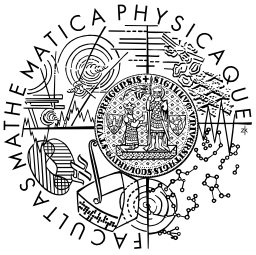


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Daniel Šipoš

Predikcia športových zápasov pomocou neurónových sietí

Katedra softwaru 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

Ďakujem.

Název práce: Predikcia športových zápasov pomocou neurónových sietí

Autor: Daniel Šipoš

Katedra: Katedra softwaru a výuky informatiky

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

Abstrakt: Práca sa zameriava na vytvorenie modelov dvoch odlišných druhov neurónových sietí slúžiacich na predpovedanie výsledkov vybraných futbalových a tenisových zápasov a porovnanie týchto modelov z hľadiska percentuálnej úspešnosti a potenciálneho zisku, ak by sme na dané zápasy uzatvárali stávku v priemernej medzinárodnej stávkovej kancelárii. Porovnávané druhy neurónových sietí sú dopredná neurónová sieť a rekurentná neurónová sieť. Predikované futbalové zápasy sú tvorené ligovými zápasmi z troch európskych ligách. Špecifikom je sledovanie úspešnosti na zápasoch, v ktorých ani jeden z tímov nie je jasným favoritom podľa stávkových kancelárií.

Klíčová slova: neurónová sieť, športové stávky, 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: This thesis focuses on creating models of two different types of neural network used for predicting results of selected football and tennis matches and comparing these two models in terms of their accuracy and potential profit, if we had bet on those games in an average betting agency. Compared types of neural networks are feed-forward and recurrent neural network. Predicted football matches consist of league matches of three European leagues. Specific feature of this thesis is tracking accuracy in predicting matches, where neither team is clear favorite to win according to the bookmakers.

Keywords: neural network, sports betting, football, tenis

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 to 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 vypisujú kurzy, ktoré v prípade, že tieto udalosti nastanú, zaručia stávkujúcemu výhru. Ich ziskovosť je založená na vypisovaní kurzov tak, aby boli lákavé pre bežných ľudí. V podstate sa snažia uhádnuť, s akou pravdepodobnosťou nastane daná udalosť, napríklad predikovať výsledok. Stávkové kancelárie používajú na tieto odhady nejaké dáta, 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 dáta 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 dátasetom s veľkosťou 76 zápasov, z ktorého časť delili na trénovacie a časť na testovacie dáta. 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 v rozmedzí 30 – 38 % pri disjunktných testovacích a trénovacích dátach.

V roku 2011 sa dvojica Hucaljuk a Rakipović zameriavala na výber príznakov pri predikcii výsledkov futbalovej Ligy majstrov (Hucaljuk a Rakipović, 2011). Pracovali s dátami 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 dátaset bolo aplikovaných 6 rôznych metód strojového učenia, menovite: naivný bayesovský klasifikátor, bayesovské siete, LogitBoost, k-NN, random forest a neurónové siete. Najvyššia dosiahnutá úspešnosť bola 68 %, ktorú dosiahli použitím neurónových sietí.

V roku 2014 použila dvojica Igiri a Nwachukwu nástroj 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 úspešnost 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 dáta 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 týka hlavne športových hier, kde je dôležité, aby sa hodnotenie hráča čo najviac približovalo realite. FIFA 2015 používa rôzne atribúty na ohodnotenie hráča, ako napríklad zrýchlenie, strely z diaľky alebo reflexy pre post brankára. Tieto dáta 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 dáta a virtuálny prediktor, ktorý využíval práve dáta z popísanej videohry. Obe využívali logistickú regresiu a metódu podporných vektorov. Reálny prediktor dosiahol úspešnosť 75 %, virtuálny 80 %, čo podľa autorov dokazuje, že dáta 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 rovnosti bodov záleží vo futbale aj na rozdiele v počte strelených a inkasovaných gólov. 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 počas 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 4 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 z tímu 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 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 dátaset viac ako 6200 riadkov a je vytvorený z informácií z viac ako 55 000 zápasov.

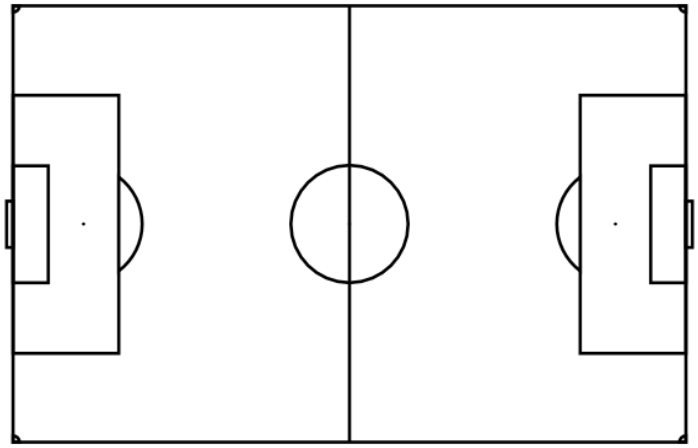
Ďalší atribut, ktorým sa táto práca odlišuje od ostatných predstavuje použité dáta. V tejto práci budú použité výhradne výsledky a prostredie zápasov, z ktorých sú následne kalkulované ostatné informácie. Nebudú použité abstraktné dáta ako sila hráčov alebo tímu, ani ohodnotenia žiadnych hráčov, ani dáta 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 dát.

Žiadna z vyššie spomínaných prác nepoužíva ako jednu z metód vyhodnocovania sietí kurzy stávkových kancelárií. V tejto práci nás zaujímajú hlavne zápasy, v ktorých ani jeden z tímov nie je favoritom z hľadiska kurzov stávkových kancelárií. Vyhodnocovať teda budeme celkovú úspešnosť siete; úspešnosť siete v zápasoch bez jasného favorita a zisk, ktorý by sme dosiahli stávkovaním výhradne na zápasy, v ktorých nie je jasný favorit. Zápasy bez jasného favorita definujeme podľa stávkových kancelárií. Sú to zápasy, pri ktorých je rozdiel kurzov na výhru jedného a druhého tímu menší alebo rovný 1 v prípade futbalu a v prípade tenisu to sú zápasy, pri ktorých je rozdiel kurzov na výhru jeného a druhého hráča menší alebo rovný 0,6.

