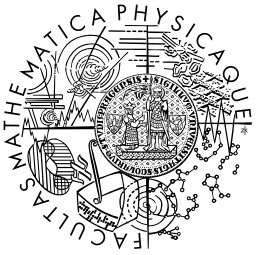


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Daniel Šipoš

Predikce sportovních utkání pomocí neuronových sítí

Katedra softvéru a výuky informatiky

Vedoucí bakalářské práce:	Mgr. David Kuboň
Studijní program:	Informatika
Studijní obor:	Obecná informatika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Poděkování.

Název práce: Predikce sportovních utkání pomocí neuronových sítí

Autor: Daniel Šipoš

Katedra: Katedra softvéru a výuky informatiky

Vedoucí bakalářské práce: Mgr. David Kuboň, Katedra softvéru a výuky informatiky

Abstrakt: Futbal a tenis patria k najpopulárnejším športom na tejto planéte. Platí to hlavne vďaka jednoduchosti pravidiel a nenáročnosti na vybavenie. Obe športy môže hrať prakticky ktokoľvek. Sú ale ľudia, ktorým tieto športy idú lepšie ako ostatným, takzvaní profesionáli. Títo profesionáli potom chodia po rôznych turnajoch, resp. ligách, kde hrajú zápasy, aby sa ukázalo, kto je najlepší. Ľudská chamtivosť na jednej strane a závislosť na hazarde na druhej podporujú vznik rôznych spoločností, stávkových kancelárií, ktoré umožňujú tipovať výsledky týchto zápasov za peniaze. Táto práca sa zameriava na predpovedanie takýchto výsledkov futbalových líg a tenisových turnajov pomocou dvoch mierne odlišných druhov neurónových sietí, porovnanie jednotlivých predpovedných modelov a taktiež porovnanie predvídateľnosti futbalu a tenisu.

Klíčová slova: neurónová sieť rekurentná neurónová sieť športové stávkovanie športové kurzy futbal tenis

Title: Prediction of sports results using neural networks

Author: Daniel Šipoš

Department: Department of Software and Computer Science Education

Supervisor: Mgr. David Kuboň, Department of Software and Computer Science Education

Abstract: Abstract.

Keywords: neural network recurrent neural network football tenis sport betting sport odds

Obsah

Úvod

Šport je súčasťou zábavného priemyslu hlavne pre relatívnu nepredvítateľnosť jeho výsledkov. Stať sa môže v podstate čokoľvek. Vyhrať môže favorit udalosti alebo osoba/tím, od ktorej sa o vôbec neočakávalo. Môže začať pršať alebo na ihrisko vbehnúť exhibicionista s kontroverznou myšlienkou.

Táto nepredvídateľnosť podnietila vznik stávkových kancelárií, ktoré na tieto a na rôzne ďalšie udalosti vypisuje kurzy, ktoré v prípade, že tieto udalosti nastanú, zaručia stávkujúcemu výhru. Tieto stávkové kancelárie ročne zarábajú milióny tým, ako vypisujú kurzy, aby boli lákavé pre bežných ľudí. V podstate sa snažia uhádnuť, s akou pravdepodobnosťou nastane daná udalosť, napríklad predikovať výsledok. Stávkové kancelárie určite používajú na tieto odhady nejaké data, ale pravdepodobnosti daných udalostí zvykne predpovedať odborník, bookmaker. Je možné nájsť nejakú množinu dát, na základe ktorej vieme naučiť počítač predikovať výsledky jednotlivých športových udalostí s určitou presnosťou?

Súvisiace práce

V minulosti boli použité rôzne metódy na predikciu športových výsledkov. V roku 2005 sa o predpoveď 6 rôznych udalostí týkajúcich sa austrálskej kriketovej ligy a AFL, ligy v austrálskom futbale, pokúsil Bailey (Bailey a kol., 2005). Na austrálsky futbal použil data zo zápasov zo 100 sezón odohraných pred rokom 1997 a testoval to na zápasoch od sezóny 1997 do 2003 použitím rôznych modelov lineárnej regresie. Dokázal získať presnosť 66.7%.

V roku 2006 Joseph, Fenton a Neil vyskúšali viaceré druhy strojového učenia na predikciu výsledkov zápasov tímu Tottenham Hotspur F.C. v najvyššej anglickej futbalovej lige, Premier League, v sezónach 1995/1996 a 1996/1997 (Joseph a kol., 2006). To znamená, že pracovali s datasetom o veľkosti 76 zápasov, z ktorej časť delili na trénovacie a časť na testovacie data. Použité metódy zahŕňali expertmi konštruované bayesovské siete, naivný bayesovký klasifikátor, rozhodovacie stromy a k-NN (k nearest neighbours clustering). Použili pri tom 30 príznakov, ale 28 sa viazalo iba na to, či daný hráč nastúpil od začiatku na daný zápas alebo nie, zvyšné dva predstavovali silu súpera a miesto zápasu (či hral predikovaný tím na domácom štadióne alebo nie). V tomto prípade dosiahli bayesovské siete úspešnosť niečo vyše 59%, zvyšné metódy sa pohybovali medzi 30 – 38% pri disjunktných testovacích a trénovacích datach .

V roku 2011 sa dvojica Hucaljuk a Rakipović zameriavala na výber príznakov pri predikcii výsledkov futbalovej Ligy majstrov (Hucaljuk a Rakipović, 2011). Pracovali s datami z 96 zápasov, ktoré manuálne ohodnotili podľa 30 príznakov. Vybrané príznaky predstavovali formu oboch tímov v posledných 6 zápasoch, výsledok posledného vzájomného zápasu týchto dvoch tímov, postavenie v rebríčku, počet zranených hráčov a priemerný počet strelených a inkasovaných gólov. Neskôr zúžili počet príznakov na 20 a na novovzniknutý dataset bolo aplikovaných 6 rôznych metód strojového učenia, naivný bayesovský klasifikátor, bayesovské siete, LogitBoost, k-NN, random forest a neurónové siete. Najvyššia dosiahnutá úspešnosť bola 68%, dosiahli ju neurónové siete.

V roku 2014 použili Igiri a Nwachukwu nástroj, ktorý zvaný Rapid Miner (Igiri a Nwachukwu, 2014). Jeho úlohou bolo predikovať výsledky anglickej Premier League. Použité techniky boli popredná neurónová sieť a lineárna regresia. Neurónová sieť dosiahla úspešnosti 85%, lineárna regresia 93%. Je potrebné dodať, že neurónová sieť predpovedala všetky typy výsledkov (výhra domácich, prehra, remíza), zatiaľ čo regresia predpovedala len zápasy, ktoré sa v konečnom dôsledku skončili výhrou alebo prehrou domáceho celku, takže celková úspešnosť bola o niečo nižšia. Autori dodali, že ak sa predpokladá, že zápas môže skončiť aj remízou, tak neurónové siete mali lepšie výsledky. K predikcii použili rôzne príznaky vrátane kurzov, priemerný počet striel, striel na bránu, rohových kopov, ale aj abstraktnejšie príznaky ako ofenzívna/defenzívna sila mužstva a ohodnotenie sily jednotlivých hráčov a kvality manažéra.

V tom istom roku sa Shin a Gasparyan pokúsili nájsť nové metódy predikcie (Shin a Gasparyan, 2014). Navrhli použiť data z videohry FIFA 2015 na predikciu španielskej La Ligy. Použitie tohto návrhu odôvodnili tým, že vydavatelia videohier v dnešnej dobe pracujú na tom, aby boli ich hry čo možno najreálnejšie. To sa hlavne týka športových hier, kde je dôležité, aby bol každý hráč ohodnotený čo možno najpresnejšie, aby sa to podobalo realite. FIFA 2015 používa rôzne atribúty na ohodnotenie hráča ako napríklad zrýchlenie, strely z diaľky alebo reflexy pre post brankára. Tieto data sa získavajú oveľa jednoduchšie ako z iných zdrojov. Autori vytvorili dva typy modelov: učenie s učiteľom a bez učiteľa. Pri učení s učiteľom vytvorili 2 prístupy, reálny prediktor, ktorý využíval reálne data a virtuálny prediktor, ktorý využíval práve data z popísanej videohry. Obe využívali logistickú regresiu a metódu podporných vektorov. Reálny prediktor dosiahol úspešnosť 75%, virtuálny 80%, čo podľa autorov dokazuje, že data získané z videohier sa dajú používať aj v reálnom svete. Učenie bez učiteľa analyzovalo stratégie tímov podľa typov hráčov, ktorí sú v danom tíme pomocou k-means clusteringu. Zistili, že lepšie tímy zvyknú mať útočnejšie stratégie a slabšie tímy dokážu uhrať lepšie výsledky proti silnejším tímom, ak majú defenzívnejšiu stratégiu.

