}$$

Parametar T se naziva ***temperatura***, koji u slučaju: (1) $T > 1$ "razblažuje" softmax, tj. povećava entropiju izlaza; (2) $T \to \infty$ vjerojatnosti $\hat{q}_i$ se približavaju $1/K$, tj. svaka

od K klasa je gotovo jednako vjerojatna, što predstavlja maksimalnu nesigurnost; (3) $T = 1$ daje originalnu vjerojatnost $\hat{p}_i$; (4) $T \to 0$ sažima vjerojatnosti u jednu klasu $\hat{q}_i = 1$ [29]. Parametar T se optimira s obzirom na NLL na validacijskom setu, a budući da parametar T ne mijenja maximum softmax funkcije, predikcija osnovnog modela $\hat{y}'$ ostaje ne promijenjena[29]. To znači da TS ne utječe na prediktivnu točnost modela. TS smanjuje kalibracijsku pogrešku na razini cijelog modela, što je korisno u slučaju da je cijeli model pretjerano siguran (eng. overconfident) ili nesiguran (eng. underconfindent) [42].

Skalirajući kalibrator pretinaca (eng. scaling-binning calibrator) je metoda kalibriranja modela predložena u članku [36]. U ovoj metodi se povezuju dva različita pristupa, skaliranje i histogram pretinca, kako bi se iskoristile pozitivne, a smanjile negativne karakteristike tih pristupa. Skalirajuće metode zahtjevaju manji broj uzoraka ali daju slabiju kalibraciju, te ne mogu procijeniti koliko su dobro kalibrirane. S druge strane histogram pretinaca daje mjerljivu kalibracijsku pogrešku, ali je neefikasan što se tiče potrebnog broja uzoraka. Skalirajući kalibrator pretinaca prilagođava parametrijsku funkciju tako da smanjuje varijancu, potom, vrijednosti te funkcije raspodjeljuje u pretince kako bi se osigurala dobra kalibracija [36]. Kod ove metode, skup podataka T s brojem uzoraka n dijeli se na 3 podskupa: T_1, T_2, T_3. Skalirajući kalibrator pretinaca sastoji se od tri koraka, a u svakom koraku koristimo drugi podskup.

Prvi korak je ***prilagođavanje parametrijske funkcije***:

$$g = \arg\min_{g \in G} \sum_{(x,y) \in T_1} (y - g(x))^2. \tag{9.23}$$

Drugi korak je ***konstrukcija sheme pretinaca***. Intervali pretinaca $\mathcal{I}_j$ se odabiru tako da podjednaki broj uzoraka s vrijednostima $g(x_i)$, za $(x_i, y_i) \in T_2$, upada u svaki od pretinaca B_j, za $j \in \{1, ..., m\}$, dakle strategijom pretinaca s ravnomjernim brojem uzoraka.

Treći korak je ***diskretizacija***. Diskretiziramo funkciju g tako da njene izlazne vrijednosti usrednjujemo za svaki pretinac. Ako je $\mu(S) = \frac{1}{|S|} \sum_{s \in S} s$ srednja vrijednost skupa vrijednosti S, a B_j skup svih uzoraka iz skupa T_3 čija $g(z_i)$ vrijednost upada unutar intervala $\mathcal{I}_j$, tada je srednja vrijednost j-tog pretinca:

$$\hat{\mu}[j] = \mu(\{g(x_i) | g(x_i) \in \mathcal{I}_j \wedge (x_i, y_i) \in T_3\}). \tag{9.24}$$

Sada definiramo funkciju β koja daje indeks pretinca u kojem se uzorak nalazi, dakle $\beta(x) = j$, ako je $x \in B_j$. Tada definiramo skalirajući kalibrator pretinaca:

$$\hat{g}_B(z) = \hat{\mu}[\beta(g(x))]; \tag{9.25}$$

što znači da u za kalibriranu vrijednost dajemo srednju vrijednost pretinca u koji upada vrijednost $g(x)$.

10. Literatura

[1] Adaptive binning. `https://github.com/yding5/AdaptiveBinning`. Accessed: 2021-08-03.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. U ***12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)***, stranice 265–283, 2016.

[3] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, i Ryan J Tibshirani. Predictive inference with the jackknife+. ***The Annals of Statistics***, 49(1):486–507, 2021.

[4] José M Bernardo i Adrian FM Smith. ***Bayesian theory***, svezak 405. John Wiley & Sons, 2009.

[5] Leo Breiman. Bagging predictors. ***Machine learning***, 24(2):123–140, 1996.

[6] Leo Breiman. Random forests. ***Machine learning***, 45(1):5–32, 2001.

[7] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. ***Monthly weather review***, 78(1):1–3, 1950.

[8] Lars Carlsson, Martin Eklund, i Ulf Norinder. Aggregated conformal prediction. U ***IFIP International Conference on Artificial Intelligence Applications and Innovations***, stranice 231–240. Springer, 2014.