1. Základné pojmy

1.1 Futbal

Futbal je šport, pri ktorom na hracej ploche, futbalovom ihrisku (na obrázku 1.1), 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 z 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).



Obr. 1.1: Vzhľad futbalového ihriska, pre medzinárodné zápasy musí mať dlhšia strana 115 – 120 m, kratšia 64 – 95 m

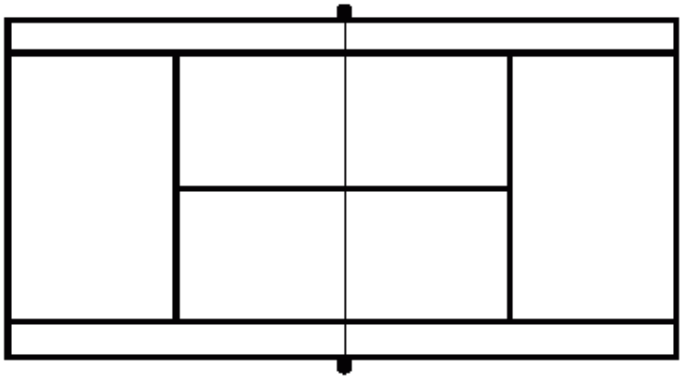
1.1.1 Futbalové ligy

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 z 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 tímy 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 (obrázok 1.2) 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 oba 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 zaujímať hlavne dvojhry, teda zápasy jeden proti jednému. (Táborský, 2005).



Obr. 1.2: Vzhľad tenisového kurtu, kurt je dlhý 23,77 m, široký 10,97 m, dodatočný prázdny priestor okolo kurtu je vyhradený, aby hráči mali možnosť dosiahnuť na loptičky, ktoré sú v hre, ale nachádzajú sa mimo kurtu. Sieť je vysoká 1.07 m na krajoch kurtu, 0,91 m v strede.

1.2.1 Turnaje ATP Tour

ATP Tour je tenisový okruh najvyššej celosvetovej úrovne organizovaný asociáciou ATP (Association of Tennis Professionals). Profesionálni hráči sa schádzajú 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 pred tým, 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

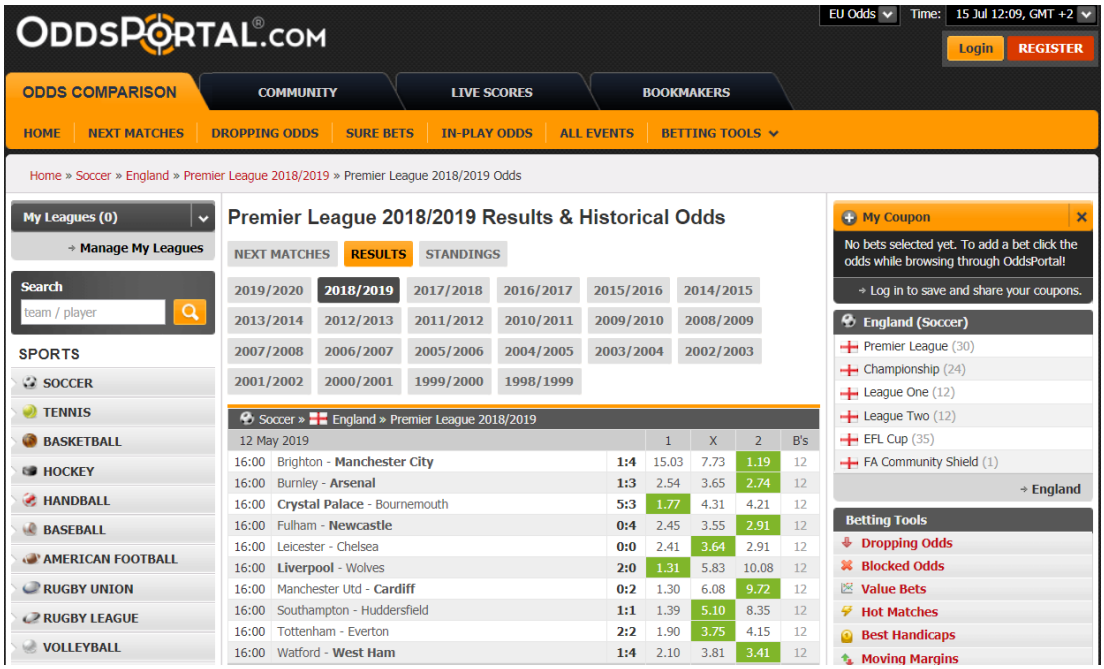
vo fyzickom kontakte medzi sebou, zatiaľ čo pri tenise sú protihráči vždy na opačných stranách tenisového kurtu. Rozdielny je aj počet hráčov v jednom tíme, vo futbale je maximálny počet hráčov hrajúcich v jednom momente za jeden tím 11, v tenise to je buď jeden alebo dvaja. Spoločný je napríklad fakt, 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 niekoľkých hodín (Koromházová, 2008).

1.4 Kurzy stávkových 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 futbalových zápasov alebo automobilových pretekov. Stávkové kancelárie ale vypisujú kurzy aj na nešportové udalosti, kde medzi tie známejšie patria prezidentské voľby (Bieliková, 2019) alebo ohlásenie mena novorodeného dieťaťa 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 republika) 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).



Obr. 1.3: Takto vyzerá stránka OddsPortal.com, z ktorej som získaval dáta pre potreby tejto práce. Znázornené je posledné kolo predikovanej časti anglickej *Premier League*.