Taktiež v roku 2014 sa v Iráne skupina výskumníkov pokúsila predpovedať výsledky posledného kola najvyššej iránskej futbalovej ligy IPL zo sezóny 2013/2014 (Arabzad a kol., 2014). Pred posledným kolom nebolo nič rozhodnuté a väčšina z 16 tímov v lige bojovala o lepšie umiestnenie, 5 tímov bojovalo dokonca o titul. Pri vyrovnanosti bodov záleží vo futbale aj na rozdiele v počte strelených a inkasovaných gólov a kvôli vyrovnanosti ligy sa títo výskumníci pokúsili predikovať presné výsledky, teda presný počet gólov strelených domácim i hosťujúcim mužstvom vo všetkých 8 zápasoch. Získali informácie z viac ako 1800 predchádzajúcich zápasov ligy a k predikcii použili rôzne príznaky vrátane počtu získaných bodov v sezóny, počtu získaných bodov v posledných 4 zápasoch a kvality súpera počas posledných 4 zápasov, spolu aj s identifikačnými kódmi jednotlivých tímov a kolom, v ktorom sa daný zápas odohral. Celkovo použili 10 príznakov, na predikciu použili neurónovú sieť. Vo výsledku správne predpovedali víťaza ligy, vzájomné poradie medzi štyrmi z 5 tímov, ktoré bojovali o víťazstvo v lige a presné poradie posledných 5 tímov v tabuľke.

V roku 2016 vyskúšali logistickú regresiu na predikciu výsledkov futbalovej Premier League výskumníci okolo Prasetia (Prasetio a kol., 2016). Stavali na výsledkoch svojich predchodcov a vybrali 4 príznaky, ktoré hrali v predchádzajúcich prácach najväčšiu rolu, konkrétne ohodnotenia pre obranu a útok, pre domácich

aj hostí. Dosiahli úspešnosti v najlepšom prípade 69,5%.

Prínos tejto práce

V tejto práci budeme predikovať futbal a tenis pomocou popredných a rekurentných neurónových sietí. Tenis nie je predikovaný v žiadnej z prác spomínaných v predchádzajúcej sekcii. Futbal je síce predikovaný, ale ani raz štýlom, aký bude prezentovaný v tejto práci.

Väčšina prác má oveľa menšiu trénovaciu vzorku pre siete. Pre túto prácu boli použité informácie z viac ako 5000 futbalových zápasov, z toho trénovacia množina tvorila viac ako 3000 vstupov pre každú ligu. Pre tenis obsahuje dataset viac ako 6200 riadkov a je vytvorený z informácií z viac ako 55000 zápasov.

Ďalšou vecou, ktorou sa táto práca odlišuje od ostatných je to, aké data sú použité. V tejto práci budú použité výhradne výsledky a prostredie zápasov, z ktorých sú následne kalkulované ostatné informácie. Nebudú použité abstraktné data ako sila hráčov alebo tímu ani ohodnotenia žiadnych hráčov ako ani data o počte rohových kopov, žltých kariet, es alebo nevynútených chýb. V tomto ohľade je najpodobnejšia práca od iránskych výskumníkov (Arabzad a kol., 2014), ale aj tam sú značné rozdiely v použití dat.

Žiadna z vyššie spomínaných prác nepoužíva ako jednu z metód vyhodnocovania sietí kurzy stávkových kancelárií. V tejto práci nás hlavne zaujímajú zápasy, v ktorých ani jeden z tímov nie je favoritom z hľadiska kurzov stávkových kancelárií. Vyhodnocovať teda budeme celkovú úspešnosť siete; zisk, ktorý by sme dosiahli pri stávkovaní na všetky zápasy a zisk, ktorý by sme dosiahli stávkovaním výhradne na zápasy, v ktorých nie je jasný favorit.

1. Základné pojmy

1.1 Futbal

Futbal je šport, pri ktorom na hracej ploche, futbalovom ihrisku, proti sebe nastúpia dva jedenásťčlenné tímy s cieľom skórovať čo najviac gólov a inkasovať čo najmenej. Na ihrisku je vždy najviac jedna lopta, hráči ju ovládajú prevažne nohami. Gól nastáva, keď jeden s tímov pošle loptu celým objemom za bránkovú čiaru do priestoru medzi bránkové tyče, teda do súperovej bránky, vrámci pravidiel. Víťazom sa stáva tím, ktorý strelí viac gólov ako súper. Ak je počet vstrelených gólov pre obe zúčastnené strany rovnaký, nastáva remíza (Táborský, 2004).

1.1.1 Futbalové ligy

Väčšina futbalových líg na svete (vrátane tých, s ktorými sa pracuje v tejto práci) funguje aspoň z časti sezóny na systéme, ktorý môžeme nazvať *každý s každým*. To znamená, že každý tím odohrá zápas proti každému tímu v lige. Každá sezóna týchto líg sa najprv delí na kolá a až potom na zápasy. V každom kole odohrá jeden zápas každý tím (s výnimkou jedného tímu, ak liga obsahuje nepárny počet tímov, ten má v danom kole voľno). Za výhru v každom zápase sú 3 body, za remízu 1 bod a za prehru nedostane tím žiaden bod. Sledované ligy fungujú na barážovom systéme, teda najnižšie umiestnené tímy zostupujú do nižšej ligy v hierarchii líg v danej krajine a najvyššie umiestnené ligy postupujú do vyššej ligy v hierarchii.

1.2 Tenis

Tenis je šport tímov súperiacich proti sebe, skladajúcich sa z jedného alebo dvoch ľudí, hrajúcich proti sebe na tenisovom kurte. Zápasy sa delia na dvojhry, teda zápasy dvoch jednočlenných tímov, a štvorhry, zápasy dvoch dvojčlenných tímov. Hlavným cieľom tenisu je použiť tenisovú raketu na zahratie loptičky na súperovu stranu kurtu jedným úderom tak, aby mala súperiaca strana, čo najväčší problém ho vrátiť naspäť (Koromházová, 2008). Ak sa to jednému z tímov nepodarí v súlade s pravidlami, súper získa bod. Tím, ktorý získa 4 body, získa hru. Ak obe tímy získajú po 3 body skôr, ako jeden z nich získa 4, hru získa tím, ktorý získa o 2 body viac ako súper. Tím, ktorý skôr získa 6 hier, získa sadu. Ak nastane stav 5:5, hru získa tím, ktorý získa 7 hier. Zápas sa hrá na dve alebo tri víťazné sady, toto číslo je vždy určené vopred. V tejto práci nás budú hlavne zaujímať dvojhry, teda zápasy jeden proti jednému. (Táborský, 2005).

1.2.1 Turnaje ATP Tour

ATP Tour je tenisový okruh najvyššej celosvetovej úrovne organizovaný asociáciou ATP (Association of Tennis Professionals). Profesionálni hráči sa schá-

dzajú na turnajoch po celom svete. Tieto turnaje sa hrajú vyraďovacím systémom, teda hráč ktorý vyhrá v zápase postúpi do ďalšieho kola turnaja až do finále. Pár najvyšších hráčov postúpi priamo do vyraďovacej časti turnaja, ak sa doň prihlásia, zvyšní hráči ešte musia prejsť kvalifikáciou predtým, ako budú môcť hrať priamo na turnaji. Turnaje spadajúce pod ATP Tour sú turnaje typu ATP Masters 1000, ATP 500 a ATP 250. Tieto turnaje sú nazvané podľa počtu bodov, ktoré si hráč pripíše za výhru. Turnaje Grand Slam spadajú pod ITF (International Tennis Federation), víťaz ale za víťazstvo na týchto turnajoch dostane 2000 bodov. ATP publikuje rebríček profesionálnych hráčov týždenne, hráči sú zoradení zostupne podľa počtu získaných bodov v poslednom roku.

1.3 Porovnanie futbalu a tenisu

Z predchádzajúcich kapitol je zrejmé, že futbal a tenis majú veľa spoločných a veľa rozdielnych vlastností. Futbal je kontaktný šport, teda protihráči sú často vo fyzickom kontakte medzi sebou, zatiaľ čo pri tenise sú protihráči vždy na opačných stranách tenisového kurtu. Rozdielny je aj počet hráčov v jednom tíme, vo futbale je maximálny počet hráčov hrajúcich v jednom momente za jeden tím 11, v tenise to je buď jeden alebo dvaja. Spoločný je napríklad fakt, že sa jedná o loptový šport. Na druhej strane, vo futbale je povolené loptu zasiahnuť ktoroukoľvek časťou tela okrem rúk (s výnimkou brankára), v tenise je zakázané dotknúť sa tenisovej loptičky akoukoľvek časťou tela, loptičku je povolené zahrať len tenisovou raketou. Ďalším rozdielom je hrací čas. Vo futbale má každý zápas fixnú dĺžku (2 polčasy po 45 minút s maximálne 15 minútovou prestávkou medzi nimi), rozhodca na konci každého polčasu nadstaví čas, po ktorý sa nehralo kvôli rôznym prerušeniam v hre (Táborský, 2004). V tenise môže zápas vďaka pravidlám trvať od desiatok minút do niekoľko hodín (Koromházová, 2008).