[9] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. ***R package version 0.4-2***, 1(4): 1–4, 2015.

[10] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, i Zheng Zhang. Mxnet: A flexible and efficient

machine learning library for heterogeneous distributed systems. ***arXiv preprint arXiv:1512.01274***, 2015.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, i Li Fei-Fei. Imagenet: A large-scale hierarchical image database. U ***2009 IEEE conference on computer vision and pattern recognition***, stranice 248–255. Ieee, 2009.

[12] Yukun Ding, Jinglan Liu, Jinjun Xiong, i Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. U ***Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops***, stranice 4–5, 2020.

[13] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, i Vladimir Vovk. Plug-in martingales for testing exchangeability on-line. ***arXiv preprint arXiv:1204.3251***, 2012.

[14] Tiago M Fragoso, Wesley Bertoli, i Francisco Louzada. Bayesian model averaging: A systematic review and conceptual classification. ***International Statistical Review***, 86(1):1–28, 2018.

[15] Yoav Freund i Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. ***Journal of computer and system sciences***, 55(1):119–139, 1997.

[16] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. U ***icml***, svezak 96, stranice 148–156. Citeseer, 1996.

[17] Jerome H Friedman. Stochastic gradient boosting. ***Computational statistics & data analysis***, 38(4):367–378, 2002.

[18] Jerome H Friedman. ***The elements of statistical learning: Data mining, inference, and prediction***. springer open, 2017.

[19] Yarin Gal. ***Uncertainty in deep learning***. Doktorska disertacija, University of Cambridge, 2016.

[20] Yarin Gal i Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. U ***international conference on machine learning***, stranice 1050–1059. PMLR, 2016.

[21] A. Gammerman, V. Vovk, i V. Vapnik. Learning by transduction. U ***Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence***, UAI'98, stranica 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

[22] Seymour Geisser. The predictive sample reuse method with applications. ***Journal of the American statistical Association***, 70(350):320–328, 1975.

[23] Pierre Geurts, Damien Ernst, i Louis Wehenkel. Extremely randomized trees. ***Machine learning***, 63(1):3–42, 2006.

[24] Asma Ghandeharioun, Brian Eoff, Brendan Jou, i Rosalind Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. U ***2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)***, stranice 4202–4206. IEEE, 2019.

[25] Ryan Giordano, Michael I Jordan, i Tamara Broderick. A higher-order swiss army infinitesimal jackknife. ***arXiv preprint arXiv:1907.12116***, 2019.

[26] Tilmann Gneiting i Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. ***Journal of the American statistical Association***, 102(477):359–378, 2007.

[27] Ian J Goodfellow, Jonathon Shlens, i Christian Szegedy. Explaining and harnessing adversarial examples. ***arXiv preprint arXiv:1412.6572***, 2014.

[28] Charles Miller Grinstead i James Laurie Snell. ***Introduction to probability***. American Mathematical Soc., 2012.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, i Kilian Q Weinberger. On calibration of modern neural networks. U ***International Conference on Machine Learning***, stranice 1321–1330. PMLR, 2017.

[30] James A Hanley i Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. ***Radiology***, 143(1):29–36, 1982.

[31] Max Hinne, Quentin F Gronau, Don van den Bergh, i Eric-Jan Wagenmakers. A conceptual introduction to bayesian model averaging. ***Advances in Methods and Practices in Psychological Science***, 3(2):200–215, 2020.

[32] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, i Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. U ***Proceedings of the 22nd ACM international conference on Multimedia***, stranice 675–678, 2014.

[33] Marc C Kennedy i Anthony O'Hagan. Bayesian calibration of computer models. ***Journal of the Royal Statistical Society: Series B (Statistical Methodology)***, 63 (3):425–464, 2001.

[34] Aditya Khosla, Akhil S Raju, Antonio Torralba, i Aude Oliva. Understanding and predicting image memorability at a large scale. U ***Proceedings of the IEEE international conference on computer vision***, stranice 2390–2398, 2015.

[35] Alex Krizhevsky, Ilya Sutskever, i Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. ***Advances in neural information processing systems***, 25:1097–1105, 2012.

[36] Ananya Kumar, Percy Liang, i Tengyu Ma. Verified uncertainty calibration. ***arXiv preprint arXiv:1909.10155***, 2019.

[37] Balaji Lakshminarayanan, Alexander Pritzel, i Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. ***arXiv preprint arXiv:1612.01474***, 2016.

[38] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, i Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. ***arXiv preprint arXiv:1511.06314***, 2015.

[39] Henrik Linusson, Ulf Norinder, Henrik Boström, Ulf Johansson, i Tuve Löfström. On the calibration of aggregated conformal predictors. U ***Conformal and probabilistic prediction and applications***, stranice 154–173. PMLR, 2017.

[40] Andrey Malinin i Mark Gales. Predictive uncertainty estimation via prior networks. ***arXiv preprint arXiv:1802.10501***, 2018.