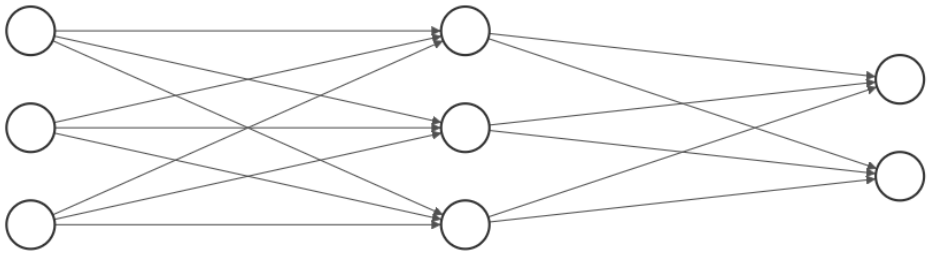
2. Neurónové siete

Neurónová sieť je založená na orientovanom grafe (ako je možné vidieť na obrázku 2.1). 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:

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

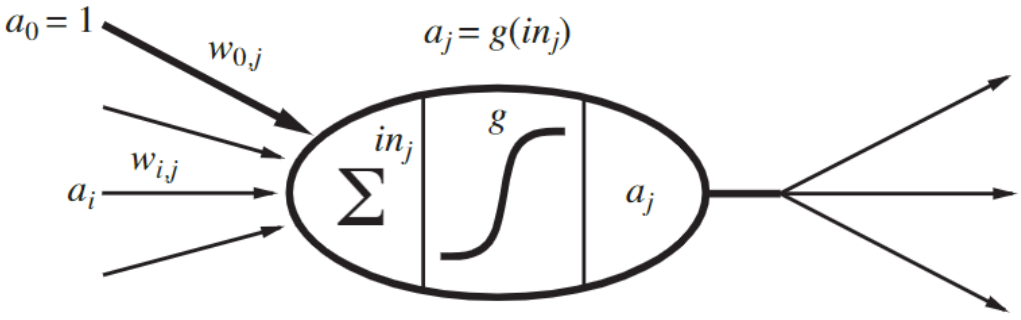
Potom sa na výsledok aplikuje aktivačná funkcia g, tým získame výstup z uzlu:

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



Obr. 2.1: Ukážka jednej z neurónových sietí

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.



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

Takto teda vyzerá matematický model jedného uzlu (v tomto prípade zvaného neurón) v sieti. Spájanie týchto neurónov vytvorí sieť. Existujú rôzne spôsoby, akými sa dajú jednotlivé neuróny spojiť do siete. Dva z nich sú dôležité pre túto prácu, pretože obe použijeme a porovnáme medzi sebou. Tieto dva prístupy sú dopredná a rekurentná neurónová sieť.

2.1 Dopredné neurónové siete

Dopredná neurónová sieť (feed-forward neural network) má spojenia len v jednom smere, takže tvorí orientovaný acyklický graf (obrázok 2.1 zobrazuje práve tento typ siete). Ak si graf topologicky usporiadame, tak každý uzol dostane vstup z niektorých z predchádzajúcich uzlov a predá výstup niektorým z nasledujúcich uzlov. Dopredná neurónová sieť teda predstavuje funkciu jej momentálneho vstupu, teda neuchováva žiaden stav, ak nepočítame váhy samotné (Russell a Norvig, 2016).

Tieto 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 .

2.1.1 Jednovrstvové siete

Jednovrstvové siete spájajú vstupné neuróny priamo s výstupnými. Tieto siete sú ale obmedzené a nevedia sa naučiť funkcie, ktoré nie sú lineárne separabilné, platí to dokonca aj pre niektoré jednoduché funkcie ako napríklad XOR (Russell a Norvig, 2016).

V Euklidovskej geometrii je lineárna separabilita vlastnosť dvoch množín bodov v priestore, ktoré sa dajú v tomto priestore presne oddeliť nadrovinou. Najjednoduchšie sa to dá predstaviť v dvojrozmernom priestore napríklad na boolovskej funkcii OR. Bodu (0,0) v súradnicovom systéme (x,y) priradí funkcia hodnotu 0, bodom (1,0), (0,1) a (1,1) priradí hodnotu 1. Ak body rozdelíme do množín podľa priradenej hodnoty, tak tieto dva množiny sa dajú presne oddeliť priamkou, napríklad $y = -x + 1/2$. Je zjavné, že množiny bodov získané z funkcie XOR (tabuľka 2.1) sa takto rozdeliť nedajú, tieto množiny teda nie sú lineárne separabilné.