1.4 Kurzy stávkových na kancelárií

Kurzové stávky sú stávky na akýkoľvek jav, na ktorý vypíše daná stávková kancelária kurz. Kurzy stanovuje bookmaker podľa toho, aká je pravdepodobnosť, že daný jav nastane, kde platí, že čím nižší kurz, tým je vyššia pravdepodobnosť nastania daného javu. Väčšinou sa tieto javy týkajú nejakej športovej udalosti, napríklad nejaké futbalové zápasy alebo automobilové preteky, ale stávkové kancelárie vypisujú kurzy aj na nešportové udalosti, kde medzi tie známejšie patria prezidentské voľby (Bieliková, 2019) alebo ohlásenie mena novorodeného princa v kráľovskej rodine, kde zvyknú byť vypísané kurzy napríklad na pohlavie, meno novorodenca alebo presný dátum narodenia (Mansaray, 2019).

Na javy, na ktoré sú vypísané kurzy môže potom zákazník staviť istú sumu peňazí , vklad, obvykle tak, že vloží tento vklad do stávkovej kancelárie. Ak daný jav nastane, zákazník dostane od tejto stávkovej kancelárie výhru, ktorá predstavuje výsledok vynásobenia daného kurzu vkladom. Ak daný jav nenastane, vklad prepadá v prospech stávkovej kancelárie. Pre príklad si vezmime tipovanie výsledku futbalového zápasu Slovensko - Česká republika, ktorý sa odohral dňa 13.10.2018. Podľa internetového portálu OddsPortal.com bol priemerný vypísaný kurz na tip domáci (v tomto prípade Slovensko) 2,06, na tip hostia (Česká repub-

lika) 3,86 a na tip remíza 3,35. Zápas skončil výhrou hostí, čo znamená, že ak by sme boli stavili 100 korún na tento výsledok, tak by sme si boli odniesli zo stávkovej kancelárie 386 korún (3,86 * 100 = 386), čo predstavuje zárobok 286 korún, pretože 100 korún predstavuje vklad. Ak by sme boli stavili 100 korún na výhru domácich alebo na remízu, tak by sme boli prehrali celý vklad.

Stávkovanie je hazardná hra, obľúbená práve preto, že každý hráč môže vyhrať a vie aj ovplyvniť svoju pravdepodobnosť úspechu tým, že danú udalosť pozná (Netík, 2005).

2. Neurónové siete

Neurónová sieť je založená na orientovanom grafe (ako je možné vidieť na obrázku 2), je teda zložená z uzlov, ktoré sú spojené orientovanými hranami (Kvasnička a kol., 2002). Spojenie uzlu i do uzlu j slúži na propagáciu aktivácie a_i z i do j. Každé takéto spojenie má priradenú váhu $w_{i,j}$, ktorá rozhoduje o sile a znamienku spojenia. Každý uzol má naviac falošný vstup $a_0 = 1$ s priradenou váhou $w_{0,j}$. Všetky uzly si potom vypočítajú váženú hodnotu vstupov, pre uzol j je táto hodnota:

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i$$

Potom sa na výsledok aplikuje aktivačná funkcia g, tým získame výstup z uzlu:

$$a_j = g(in_j) = g\left(\sum_{i=0}^{n} w_{i,j} a_i\right)$$

Aktivačná funkcia g je typicky buď pevná hranica alebo logistická funkcia. V prvom prípade sa uzly volajú perceptrony, v druhom prípade sa niekedy používa pojem sigmoid perceptron. Obe tieto typy nelineárnych aktivačných funkcií zaručujú dôležitú vlastnosť neurónovej siete, a to, že celá sieť uzlov môže reprezentovať aj nelineárnu funkciu.

Takto teda vyzerá matematický model jedného uzlu (v tomto prípade zvaného neurón) v sieti. Spájanie týchto neurónov vytvorí sieť. Existujú dva rozdielne prístupy, akými sa dajú tieto neuróny spojiť do siete. Obe nás zaujímajú pre túto prácu, pretože obe použijeme v praxi a budeme ich porovnávať medzi sebou.

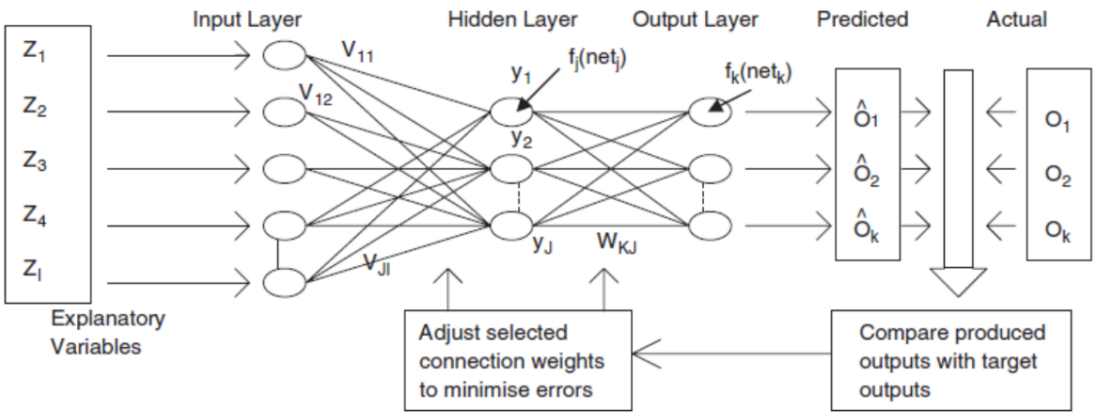
Neurónové siete sú obvykle zoradené do vstiev tak, že každý neurón dostane vstup len z neurónov z predošlej vrstvy. Podľa počtu vrstiev sa siete delia na jednovrstvové a viacvrstvové, kde jednovrstvové siete spájajú vstpné neuróny priamo s výstupnými. Viacvrstvové siete majú medzi vstupom do siete a výstupom z nej ešte jednu alebo viac vrstiev tzv. skrytých (hidden) neurónov (Obrázok 2) (Russell a Norvig, 2016).

2.1 Dopredné neurónové siete

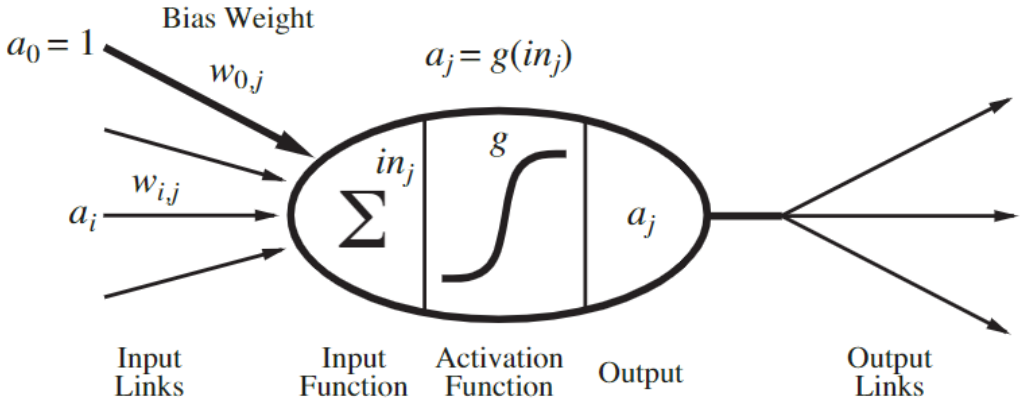
Dopredná neurónová sieť (feed-forward neural network) má spojenia len v jednom smere, takže tvorí orientovaný acyklický graf. Ak si graf topologicky usporiadame, tak každý uzol dostane vstup z niektorých z predchádzajúcich uzlov a predá výstup niektorým z nasledujúcich vrcholov. Dopredná neurónová sieť teda predstavuje funkciu jej momentálneho vstupu, teda neuchováva žiaden stav, ak nepočítame váhy samotné (Russell a Norvig, 2016).