[41] Rupert G Miller. The jackknife-a review. ***Biometrika***, 61(1):1–15, 1974.

[42] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, i Mario Lucic. Revisiting the calibration of modern neural networks. ***arXiv preprint arXiv:2106.07998***, 2021.

[43] Mahdi Pakdaman Naeini, Gregory Cooper, i Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. U ***Twenty-Ninth AAAI Conference on Artificial Intelligence***, 2015.

[44] Alexandru Niculescu-Mizil i Rich Caruana. Predicting good probabilities with supervised learning. U ***Proceedings of the 22nd international conference on Machine learning***, stranice 625–632, 2005.

[45] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, i Alex Gammerman. Inductive confidence machines for regression. U ***European Conference on Machine Learning***, stranice 345–356. Springer, 2002.

[46] Nicolas Papernot i Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. ***arXiv preprint arXiv:1803.04765***, 2018.

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, i Adam Lerer. Automatic differentiation in pytorch. 2017.

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. ***arXiv preprint arXiv:1912.01703***, 2019.

[49] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. ***Advances in large margin classifiers***, 10 (3):61–74, 1999.

[50] Maurice H Quenouille. Approximate tests of correlation in time-series 3. U ***Mathematical Proceedings of the Cambridge Philosophical Society***, svezak 45, stranice 483–484. Cambridge University Press, 1949.

[51] Maurice H Quenouille. Notes on bias in estimation. ***Biometrika***, 43(3/4):353–360, 1956.

[52] Robert E Schapire. The strength of weak learnability. ***Machine learning***, 5(2): 197–227, 1990.

[53] Ana Severiano, João A Carriço, D Ashley Robinson, Mário Ramirez, i Francisco R Pinto. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. ***PLoS One***, 6(5):e19539, 2011.

[54] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, i Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. ***The journal of machine learning research***, 15(1):1929–1958, 2014.

[55] Lukas Steinberger i Hannes Leeb. Conditional predictive inference for high-dimensional stable algorithms. ***arXiv preprint arXiv:1809.01412***, 2018.

[56] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. ***Journal of the royal statistical society: Series B (Methodological)***, 36(2):111–133, 1974.

[57] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, i Rob Fergus. Intriguing properties of neural networks. ***arXiv preprint arXiv:1312.6199***, 2013.

[58] Mingxing Tan i Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. U ***International Conference on Machine Learning***, stranice 6105–6114. PMLR, 2019.

[59] John Tukey. Bias and confidence in not quite large samples. ***Ann. Math. Statist.***, 29:614, 1958.

[60] Vladimir Vapnik. ***The nature of statistical learning theory***. Springer science & business media, 1999.

[61] Vladimir Vovk. Cross-conformal predictors. ***Annals of Mathematics and Artificial Intelligence***, 74(1):9–28, 2015.

[62] Vladimir Vovk, Ilia Nouretdinov, i Alexander Gammerman. Testing exchangeability on-line. U ***Proceedings of the 20th International Conference on Machine Learning (ICML-03)***, stranice 768–775, 2003.

[63] Vladimir Vovk, Alexander Gammerman, i Glenn Shafer. ***Algorithmic learning in a random world***. Springer Science & Business Media, 2005.

[64] Andrew Gordon Wilson. The case for Bayesian deep learning. ***arXiv preprint arXiv:2001.10995***, 2020.

[65] David H Wolpert. Stacked generalization. ***Neural networks***, 5(2):241–259, 1992.

[66] Bianca Zadrozny i Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. U ***Icml***, svezak 1, stranice 609–616. Citeseer, 2001.

[67] Bianca Zadrozny i Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. U ***Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining***, stranice 694–699, 2002.

[68] Gianluca Zeni, Matteo Fontana, i Simone Vantini. Conformal prediction: a unified review of theory and new challenges. ***arXiv preprint arXiv:2005.07972***, 2020.

[69] Zhi-Hua Zhou. Ensemble learning. ***Encyclopedia of biometrics***, 1:270–273, 2009.

11. Sažetak

Modeli dubokog učenja često se koriste u donošenju odluka s visokim ulozima, poput zdravstvene dijagnostike, autonomne vožnje i kaznenog pravosuđa. U tim zadacima svaka pogreška može imati ozbiljne posljedice. Međutim, stopa pogreške može se značajno umanjiti ako su poznati podaci o prediktivnoj pouzdanosti. U ovom seminarskom radu predstavljamo različite metode za mjerenje prediktivne nesigurnosti za modele dubokog učenja. Osim toga, prepoznajemo dvije vrste nesigurnosti prema izvoru nesigurnosti, modelsku (epistemičku) nesigurnost i podatkovnu (aleatornu) nesigurnost. Dajemo pregled metoda za procjenu nesigurnosti predikcije modela, kao i pregled kalibracije modela. Odabrane metode također empirijski ispitujemo nad dubokim modelima za procjenu pamtljivosti slike.