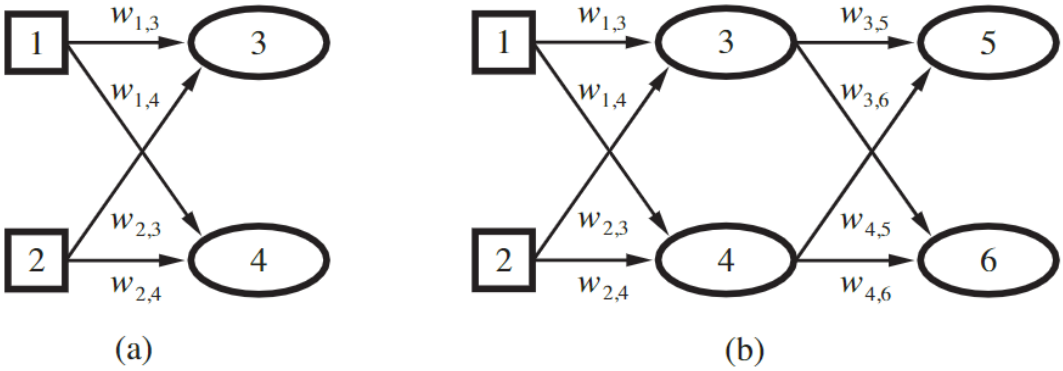
x	y	x XOR y
0	0	0
0	1	1
1	0	1
1	1	0

Tabuľka 2.1: Tabuľka funkcie XOR

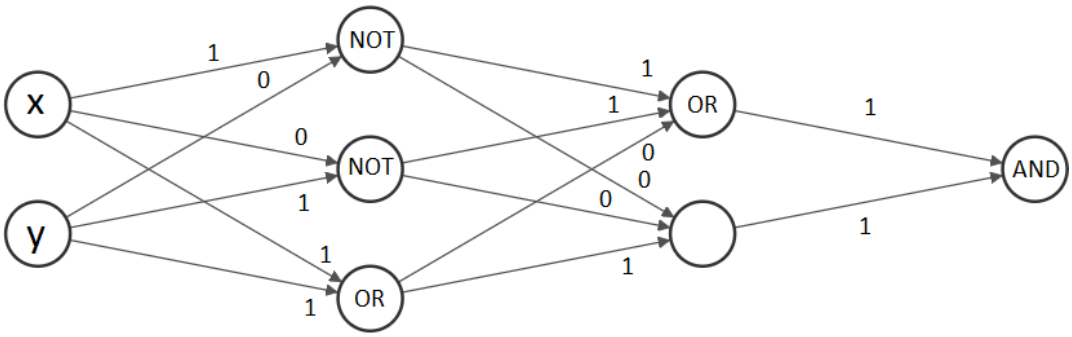
2.1.2 Viacvrstvové siete

Viacvrstvové siete majú medzi vstupom do siete a výstupom z nej ešte jednu alebo viac vrstiev tzv. skrytých (hidden) neurónov (Obrázok 2.3). Waren McCulloch a Walter Pitts (McCulloch a Pitts, 1943) vo svojom článku dokázali, že jeden neurón v sieti vie reprezentovať základné boolovské funkcie AND, OR a NOT a vyslovili, že každá dodatočná funkcionalita sa dá získať spojením väčšieho počtu neurónov do siete. Samotné XOR z predchádzajúcej podsekcie sa dá vyjadriť ako *(x OR y) AND (NOT(x) OR NOT(y))* a podľa tohto vieme vytvoriť aj jednoduchú viacvrstvovú sieť, ktorá používa len neuróny s týmito funkciami (obrázok 2.4,

matematika týchto viacvrstvových modelov nám však dovoľuje vytvoriť aj jednoduchšie siete pre rozoznávanie XOR, je len potrebné zmeniť aktivačnú funkciu a váhy jednotlivých spojení).



Obr. 2.3: Ukážka rozdielu medzi jednovrstvou(a) a viacvrstvovou sieťou (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), falošné vstupy do každého neuŕony nie sú ukázané (Russell a Norvig, 2016).



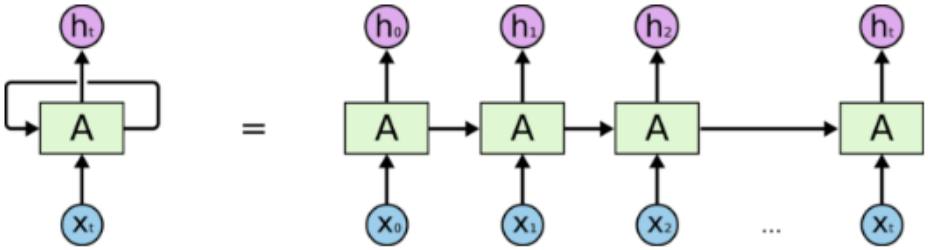
Obr. 2.4: Jednoduchá ukážka viacvrstvovej neurónovej siete rozoznávajúcej XOR.

2.2 Rekurentné neurónové siete

Na druhej strane máme rekurentnú neurónovú sieť (*RNN*). Tento typ siete posúva svoj výstup naspäť do svojho vlastného vstupu (obrázok 2.5). Z toho vyplýva, že aktivačné úrovne siete tvoria dynamický systém, ktorý môže dosiahnuť stabilný stav, oscilovať či sa dokonca správať chaoticky.

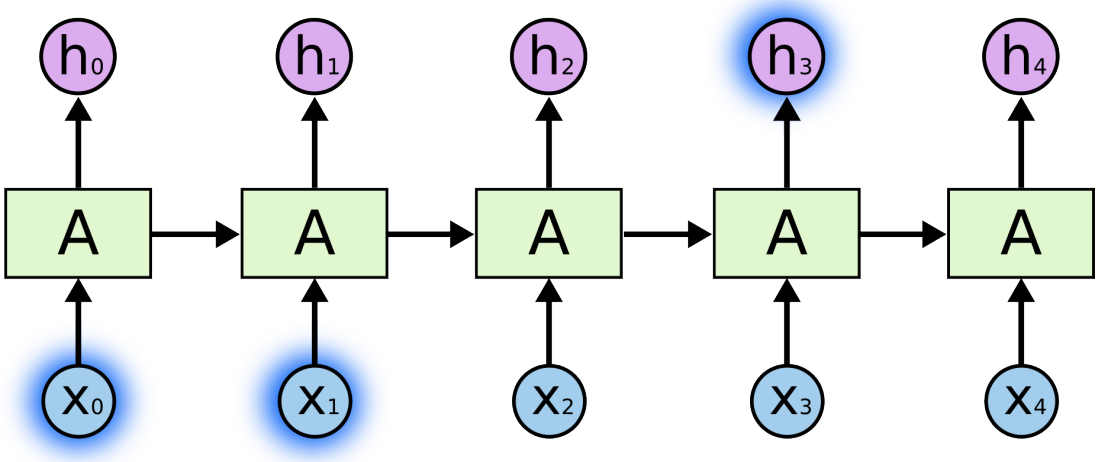
Výstup siete závisí na vstupe. Pri tomto type siete môže výstup závisieť aj na predchádzajúcich výstupoch, tranzitívne teda aj na predchádzajúcich vstupoch. Z toho vyplýva, že si rekurentná neurónová sieť môže vypracovať krátkodobú pamäť (Russell a Norvig, 2016).

RNN sa používa tam, kde dopredná neurónová sieť zlyháva, a to keď nám záleží na závislosti na predchádzajúcich vstupoch. Príklady použitia sú predikcia nasledujúceho slova v texte, rozpoznávanie reči alebo preklad textu medzi jazykmi. Ako príklad si môžeme predstaviť slovné spojenie „mraky sú na nebi“.



Obr. 2.5: Jednoduchá ukážka rekurentnej neurónovej siete. A je sieť, x_i je jej vstup a h_i je jej výstup. Sieť si predáva medzi výpočtom výstupov stav (vodorovné šípky) (Olah, 2015).

Ak by sme predpovedali posledné slovo z tohto slovného spojenia, tak by nám nestačilo poznať len posledné slovo, ale aj pár predchádzajúcich. Tento prípad ukazuje krátkodobé závislosti jednotlivých slov, teda slová, ktoré na sebe závisia sú v krátkej vzdialenosti od seba (na obrázku 2.6 môžeme vidieť príklad RNN na vyhodnocovanie krátkodobých závislostí).