2.2 Rekurentné neurónové siete

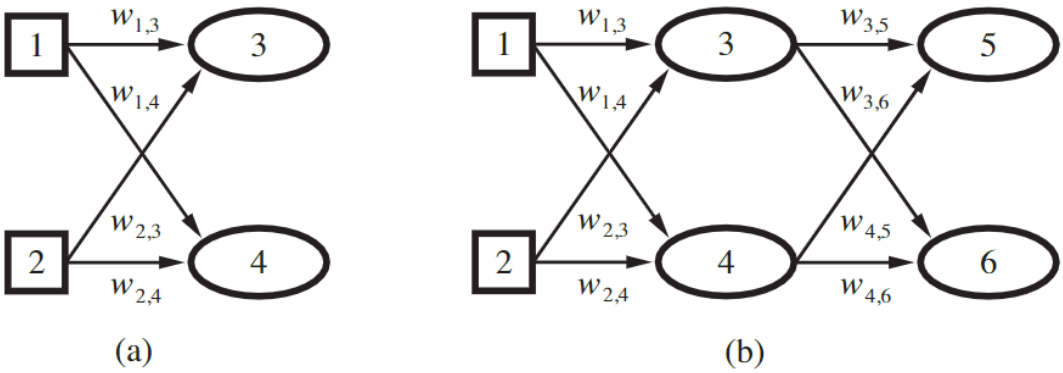
Na druhej strane je rekurentná neurónová sieť (recurrent neural network). Tento typ siete posúva svoj výstup naspäť do svojho vlastného vstupu. Z toho vyplýva, že aktivačné úrovne siete tvoria dynamický systém, ktorý môže dosiahnuť stabilný stav alebo oscilovať či sa dokonca správať chaoticky.



Obr. 2.1: Na obrázku je ukážka jednej z neurónových sietí. Konkrétne to je sieť, ktorú použili v roku 2014 iránski výskumníci pri predikcii výsledkov najvyššej iránskej ligy (Arabzad a kol., 2014).



Obr. 2.2: Takto vyzerá jeden uzol siete (neurón) (Russell a Norvig, 2016).



Obr. 2.3: Ukážka rozdielu medzi jednovrstvou sieťou (a) a viacvrstvovou (b). Obe majú 2 vstupné a 2 výstupné neuróny, viacvrstvová má ešte medzi nimi ďalšie vrstvy skrytých neurónov (v tomto prípade jednu vrstvu s 2 skrytými neurónmi) (Russell a Norvig, 2016).

Výstup siete závisí na vstupe. Pri tomto type siete môže výstup závisieť aj na predchádzajúcich výstupoch, tranzitívne teda aj na predchádzajúcich vstupoch. Z toho vyplýva, že si rekurentná neurónová sieť môže vypracovať krátkodobú pamäť (Russell a Norvig, 2016).

3. Datasety

Data, s ktorými budeme pracovať, sú výhradne len výsledky a konečné stavy jednotlivých zápasov.

3.1 Futbal

Pre futbal data predstavujú pre každú ligu dataset všetkých zápasov odohraných len vrámci ligy za pár posledných sezón. Nebudeme používať žiadne data informujúce o hráčoch, ktorí sú v oficiálnej súpiske na zápas ani data priamo len o základnej zostave na daný zápas. Taktiež vzhľadom na to, že tímy v jednotlivých ligách hrajú zápasy aj mimo ligy, prinajmenšom zápasy v ligovom pohári, tak nebudú použité ani informácie o oddychu pred daným zápasom, teda koľko dní pred zápasom mali zúčastnené tímy voľno.

Dataset pre každú ligu je tabuľka, kde riadky predstavujú jednotlivé zápasy zoradené podľa dátumu, v ktorom bol zápas odohraný, zostupne. Stĺpce sú v poradí:

1. Jednoznačný názov domáceho tímu (nemusí byť celý názov, stačí skrátený, ale jednoznačný a, pokiaľ možno, v celom datasete konzistentný),
2. Jednoznačný názov hosťujúceho tímu,
3. Id zápasu,
4. Ligové kolo, v ktorom sa zápas odohral (0, ak sa nevie),
5. Id domáceho tímu,
6. Id hosťujúceho tímu,
7. Počet gólov strelených domácim tímom v zápase,
8. Počet gólov strelených hosťujúcim tímom v zápase,
9. Dátum zápasu,
10. Sezóna,
11. Kurz na výhru domácich,
12. Kurz na remízu,
13. Kurz na výhru hostí.

Tento dataset potom predáme programu *DataMaker.exe* (TODO!) písanom v jazyku C#, ktorý pretransformuje tieto data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných neurónov je 44, ich význam a poradie je popísané v sekcii Prílohy (A.1).

Skóre je pokus čo najlepšie ohodnotiť formu tímu jedným údajom. V sekcii (Vylepšovanie siete) (TODO!) sa budeme snažiť znížiť počet vstupných neurónov

Brighton	Manchester City	10000	0	7	13	1	4	2019-05-12	2018/2019	15.03	7.73	1.19
Burnley	Arsenal	10001	0	0	5	1	3	2019-05-12	2018/2019	2.54	3.65	2.74
Crystal Palace	Bournemouth	10002	0	2	1	5	3	2019-05-12	2018/2019	1.77	4.31	4.21
Fulham	Newcastle	10003	0	27	10	0	4	2019-05-12	2018/2019	2.45	3.55	2.91
Leicester	Chelsea	10004	0	17	11	0	0	2019-05-12	2018/2019	2.41	3.64	2.91
Liverpool	Wolves	10005	0	6	32	2	0	2019-05-12	2018/2019	1.31	5.83	10.08
Manchester Utd	Cardiff	10006	0	8	26	0	2	2019-05-12	2018/2019	1.30	6.08	9.72
Southampton	Huddersfield	10007	0	12	4	1	1	2019-05-12	2018/2019	1.39	5.10	8.35
Tottenham	Everton	10008	0	16	19	2	2	2019-05-12	2018/2019	1.90	3.75	4.15
Watford	West Ham	10009	0	9	18	1	4	2019-05-12	2018/2019	2.10	3.81	3.41

Obr. 3.1: Ukážka prvých 10 riadkov z tabuľky všetkých zápasov anglickej Premier League ilustrujúcich členenie tabuľky

a ponechať len tie, ktoré sú dôležité. Čím väčší počet vstupných neurónov, tým väčšia je šanca, že sieť sa pokúsi medzi datami nájsť nejakú súvislosť, ktorá tam nie je, čo môže pri testovacích datach vyústiť v nesprávne výsledky (pretrénovanie dat) (TODO! cit).

Posledné 3 stĺpce tohto súboru predstavujú kurzy na dané výsledky. Tieto ale nie sú pri trénovaní siete využívané.

Program tiež vytvorí ďalší súbor, ktorý obsahuje testovacie data, teda data, ktoré sa nevyužívajú pri trénovaní siete, ale len pri vyhodnocovaní výsledkov. Tieto data sú v rovnakom poradí a musia obsahovať kurzy na dané výsledky a aj výsledok zápasu vo forme troch stĺpcov v poradí domáci, remíza, hostia, kde výsledok, ktorý nastal je ohodnotený 1, zvyšné sú 0. Je to potrebné pre vyhodnocovanie, pretože neurónová sieť bude mať 3 výstupné neuróny v rovnakom poradí a predikciu ohodnotí na 1.

Data v jednom súbore predstavujú pár posledných sezón a prvú polovicu sezóny 2018/2019, ktorá predstavuje všetky odohrané zápasy od začiatku sezóny až po odohratie posledného zápasu pred začiatkom kola, ktoré je numericky už v druhej polovici sezóny. Napríklad najvyššia anglická futbalová liga, Premier League, má 38 kôl každú sezónu, do úvahy sa bude brať posledných pár sezón pred sezónou 2018/2019 a všetky zápasy odohrané pred prvým zápasom 20. kola sezóny 2018/2019 (s výnimkou predohrávok, teda zápasov, ktoré boli preložené na dátum pred dátumom, v ktorom daný zápas figuroval v predsezónnom rozpise zápasov). Túto hranicu pre každú predikovanú ligu uvediem ručne do zdrojového kódu programu DataMaker.exe, pretože neviem o nejakom reálnom funkčnom algoritme, ktorý by to vedel s absolútnou istotou určiť a predstavuje to len jednu sezónu pre 5 líg.

Tento súbor je potom predaný programu v0.py (TODO!), ktorý data pripraví, vytvorí neurónovú sieť s danými parametrami (bližšie o presných parametroch v kapitole Príprava siete) a naučí ju dané data, ktoré nakoniec vyhodnotí podľa rôznych kritérií ako dôvera v daný tip alebo kurzovo vyrovnané zápasy, teda zápasy, kde na výhru domácich a vyhru hostí je dostatočne podobný kurz.

3.2 Tenis

Pre tenis budeme používať data pre najlepších 100 hráčov na začiatku každého roka v rebríčku ATP. Data v tomto prípade predstavujú zápasy z turnajov

typu ATP 250, ATP 500, ATP Masters 1000, Grandslam, Finals a záverečných finálových turnajov.

Dataset je tabuľka, každý zápas predstavuje jeden riadok tabuľky, zápasy sú zoradené do turnajov od najskôr odohraných turnajov po tie najbližšie súčasnosti (ak sa obe turnaje začali a končili hrať v rovnaký deň, tak sú v ľubovoľnom poradí, nie je možné, aby poradie zmenilo nejaké data, pretože nie je možné hrať na dvoch turnajoch takéhoto typu zároveň). Zápasy v turnajoch sú zoradené od finále po prvé kolo, teda intuitívne opačne. V tomto prípade na poradí nezáleží, dôležité je, že je v tom systém. Program na spracovanie dat (ATPDataMaker.exe) si tie poradie dat upraví tak, aby mu vyhovovali. Stĺpce tabuľky sú v poradí:

1. Názov turnaja,
2. Počet bodov, ktoré víťaz obrdží za výhru v turnaji (ak to je neznáme, tak je tam nápis N/A)
3. Rok, v ktorom sa turnaj odohral,
4. Povrch kurtov na turnaji (tvrdý *(Hard)*, antukový*(Clay)* alebo trávnatý povrch *(Grass)*),
5. Meno víťaza zápasu,
6. Meno hráča, ktorý zápas prehral,
7. Kolo turnaja, v ktorom sa zápas odohral od najdôležitejšieho (1 značí finále, 2 semifinále, apod.),
8. ID zápasu,
9. ID víťaza (ak ID hráča je NULL, znamená to, že hráč sa doposiaľ ani raz neumiestnil v Top 100 rebríčka ATP),
10. ID porazeného hráča,
11. Počet setov, ktoré v zápase získal víťaz,
12. Počet setov, ktoré v zápase získal porazený hráč,
13. Počet hier, ktoré v zápase získal víťaz v jednotlivých setoch oddelené znakom |,
14. Počet hier, ktoré v zápase získal porazený hráč v jednotlivých setoch oddelené znakom |,
15. Predzápasový kurz na výhru víťaza zápasu,
16. Predzápasový kurz na výhru porazeného hráča v zápase.

ID hráčov sa nachádzajú v ďalšom súbore (*atpranking.csv*), ktorý sa predáva aplikácii na tvorbu vstupných neurónov do neurónovej siete. Tento súbor obsahuje ID jednotlivých hráčov, ich mená a ich poradie v koncoročných rebríčkoch hodnotenia ATP za roky 1999–2018. Poradie berieme len ak sa hráč umiestnil na

Halle	500	2018	Grass	Nikoloz Basilashvili	Elias Ymer	7	69795	20	NULL	2	0 6|6	4|3	2.11	1.71
Halle	500	2018	Grass	Matthias Bachinger	Vasek Pospisil	7	69796	222	69	2	1 5|7|7	7|6|5	3.42	1.31
Halle	500	2018	Grass	Lukas Lacko	Ivo Karlovic	7	69797	122	99	2	0 7|7	6|6	2.93	1.40
Halle	500	2018	Grass	Mikhail Youzhny	Ruben Bemelmans	7	69798	121	NULL	2	0 7|6	6|2	1.63	2.26
London / Queen's Club	500	2018	Grass	Marin Cilic	Novak Djokovic	1	69799	6	0	2	1 5|7|6	7|6|3	2.22	1.69
London / Queen's Club	500	2018	Grass	Marin Cilic	Nick Kyrgios	2	69800	6	34	2	0 7|7	6|6	1.69	2.22
London / Queen's Club	500	2018	Grass	Novak Djokovic	Jeremy Chardy	2	69801	0	39	2	0 7|6	6|4	1.13	6.47
London / Queen's Club	500	2018	Grass	Marin Cilic	Sam Querrey	3	69802	6	50	2	0 7|6	6|2	1.35	3.30
London / Queen's Club	500	2018	Grass	Jeremy Chardy	Frances Tiafoe	3	69803	39	38	2	0 6|6	4|4	1.52	2.58
London / Queen's Club	500	2018	Grass	Novak Djokovic	Adrian Mannarino	3	69804	0	41	2	0 7|6	5|1	1.10	7.76
London / Queen's Club	500	2018	Grass	Nick Kyrgios	Feliciano Lopez	3	69805	34	63	2	0 7|7	6|6	1.50	2.65

Obr. 3.2: Ukážka vybraných pár riadkov z tabuľky zápasov v okruhu ATP ilustrujúcich stĺpce a riadky.

miestach 1–100. Predzápasové kurzy môžu byť prázdne (vyplnené kurzom 0.0), ale len, ak nás na daný zápas nezaujímajú kurzy (zaujímajú nás len pre posledné dva sezóny, prvá je testovacia a druhá vyhodnocovacia sezóna).

Zápasy obsiahnuté v súbore *atpresults.csv* sú len zápasy, v ktorých aspoň jeden hráč bol na konci aspoň raz v daných rokoch na miestach 1–100 v hodnotení ATP. Predikovať sa budú len zápasy medzi takýmito hráčmi, ale kvôli rôznym výpočtom je potrebné mať všetky data o takýchto hráčoch z turnajov, ktoré sú obsiahnuté v súbore.

Tieto datasety sa potom predajú súboru *ATPDataMaker.exe*, ktorý ich pretransformuje na data pre vstupné neuróny neurónových sietí. Všetkých vstupných neurónov je 37, súbor ku každému vstupnému neurónu vydá aj očakávaný výstup (1?, 2?) a pre predikovanú časť dodá aj kurzy stávkových kancelárií na daný výsledok (1B, 2B). Poradie a význam stĺpcov je popísaný v sekcii Prílohy (A.2).

Skóre je pokus výraznejšie ohodnotiť formu hráča ako len počtom výhier a prehier. Pri pokusoch a vylaďovaní siete budeme v sekcii (Vylepšovanie siete) (TODO!) selektovať dané vstupné neuróny podľa rôznych kritérií a vyskúšame tiež aj ako sa bude sieť správať, ak nahradíme všetky stĺpce obsahujúce data o forme rozdielom v skóre. Teoreticky tým ušetríme 6 vstupných neurónov, príliš veľa vstupných neurónov môže viesť k rýchlejšiemu pretrénovaniu siete (TODO! cit), čomu sa budeme snažiť zabrániť selektovaním len tých dôležitých neurónov.

4. Stavba siete

Všetky siete boli napísané v programovacom jazyku *Python* s použitím knižníc *numpy* a *tensorflow*. Na začiatku bola sieť skonštruovaná so všetkými 44 príznakmi, ktoré sme dostali z transformačnej časti práce (Príloha A.1). Sieť obsahovala teda vstup o veľkosti 44 príznakov, 2 skryté vrstvy neurónov (s 25, resp. 15 neurónmi) a troj-neurónový výstup typu softmax, ktorý vyberie najpravdepodobnejšiu možnosť, nastaví daný výstup neurónu na 1 a zvyšné nastaví na 0. Na vylepšovanie siete sme, ako je napísané v jednej z predošlých kapitol (5), použili údaje, ktoré sa napokon budú pri vyhodnocovaní nachádzať medzi testovacími datami.

V prípade futbalu data prišli v dvoch súboroch, ako trénovacie a finálne vyhodnocovacie, takže sme trénovacie data až v tomto programe rozdelili podobne ako budú rozdelené pri vyhodnocovaní výsledkov. Konkrétne, pre futbal data predstavovali 7 celých sezón a prvú polovicu ďalšej sezóny (ako popísané v 3.1), vyhodnocovacie data predstavujú teda druhú polovicu tejto sezóny. Takže sme trénovacie data rozdelili na tri časti, 6 celých sezón a polovicu ďalšej (trénovacie data), druhú polovicu siedmej sezóny (trénovacie data) a zvyšnú prvú polovicu ôsmej sezóny (nepoužité data). V prípade tenisu prišli data už priamo z transformačnej časti v troch súboroch, trénovacie, testovacie a vyhodnocovacie data.