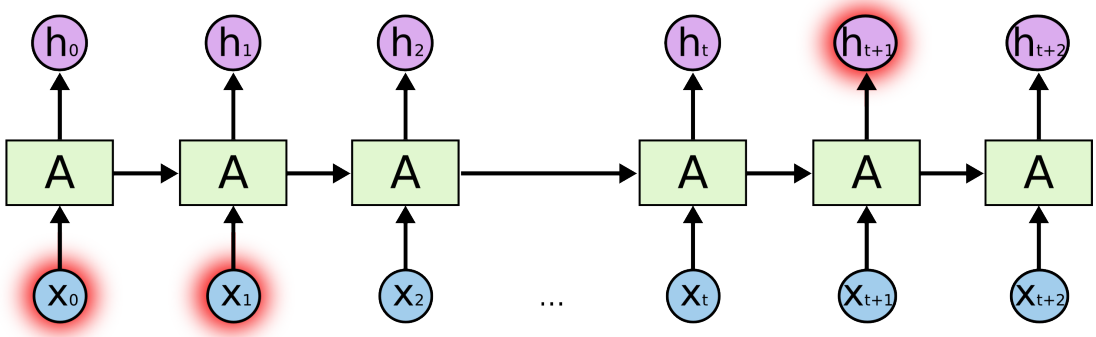
Obr. 2.6: Ukážka krátkodobej závislosti, výstup h_3 závisí na vstupoch x_0 a x_1 (Olah, 2015).

Pri doprednej neurónovej sieti by sme to vedeli dosiahnuť, ale potrebovali by sme zafixovať počet slovných n-gramov, ktoré by sme použili a každý ďalší by zvýšil výpočetnú náročnosť. Ak by sme mali napríklad slovné spojenie „mraky sú na modrom nebi“, tak na predikciu posledného slova by sme si museli uchovávať informáciu o posledných 4 slovách, čo je o jedno viac ako pri poslednom príklade.

V prípade RNN nám stačí vždy vyhodnocovať posledné slovo a predávať si nejaký stav, ktorý sieť má. Pre náš príklad by sme si mohli posúvať informáciu o tom, že hovoríme o mraku. Všeobecne pri predikcii nasledujúceho slova v texte by sme si ideálne chceli uchovávať informáciu o gramatických kategóriách dôležitých slov, pretože v slovenčine aj češtine je táto informácia dôležitá pre vytvorenie správneho tvaru predikovaného slova, nakoľko sa gramatické kategórie medzi vetnými členmi, ktoré sú spolu v syntaktickom vzťahu, musia zhodovať.

Teoreticky sme schopní vytvoriť RNN, ktorá si pamätá závislosť na jednotli-

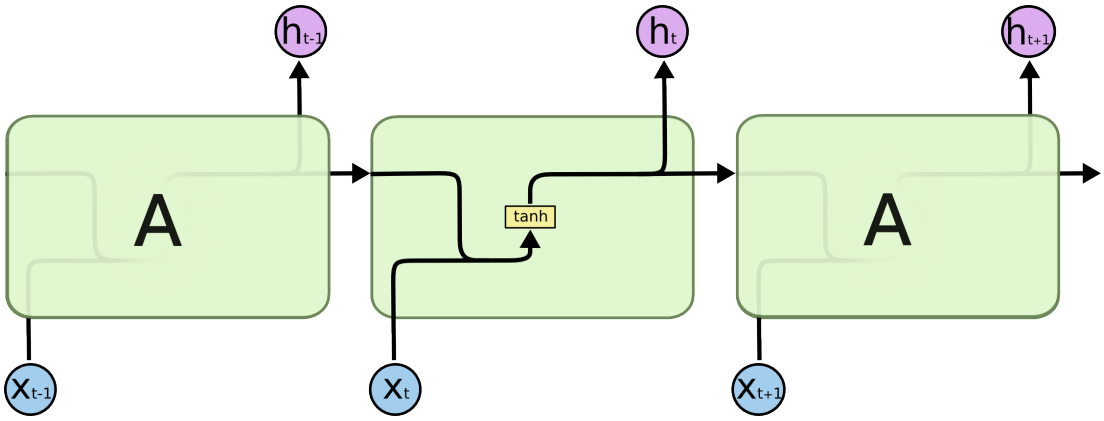
vých slovách s ľubovoľnou vzdialenosťou medzi jednotlivými slovami (ľubovoľným počtom slov medzi nimi). Nanešťastie v praxi to nefunguje až tak ideálne, vo svojej práci to ukázali Bengio, Simard a Frasconi už v roku 1994 (Bengio a kol., 1994). Pre príklad si môžeme predstaviť text, kde sa niekde na začiatku objaví veta „Narodil som sa vo Francúzsku.“, potom nasleduje nejaký text a o pár viet ďalej nasleduje časť „...hovorím plynule francúzsky“. Ak by sme sa pokúsili predpovedať posledné uvedené slovo, možno by sme sa dozvedeli, že chceme nejaký jazyk, ale v šume zo všetkých slov medzitým by sme stratili informáciu o krajine (Olah, 2015). Tento príklad predstavuje tzv. dlhodobé závislosti (na obrázku 2.7 môžeme vidieť príklad dlhodobej závislosti v RNN).



Obr. 2.7: Ukážka dlhodobej závislosti, výstup h_{t+1} závisí na vstupoch x_0 a x_1 (Olah, 2015).

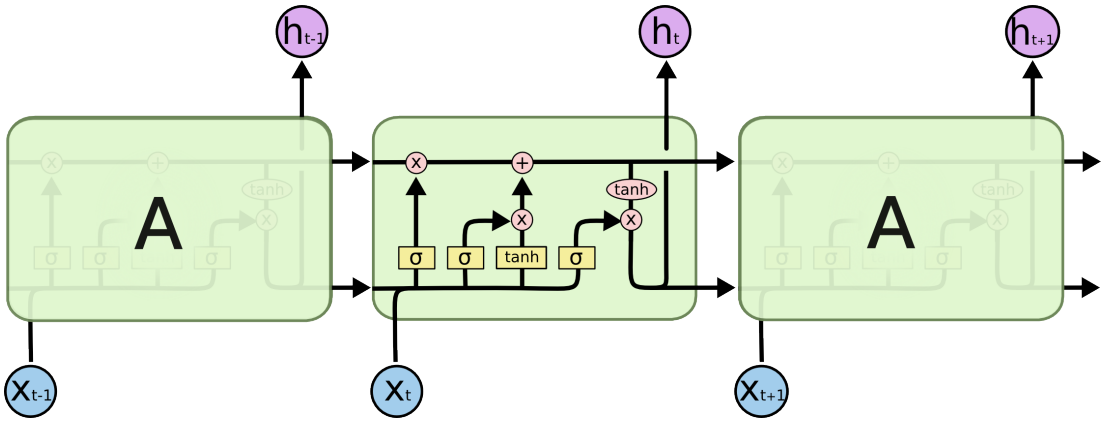
2.2.1 LSTM

LSTM (skratka pre *Long Short-Term Memory* voľne preložiteľné ako ďaleká krátkodobá pamäť) je špeciálnym typom RNN skonštruovaným tak, aby mal čo najmenšie problémy pri dlhodobej závislosti (Olah, 2015). Predstavili ich v roku 1997 vo svojej práci Hochreiter a Schmidhuber (Hochreiter a Schmidhuber, 1997). Všetky štandartné RNN majú veľmi jednoduchú reťazovú štruktúru, napríklad na vstup a stav siete (posledný výstup siete) aplikujú funkciu *tanh* a výsledok pošlú na výstup a ako stav ďalšej sieti (príklad je vidieť na obrázku 2.8).