Každá sieť mala svoje nedostatky v celkovej úspešnosti, ale doposiaľ neexistuje efektívne nastavenia neurónových sietí pre každú situáciu (Gandhi, 2018), takže každá sieť sa musela vylepšovať osobitne a manuálne vzhľadom na rozdiely v prístupoch. Cieľom práce nie je porovnať rovnakú architektúru a viac typov neurónových sietí, ale pokúsiť sa získať, čo možno najlepšie výsledky a porovnať potenciály doprednej a rekurentnej neurónovej siete.

5. Dokumentácia

Z programátorského hľadiska je práca rozdelená na tri časti. Prvú časť predstavuje získavanie výsledkov a kurzov jednotlivých zápasov. Druhú časť programu predstavuje transformácia dat na údaje priamo vložiteľné do vstupných neurónov daných neurónových sietí. Poslednú časť tvorí stavba daného typu neurónovej siete pre daný šport.

Transformačná časť v prípade tenisu rozdelí data na 3 časti, trénovacie data, testovacie data (pre optimalizovanie siete) a vyhodnocovacie data. Je teda zaručené, že žiadna sieť neuvidí vyhodnocovacie data vopred pred finálnym vyhodnocovaním. V prípade futbalu data rozdelí na 2 časti, trénovacie a vyhodnocovacie data. Program v poslednej časti si trénovacie data rozdelí podľa potreby. Vyhodnocovacie data sú použité až pre získavanie výsledkov použitých v tejto práci, a teda použili sa až v poslednej fáze.

Údaje, ktoré sa objavia vo výstupe sú celková úspešnosť a celkový zisk, úspešnosť a zisk siete pri vyrovnaných zápasoch (a vyrovnaný zápas považujem zápas, kde kurzy na výhru jedného alebo druhého tímu sa líšia najviac o 1) a úspešnosť a zisk siete pri výhradnom tipovaní zápasov, na ktoré máme istú dôveru (od hodnoty, ktorá je počítaná ako rozdiel dvoch najvyšších čísel, ktoré sieť vydá na výstup, jednoducho povedané, rozdiel najpravdepodobnejšej a druhej najpravdepodobnejšej možnosti výsledku zápasu z hľadiska siete). Táto hodnota dôvery bola tiež vyoptimalizovaná pre každú sieť/program osobitne.

5.1 Futbal

Podmienkou pre futbal je získať všetky zápasy sezóny pre danú ligu. Musia byť všetky, pretože v ďalšej časti sa počíta na základe už odohraných zápasov a jeden zápas by mohol skresliť výsledky. Jednotlivé ligy boli teda vyberané nielen na základe kvality, ale aj na základe toho, že v pár posledných sezónach sa ani raz nestalo, že zápas musel byť z nejakého hľadiska udelený kontumačne jednému z tímov (ako sa napríklad stalo vo francúzskej lige v roku 2007 pre problémy s divákmi (Ligue 1.com)) alebo celá sezóna bola poznačená korupčným škandálom ako v prípade talianskej ligy v sezóne 2006 ((Macek a Vojtaššák, 2006)). Takéto výsledky by nemuseli skresliť stavbu neurónovej siete, ale všeobecne je lepšie, ak sa takýmto situáciám vyhneme.

Údaje o týchto zápasoch sa dajú stiahnuť jednoducho, spustením programu oddscaper.py a zadaním skratky danej ligy pre futbal ako parameter. Skratky sú:

1. ENG - najvyššia anglická liga (Premier League)
2. GER - najvyššia nemecká liga (Bundesliga)
3. SPA - najvyššia španielska liga (La Liga)

Program stiahne všetky výsledky a kurzy pre všetky zápasy všetkých kompletných sezón tej-ktorej ligy zo stránky www.oddsportal.com.

Výstup tohto programu predáme programu *DataMaker.exe* (teda ako prvý parameter programu DataMaker.exe je potrebné predať cestu k súboru, ktorý je

výstupom súboru oddscraper.py, tento súbor sa volá rovnako ako skratka danej ligy s príponou *.csv*), ktorý je písaný v jazyku C# a pretransformuje tieto data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných neurónov je 44. Presné poradie aj popis sa dá nájsť v sekcii Prílohy (Príloha A.1). Tieto údaje boli vybrané špecificky aj s pomocou súvisiacich prác ako údaje, ktoré popisujú stav oboch tímov, ktoré hrajú proti sebe zápas. Bonus predstavujú vstupy označené ako skóre, tieto boli vytvorené mnou ako pokus o jednoduchý a presnejší popis formy pomocou jedného údaju namiesto 10. Ak bude mať teda jeden z týchto neurónov (alebo obe spoločne) úspech, tak bude možné skrátiť počet vstupných neurónov o 10. Pre upresnenie, výstupom súboru je opäť tabuľka formátu csv, názov je zložený zo skratky pre názov danej ligy a slova *input*.

Cestu na dané súbory potom ako prvé dva parametre (v poradí, v akom sú uvedené v úvode kapitoly) predáme programu *ffnnfootball.py* alebo *rnnfootball.py* podľa toho, či chceme, aby dané údaje vyhodnocovala popredná rekurentná neurónová sieť. Výsledky vypíše na štandardný výstup a uloží ich aj do logu, ktorý pozostáva z typu siete, názvu ligy a časovej známky vo formáte *txt*.

Hodnoty dôvery pre ffnnfootball.py a rnnfootball.py sú

TODO!!!

resp.

TODO!!!

5.2 Tenis

Údaje o tenisových zápasoch sú predpripravené v súbore atpresults.csv.

Podmienkou pre tenis je získať všetky zápasy každého turnaja ATP typu 500, 1000 a Grand Slam, kde hraje aspoň jeden hráč z Top 100 rebríčka ATP pre danú sezónu. Dôvodom je fakt, že predikujeme zápasy týchto turnajov medzi hráčmi z Top 100 rebríčka ATP, ale pre týchto hráčov počítame ich momentálnu formu, takže sú pre nás dôležité aj zápasy, ktoré odohrajú proti hráčom, ktorí sa nenachádzajú v Top 100. Vzhľadom na relatívnu kvalitu turnajov ATP 250 a fakt, že množstvo hráčov z Top 100 sa ich pravidelne zúčastňuje. Čo teda logicky znamená, že niekedy nastúpia dvaja takíto hráči aj proti sebe. Takže zoberieme do úvahy aj tieto turnaje (vzájomné zápasy medzi jednotlivými hráčmi na takto ohodnotených turnajoch tiež patria medzi údaje, z ktorých sa stávajú vstupné neuróny (A.2)).

V tenise sa nemôžeme vyhnúť zápasom, ktoré boli nejakým spôsobom udelené jednému z hráčov, či už bez boja alebo po skreči súpera v priebehu zápasu, pretože zranenia sú súčasťou profesionálneho športu. Vo futbale sa to obvykle rieši prestriedaním zraneného hráča, v tenise to, prirodzene, nie je možné. Pre potreby tejto práce máme dve možnosti, buď môžeme tieto zápasy úplne ignorovať alebo ich môžeme započítavať do niektorých oblastí vstupu (ako napríklad forma alebo vzájomné zápasy) a ignorovať inde (predpovedať takéto výsledky je možno nápad pre inú prácu). Pre potreby tejto práce budeme tieto zápasy úplne ignorovať, čo

znamená, že sa nevyskytnú v trénovacích ani testovacích datach. Samozrejme, má to svoje výhody aj nevýhody. Výhodou je, že výsledky budú reálne odzrkadľovať presnosť siete na zápasoch, ktoré sa odohrali a skončili. Predpovedať zranenie nie je cieľom tejto práce. Ďalšou výhodou je spravodlivosť oblastí vstupu ako forma a vzájomné zápasy, pretože sa tam berú len zápasy, ktoré sa dohrali dokonca, takže tieto čísla sa v žiadnom okamihu nenafukujú. Napríklad ak hráč natrafí počas turnaja na dvoch/troch súperov, ktorí sa vzdajú, tak by sa mu vo forme ukázali tieto víťazstvá, aj keď to neboli plnohodnotné výhry. Nevýhodou je, že výsledky nemusia ukazovať reálne výsledky v praxi (pred zápasom nevieme určiť, či sa hráč zraní, ale sieť aj tak vydá svoju predpoveď, aj keď nebola na tieto údaje trénovaná).

Pre tabuľku *atpresults.csv* je postup podobný ako pre túto časť futbalovej predikcie. Túto tabuľku je potrebné predať programu *ATPDataMaker.exe* (opäť ako prvý parameter je potrebené predať cestu k tejto tabuľke). Program je opäť písaný v jazyku C# a opäť pretransformuje data na vstupné neuróny pre neurónovú sieť. Všetkých vstupných údajov (počet stĺpcov tabuľky) je 37. Ich presné poradie a popis sa dá nájsť v sekcii Prílohy (Príloha A2). Z popisu datasetu je vidieť, že hráči v zápasoch sú zoradení tak, že najprv je napísaný víťaz a po ňom porazený. To by nám očividne zamiešalo výsledkami a ak by to sieť zistila, tak by okamžite vypisovala úspešnosť 100 %. Presne z tohto dôvodu robí program *ATPDataMaker.exe* aj randomizovanú výmenu poradia hráčov a v ďalšom priebehu sú hráči rozlišovaní ako hráč 1 a hráč 2.

Cestu na dané súbory potom ako prvé tri parametre (v poradí, v akom sú uvedené v úvode sekcie) predáme programu *ffnnatp.py* alebo *rnnatp.py* podľa toho, či chceme, aby dané údaje vyhodnocovala popredná rekurentná neurónová sieť. Výsledky vypíše na štandardný výstup a uloží ich aj do logu, ktorý je vo formáte *txt* a ktorého názov pozostáva z typu siete, slova *atp* a časovej známky.

Hodnoty dôvery pre ffnnatp.py a rnnatp.py sú

TODO!!

resp.

TODO!!

Záver

Zoznam použitej literatúry

Arabzad, S. M., Tayebi Araghi, M., Sadi-Nezhad, S. a Ghofrani, N. (2014). Football match results prediction using artificial neural networks; the case of Iran Pro League. *Journal of Applied Research on Industrial Engineering*, **1**(3), 159–179.

Bailey, M. J. a kol. (2005). *Predicting sporting outcomes: A statistical approach*. PhD thesis, Faculty of Life and Social Sciences, Swinburne University of Technology.

Bieliková, J. (2019). Stávkovanie na voľby: Bookmakeri najviac veria Šefčovičovi a Čaputovej. URL `https://plus7dni.pluska.sk/domov/stavkovanie-volby-bookmakeri-najviac-veria-sefcovicovi-caputovej`. [cit. 2019-07-09].

Gandhi, R. (2018). Improving the Performance of a Neural Network. URL `https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d`. [cit. 2019-07-09].

Hucaljuk, J. a Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1623–1627. IEEE.

Igiri, C. P. a Nwachukwu, E. O. (2014). An improved prediction system for football a match result. *IOSR Journal of Engineering (IOSRJEN)*, **4**(12), 12–20.

Joseph, A., Fenton, N. E. a Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, **19**(7), 544–553.

Koromházová, V. (2008). *Jak dokonale zvládnout tenis*. Grada Publishing as. ISBN 978-80-247-2316-7.

Kvasnička, V., Beňušková, Ľ., Pospíchal, J., Farkaš, I., Tiňo, P. a Kráľ, A. (2002). Úvod do teórie neurónových sietí. URL `http://ics.upjs.sk/~novotnyr/home/skola/neuronove_siete/nn_kvasnicka/Uvod%20do%20NS.pdf`. [cit. 2019-05-20].

Ligue 1.com (2017). Bastia forfeit abandoned OL clash. URL `https://www.ligue1.com/ligue1/article/bastia-forfeit-abandoned-ol-clash.htm`. [cit. 2019-07-10].

Macek, Ľ. a Vojtaššák, P. (2006). Korupčný škandál položil Juventus Turín. URL `https://hnonline.sk/sport/115486-korupcny-skandal-polozil-juventus-turin`. [cit. 2019-07-10].

Mansaray, J. (2019). Any day now - odds on for ímminentťroyal baby birth. URL `https://www.reuters.com/article/us-britain-royals-baby-betting-idUSKCN1S7418`. [cit. 2019-05-09].

Netík, M. (2005). Jak sázet s pomocí internetu (1.). URL `https://www.lupa.cz/clanky/jak-sazet-s-pomoci-internetu-1/`. [cit. 2019-04-28].

Prasetio, D. a kol. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5. IEEE.

Russell, S. J. a Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited,.

Shin, J. a Gasparyan, R. (2014). A novel way to soccer match prediction. *Stanford University: Department of Computer Science*.

Táborský, F. (2004). *Sportovní hry*. Grada Publishing as. ISBN 80-247-0875-2.

Táborský, F. (2005). *Sportovní hry 2: základní pravidla, organizace, historie*. Grada Publishing as. ISBN 80-247-1330-6.

Zoznam obrázkov

Zoznam tabuliek

Seznam použitých zkratek

A. Prílohy

A.1 Vstup neurónovej siete pre futbal

Výstupný súbor z transformačnej časti predikcie pre futbal je tabuľka formátu *csv*. Riadky predstavujú jednotlivé predikované zápasy a stĺpce sú nasledovné:

1. htW - home team wins - doterajší počet výher domáceho tímu v práve evaluovanej sezóne,
2. htD - home team draws - doterajší počet remíz domáceho tímu v práve evaluovanej sezóne,
3. htL - home team loses - doterajší počet prehier domáceho tímu v práve evaluovanej sezóne,
4. htGFpG - home team goals for per game - doterajší priemerný počet strelených gólov na zápas domáceho tímu v práve evaluovanej sezóne,
5. htGApG - home team goals against per game - doterajší priemerný počet inkasovaných gólov na zápas domáceho tímu v práve evaluovanej sezóne,
6. atW - away team wins - doterajší počet výher hosťujúceho tímu v práve evaluovanej sezóne,
7. atD - away team draws - doterajší počet remíz hosťujúceho tímu v práve evaluovanej sezóne,
8. atL - away team loses - doterajší počet prehier hosťujúceho tímu v práve evaluovanej sezóne,
9. atGFpG - away team goals for per game - doterajší priemerný počet strelených gólov na zápas hosťujúceho tímu v práve evaluovanej sezóne,
10. atGApG - away team goals against per game - doterajší priemerný počet inkasovaných gólov na zápas hosťujúceho tímu v práve evaluovanej sezóne,
11. htHW - home team home wins - doterajší počet výher domáceho tímu na domácom ihrisku v práve evaluovanej sezóne,
12. htHD - home team home draws - doterajší počet remíz domáceho tímu na domácom ihrisku v práve evaluovanej sezóne,
13. htHL - home team home loses - doterajší počet prehier domáceho tímu na domácom ihrisku v práve evaluovanej sezóne,
14. htHGFpG - home team home goals for per game - doterajší priemerný počet strelených gólov na zápas domáceho tímu na domácom ihrisku v práve evaluovanej sezóne,
15. htHGApG - home team home goals against per game - doterajší priemerný počet inkasovaných gólov na zápas domáceho tímu na domácom ihrisku v práve evaluovanej sezóne,

16. atAW - away team away wins - doterajší počet výher hosťujúceho tímu v role hostí v práve evaluovanej sezóne,

17. atAD - away team away draws - doterajší počet remíz hosťujúceho tímu v role hostí v práve evaluovanej sezóne,

18. atAL - away team away loses - doterajší počet prehier hosťujúceho tímu v role hostí v práve evaluovanej sezóne,

19. atAGFpG - away team away goals for per game - doterajší priemerný počet strelených gólov na zápas hosťujúceho tímu v role hostí v práve evaluovanej sezóne,

20. atAGApG - away team away goals against per game - doterajší priemerný počet inkasovaných gólov na zápas hosťujúceho tímu v role hostí v práve evaluovanej sezóne,

21. hFW - home form wins - počet výher domáceho tímu v posledných 5 zápasoch,

22. hFD - home form draws - počet remíz domáceho tímu v posledných 5 zápasoch,

23. hFL - home form loses - počet prehier domáceho tímu v posledných 5 zápasoch,

24. hFGF - home form goals for - priemerný počet strelených gólov domáceho tímu v posledných 5 zápasoch,

25. hFGA - home form goals against - priemerný počet inkasovaných gólov domáceho tímu v posledných 5 zápasoch,

26. aFW - away form wins - počet výher hosťujúceho tímu v posledných 5 zápasoch,

27. aFD - away form draws - počet remíz hosťujúceho tímu v posledných 5 zápasoch,

28. aFL - away form loses - počet prehier hosťujúceho tímu v posledných 5 zápasoch,

29. aFGF - away form goals for - priemerný počet strelených gólov hosťujúceho tímu v posledných 5 zápasoch,

30. aFGA - away form goals against - priemerný počet inkasovaných gólov hosťujúceho tímu v posledných 5 zápasoch,

31. MW - mutual wins - počet výhier domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu (alebo všetkých vzájomných zápasoch od prvej sezóny v datasete),

32. MD - mutual draws - počet remíz domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu,

33. ML - mutual loses - počet prehier domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu,

34. MGF - mutual goals for - priemerný počet strelených gólov domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu,

35. MGA - mutual goals against - priemerný počet inkasovaných gólov domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu,