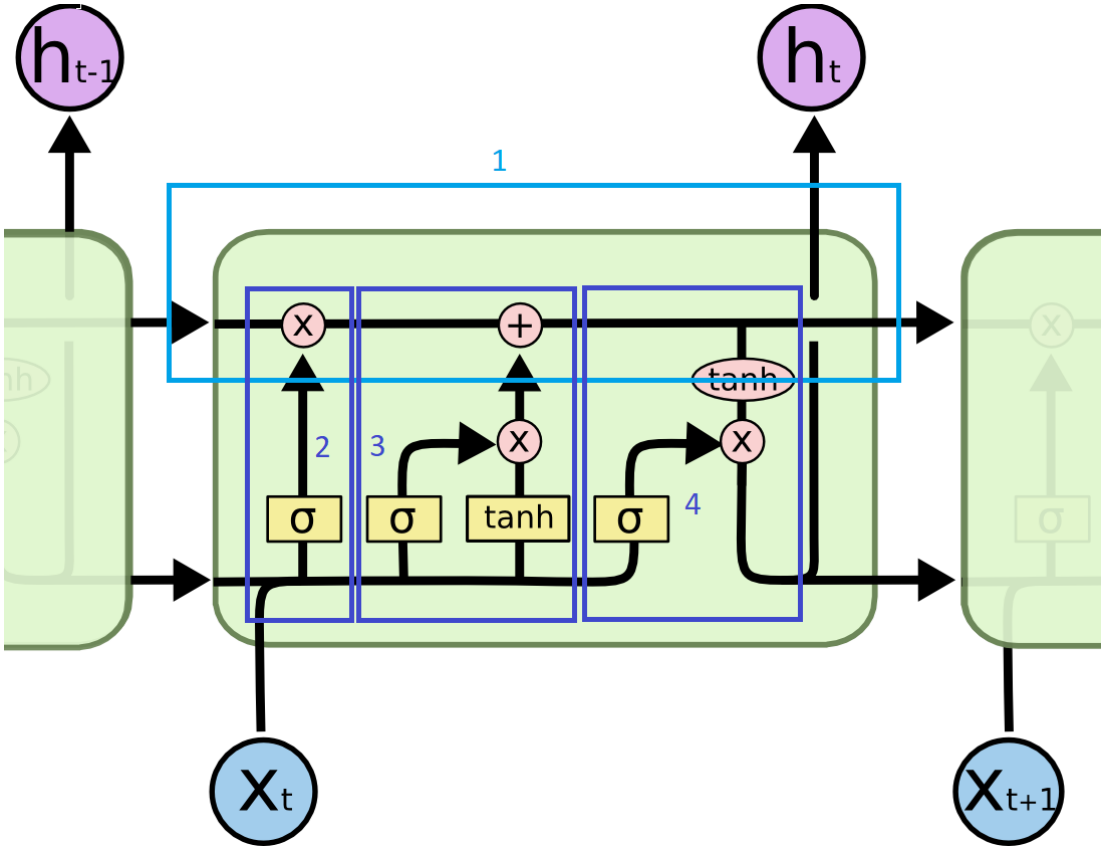


Obr. 2.8: Ukážka ako vyzerá jednoduchá RNN zvnútra (Olah, 2015).

LSTM má tiež reťazovú štruktúru, ale obvykle funguje mierne komplikovanejšie, namiesto jednej vrstvy, ktorá aplikuje nejakú funkciu na vstup a stav siete, obsahuje hneď 4 vrstvy (príklad jednej z možných prevedení LSTM je na obrázku 2.9).



Obr. 2.9: Ukážka jedného zo spôsobov implementácie LSTM zvnútra (Olah, 2015).



Obr. 2.10: Ukážka prezentovaného spôsobou RNN rozdelená na 4 časti.

Základom LSTM je stav siete (na obrázku 2.10 úsek označený číslom 1). Tento stav pokračuje cez celú sieť až na koniec iba s miernými zmenami. Veľmi jednoducho sa môže stať, že sa stav počas výpočtu skoro nezmení.

LSTM má schopnosť odstrániť stav alebo pridať do neho nejakú informáciu pomocou štruktúr zvaných brány. Brány predstavujú spôsob, ako môže informácia prejsť. Sú zložené zo sigmoidovej σ vrstvy a násobení po zložkách. Výstup σ funkcie je $x \in (0,1)$, takže číslo blízko nuly znamená, že skoro žiadna informácia sa nedostane ďalej a číslo blízko jednotky znamená, že prejde skoro všetko (na obrázku 2.10 môžeme vidieť 3 takéto brány).

Prvým krokom je rozhodnúť, ktoré informácie sa nedostanú ďalej zo stavu siete. Toto rozhodnutie robí tzv. zabúdacia brána (na obrázku 2.10 označená číslom 2). Tá sa pozrie na vstup x_t a posledný výstup h_{t-1} a na základe týchto čísel sa rozhodne, nakoľko ponechá stav siete.

Keď si predstavíme model predikcie nasledujúceho slova, tak jedným zo stavov siete môže byť rod predmetu, o ktorom je momentálny text, to je potrebné, aby sme mohli použiť správne zámeno, ak sa budeme naň odkazovať. Ak na vstup príde predmet alebo nový podmet, je pravdepodobne na čase zabudnúť starú informáciu o rode a pridať novú (Olah, 2015).

O pridanie novej informácie sa stará ďalšia vrstva v sieti (na obrázku označená číslom 3). Táto vrstva sa skladá z tzv. vstupnej brány, ktorá sa stará o to, ktoré informácie si ponecháme a *tanh* vrstva, ktorá vytvára vektor nových kandidátov, ktoré by mohli byť pridané do stavu.

Stav siete teda upravíme nasledovne: najprv ho prenásobíme hodnotou x_{f_t}, ktorá je výstupom zabúdacej brány a potom pripočítame do stavu nové informácie. Tento stav sa potom posúva ako stav do ďalšej iterácie.

Nakoniec sa musíme rozhodnúť, čo pôjde na výstup siete. Najprv prejde stav cez funkciu *tanh*, ktorá stlačí hodnoty stavu medzi -1 a 1 a potom prejde poslednou bránou (na obrázku 2.10 označenou číslom 4), tá rozhodne, ktorá časť stavu pôjde na výstup (Olah, 2015).

V našom príklade sme teda dostali ďalší predmet. Na výstup by sieť podľa toho, aký vetný člen očakáva, že bude nasledovať, mohla predať relevatné informácie o gramatických kategóriách. Stav môže naďalej obsahovať všetky tieto informácie.

2.3 Učenie

V predchádzajúcich sekciách sme hovorili o nastavení váh jednotlivých spojení medzi neurónmi a výbere aktivačnej funkcie. V praxi si ale tieto hodnoty obvykle nenastavujeme manuálne, ale nastavuje si ich sieť sama procesom zvaným učenie. Neurónová sieť je učená iteratívnym spôsobom. V každej iterácii dostane sieť množinu vstupov. Pre každý vstup vypočíta hodnotu odhadovaného výstupu, potom sa sieť pozrie na očakávaný výstup a zapamätá si rozdiel týchto hondôt. Po skončení iterácie sa sieť pozrie na hodnoty týchto rozdielov a upraví si svoje váhy tak, aby nabudúce pri rovnakých dátach vydala výstup bližší k očakávanému výstupu. Proces by mal naučiť sieť ohodnotiť celý dátaset čo najpresnejšie. Ak dáta dobre reprezentujú celý problém a nie len nejakú jeho časť, tak sa môže naučiť generalizovať, teda vydať správny výstup aj na dáta, ktoré predtým nikde nevidela.

Konkrétnejšie, proces učenia začne inicializáciou váh. Váhy sa obvykle inicializujú náhodne. Po inicializácii začne už spomínaný iteratívny proces. V každom kroku zvolíme príslušnú podmnožinu trénovacích dát (*batch*) a vyhodnotíme vy-

počítané výstupy porovnaním s očakávanými a pozmeníme jednotlivé váhy podľa toho. Vypočítané výstupy sa vyhodnocujú pomocou stratovej funkcie. Našou úlohou je hodnoty stratovej funkcie minimalizovať. Jednou z najpoužívanejších stratových funkcií je stredná štvorcová chyba (*Mean Squared Error*). Táto funkcia je definovaná ako funkcia

$$MSE(x,y,f) = \frac{1}{m} \cdot \sum_{i=1}^{m} (f(x)_i - y_i)^2$$

kde $f : \mathbb{R}^n \to \mathbb{R}^m$ je funkcia, ktorú simuluje neurónová sieť, $x \in \mathbb{R}^n$ vstupný vektor a $y \in \mathbb{R}^m$ očakávaná hodnota výstupu. To je hodnota MSE pre jeden vstup, ale my trénujeme v podmnožinách vstupných dát veľkosti k a až potom vyhodnocujeme. Pre tento prístup je MSE definovaná

$$MSE(X,Y,f) = \frac{1}{k} \cdot \sum_{j=1}^{k} (MSE(X_j,Y_j,f)) = \frac{1}{k} \cdot \sum_{j=1}^{k} (\frac{1}{m} \cdot \sum_{i=1}^{m} (f(X_j)_i - Y_{j_i})^2))$$

kde $X \in \mathbb{R}^k \times \mathbb{R}^n$ je množina vstupných vektorov a $Y \in \mathbb{R}^k \times \mathbb{R}^m$ je množina očakávaných výstupov.

Na riešenie klasifikačných problémov, teda problémov, kde výstupom je vektor, ktorý obsahuje samé 0 a jednu 1, ktorá určuje, do ktorej triedy je vstup zaradený (ako v našom prípade, kde klasifikujeme zápasy na výhry, prehry a v prípade futbalu aj remízy), sa používa softmaxová strata. Softmaxová strata (niekde v literatúre aj *cross-entropy loss*) sa aplikuje len, keď je aktivačná funkcia vo výstupných neurónoch softmax σ. Softmax $\sigma : \mathbb{R}^k \to \mathbb{R}^k$ je daná ako

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

Vlastnosti tejto funkcie ukazujú, že súčet všetkých zložiek výstupného vektoru je 1, takže táto funkcia v prípade neurónových sietí ukazuje názor siete na to, ako pravdepodobné je, že sa vstup nachádza v triede $i \; \forall i \in 1, \ldots, k$. Na konci sa vždy upraví výstup softmaxu tak, aby bol výsledný vektor vhodný k výstupu (teda najvyššia hodnota sa premení na 1, zvyšné na 0). Označme si tento výstupný vektor ako v. Keď je súčet zložiek rovný 1, vieme použiť funkciu na výpočet entropie ($H = -\sum_{i=1}^{k} (x_i \cdot log_2(x_i))$). Takže softmaxová stratová funkcia vyzerá nasledovne ($p \in 1, \ldots, k$ je $argmax_{\sigma(z)}$)

$$L(x,y) = -\sum_{i=1}^{k} (v_i \cdot log_2(\sigma(z)_i)) = -\sum_{i=1}^{k} log_2 \left(\frac{e^{z_p}}{\sum_{j=1}^{k} e^{z_j}} \right)$$

Ďalšie masívne používané stratové funkcie sú absolútna strata ($L(x,y,f) = \frac{1}{m} \sum_{i=1}^{m} |f(x)_i - y_i|$), ϵ-necitlivá strata ($L(x,y,f) = \frac{1}{m} \sum_{i=1}^{m} (|f(x)_i - y_i| - \epsilon)$), logistická strata ($L(x,y,f) = \frac{1}{m} \sum_{i=1}^{m} \left((ln(2))^{-1} \cdot ln(1 + e^{-f(x)_i y_i}) \right)$) (Rosasco a kol., 2004).