36. MhW - mutual home wins - počet výhier domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu hraných na domácom ihrisku,

37. MhD - mutual home draws - počet remíz domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu hraných na domácom ihrisku,

38. MhL - mutual home loses - počet prehier domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu hraných na domácom ihrisku,

39. MhGF - mutual home goals for - priemerný počet strelených gólov domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu hraných na domácom ihrisku,

40. MhGA - mutual home goals against - priemerný počet inkasovaných gólov domáceho tímu v posledných 5 vzájomných zápasoch proti hosťujúcemu tímu hraných na domácom ihrisku,

41. htLTS - home team long-time strength - dlhodobá sila domáceho mužstva (počítaná ako priemerný počet získaných bodov vo všetkých doterajších sezónach od prvej sezóny v datasete),

42. atLTS - away team long-time strength - dlhodobá sila hosťujúceho mužstva,

43. dFS - difference form score - rozdiel v skóre formy medzi domácim a hosťujúcim tímom,

44. dFCS - difference form current score - rozdiel v momentálnom skóre formy medzi domácim a hosťujúcim tímom,

45. W - win - hodnota určujúca konečný výsledok zápasu; 1, ak skončil víťazstvom domáceho tímu, 0 inak,

46. D - draw - hodnota určujúca konečný výsledok zápasu; 1, ak skončil remízou, 0 inak,

47. L - lose - hodnota určujúca konečný výsledok zápasu; 1, ak skončil prehrou domáceho tímu, 0 inak.

Skóre formy oboch tímov je vypočítané ako súčet cez posledných 5 zápasov počet bodov súpera tímu v momente ukončenia zápasu vynásobený počtom bodov získaných z daného zápasu. To by malo ukázať silu výsledku a dať dôraz na neskôr odohrané zápasy. Momentálne skóre formy funguje podobne s výnimkou toho, že je prepočítavané pred evaluovaným zápasom a nie v momente ukončenia zápasu, čo by malo viac ukázať silu výsledku s odstupom času.

Súbory používané na testovanie a vyhodnocovanie siete obsahujú ešte 3 stĺpce pre každý riadok, v poradí kurz na výhru domáceho mužstva, kurz na remízu a kurz na výhru hosťujúceho mužstva.

A.2 Vstup neurónovej siete pre tenis

Výstupný súbor z transformačnej časti predikcie pre tenis je tabuľka vo formáte *csv*. Riadky predstavujú jednotlivé predikované zápasy a stĺpce sú nasledovné:

1. 1W - player 1 wins - počet výher hráča 1 v práve vyhodnocovanej sezóne,
2. 1L - player 1 loses - počet prehier hráča 1 v práve vyhodnocovanej sezóne,
3. 1GDpS - player 1 game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 v práve vyhodnocovanej sezóne,
4. 2W - player 2 wins - počet výher hráča 2 v práve vyhodnocovanej sezóne,
5. 2L - player 2 loses - počet prehier hráča 2 v práve vyhodnocovanej sezóne,
6. 2GDpS - player 2 game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 2 v práve vyhodnocovanej sezóne,
7. 1FW - player 1 form wins - počet výher hráča 1 v jeho posledných 10 zápasoch,
8. 1FL - player 1 form loses - počet prehier hráča 1 v jeho posledných 10 zápasoch,
9. 1FGDpS - player 1 form game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 v jeho posledných 10 zápasoch,
10. 2FW - player 2 form wins - počet výher hráča 2 v jeho posledných 10 zápasoch,
11. 2FL - player 2 form loses - počet prehier hráča 2 v jeho posledných 10 zápasoch,
12. 2FGDpS - player 2 form game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 2 v jeho posledných 10 zápasoch,
13. 1SW - player 1 surface wins - počet výher hráča 1 na danom povrchu v práve vyhodnocovanej sezóne,
14. 1SL - player 1 surface loses - počet prehier hráča 1 na danom povrchu v práve vyhodnocovanej sezóne,
15. 1SGDpS - player 1 surface game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 na danom povrchu v práve vyhodnocovanej sezóne,

16. 2SW - player 2 surface wins - počet výher hráča 2 na danom povrchu v práve vyhodnocovanej sezóne,

17. 2SL - player 2 surface loses - počet prehier hráča 2 na danom povrchu v práve vyhodnocovanej sezóne,

18. 2SGDpS - player 2 surface game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 2 na danom povrchu v práve vyhodnocovanej sezóne,

19. 1SFW - player 1 surface form wins - počet výher hráča 1 v jeho posledných 10 zápasoch odohraných na danom povrchu,

20. 1SFL - player 1 surface form loses - počet prehier hráča 1 v jeho posledných 10 zápasoch odohraných na danom povrchu,

21. 1SFGDpS - player 1 surface form game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 na danom povrchu v jeho posledných 10 zápasoch,

22. 2SFW - player 2 surface form wins - počet výher hráča 2 v jeho posledných 10 zápasoch odohraných na danom povrchu,

23. 2SFL - player 2 surface form loses - počet prehier hráča 2 v jeho posledných 10 zápasoch odohraných na danom povrchu,

24. 2SFGDpS - player 2 surface form game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 na danom povrchu v jeho posledných 10 zápasoch,

25. 1MW - player 1 mutual wins - počet výher hráča 1 vo vzájomných zápasoch*,

26. 1ML - player 1 mutual loses - počet prehier hráča 1 vo vzájomných zápasoch*,

27. 1MGDpS - player 1 mutual game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 vo vzájomných zápasoch*,

28. 1MSW - player 1 mutual surface wins - počet výher hráča 1 vo vzájomných zápasoch* odohraných na danom povrchu,

29. 1MSL - player 1 mutual surface loses - počet prehier hráča 1 vo vzájomných zápasoch* odohraných na danom povrchu,

30. 1MSGDpS - player 1 mutual surface game difference per set - priemerný rozdiel v počte vyhraných a prehraných hier za set hráča 1 vo vzájomných zápasoch* odohraných na danom povrchu,

31. 1R - player 1 rank - umiestnenie hráča 1 v poslednom koncoročnom rebríčku ATP

32. 2R - player 2 rank - umiestnenie hráča 2 v poslednom koncoročnom rebríčku ATP

33. H - hard - kategorická hodnota určujúca povrch, na ktorom sa odohral zápas; 1, ak sa hral na tvrdom povrchu, 0 inak

34. C - clay - kategorická hodnota určujúca povrch, na ktorom sa odohral zápas; 1, ak sa hral na antuke, 0 inak

35. G - grass - kategorická hodnota určujúca povrch, na ktorom sa odohral zápas; 1, ak sa hral na trávnatom povrchu, 0 inak

36. dSc - difference in score - rozdiel v skóre oboch hráčov**,

37. dSSc - difference in surface score - rozdiel v skóre oboch hráčov na danom povrchu**,

38. 1? - did player 1 win - hodnota určujúca víťaza; 1, ak vyhral hráč 1, 0 inak

39. 2? - did player 2 win - hodnota určujúca víťaza; 1, ak vyhral hráč 2, 0 inak

* - vzájomné zápasy sú prepočítavané len pre sezóny, odkiaľ sú data; tie sú od sezóny 2003, najstaršie trénovacie data obsahujú sezónu 2012, takže to teoreticky môže ovplyvniť len zápasy medzi hráčmi, ktorí hrajú profesionálne viac ako 9 rokov a aspoň jeden z nich sa už vtedy umiestnil v Top 100 rebríčka ATP a v práve evaluovanej sezóne sa tam umiestnili obaja; to sa nestávalo často, efekt to malo na minimum vyhodnocovaní a teória hovorí, že posledné vzájomné zápasy sú aj ak dôležitejšie, takže teoreticky nevadí, že je konečná sezóna, ak je ďaleko od práve vyhodnocovanej.
** - skóre je pokus ohodnotiť silu víťazstva, berie do úvahy formu, teda posledných 10 zápasov a počíta sa ako $(150 - rank) \cdot point$, kde rank je poradie súpera v poslednom koncoročnom rebríčku ATP a point je nastavené na 1, ak hráč vyhral, a na 0, ak prehral. Ak súper nebol v Top 100 rebríčka ATP na konci predchádzajúceho roka, tak za jeho rank je dosadené číslo 130. To je len preto, lebo teoreticky má dané víťazstvo hodnotu, musí byť teda nejak ohodnotené lepšie ako ľubovoľná prehra, ktorá je ohodnotená hodnotou 0.

Súbory používané na testovanie a vyhodnocovanie siete obsahujú ešte 2 stĺpce pre každý riadok, v poradí kurz na výhru hráča 1 a kurz na výhru hráča 2.