Jedným zo spôsobov, ktoré sa používajú na úpravu jednotlivých váh (*optimizer*) je tzv. *Gradient Descent*. Matematická analýza nám umožňuje získať smer stúpania každej diferencovateľnej funkcie. Prirodzene, hodnoty stratovej funkcie sa snažíme znížiť, teda posunúť váhy v smere klesania, teda proti smeru stúpania

stratovej funkcie. Na to potrebujeme diferencovateľnú stratovú funkciu, z čoho vyplýva, že aj neurónová sieť a aktivačné funkcie vo vnútri, musia byť diferencovateľné. Naším cieľom je vypočítať gradient stratovej funkcie. Keď ho nájdeme, zameníme v ňom znamienka, teda prenásobíme ho číslom -1. Následne môžeme v tomto smere zmeniť váhy. Váhy sa menia v čase t nasledovne:

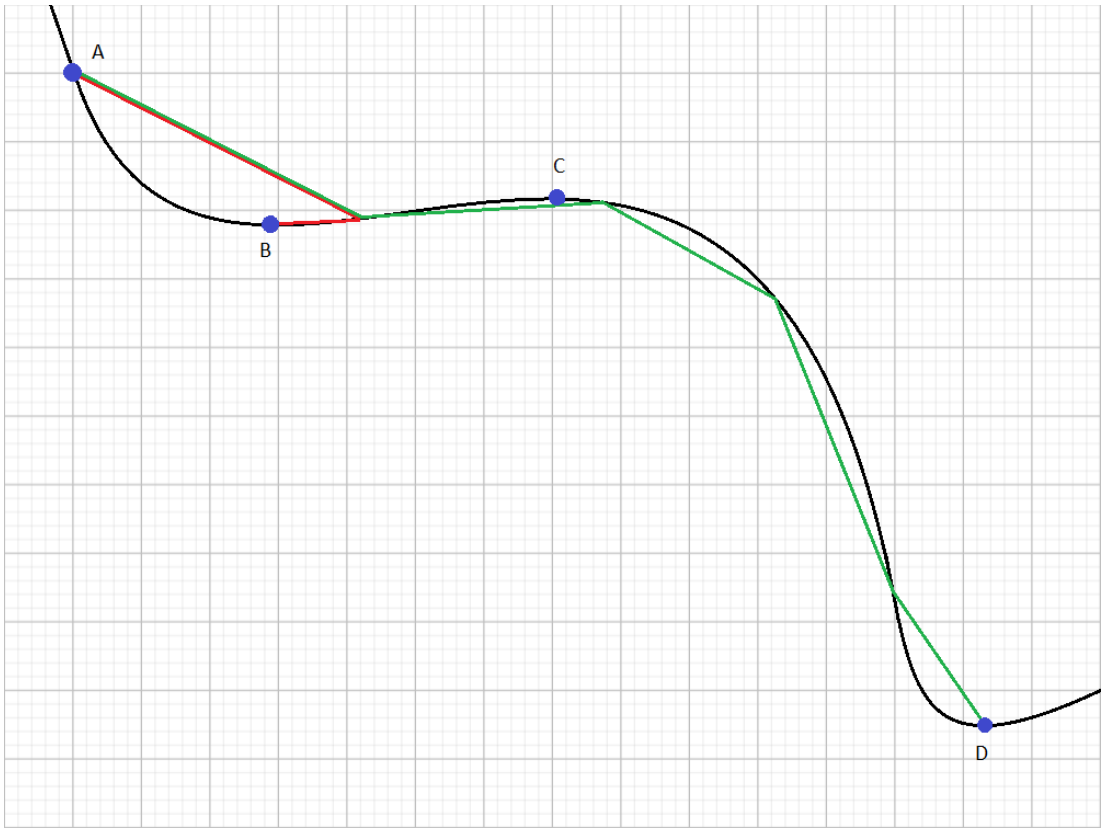
$$w_{t+1} = w_t - \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_w L(x,y,f_t)$$

kde L je stratová funkcia, f_t je funkcia, ktorú neurónová sieť simuluje v čase t a $\alpha \in \mathbb{R}^+$ je veľkosť modifikácie (*learning rate*) a jeho hodnoty sa môžu podľa algoritmu učenia meniť počtom iterácií (Bottou, 2010). Veľkosť modifikácie by sme ideálne chceli malú, aby sme náhodou minimum stratovej funkcie nepreskočili. Ak je ale veľkosť modifikácie príliš nízka, učenie trvá dlho.

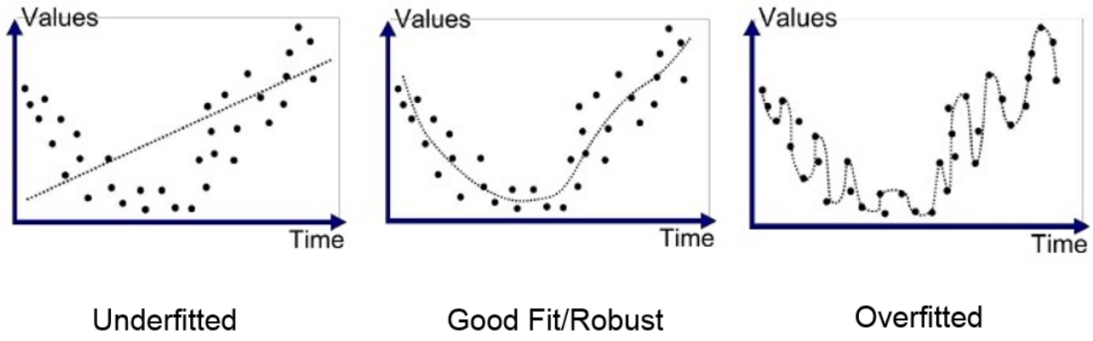
O trochu zložitejším spôsobom je *Adam* (Kingma a Ba, 2014). Názov je vytvorený zo súslovia adaptívny odhad momentu (*ADAptive Moment estimation*). Tento spôsob funguje na podobnom princípe ako *Gradient Descent* s výnimkou toho, že si pamätá a používa pár posledných gradientov, ich význam postupne znižuje. V situácii, kde *Gradient Descent* narazí za lokálne minimum, algoritmus začne meniť váhy v opačnom smere, zatiaľ čo *Adam* bude ešte jemne posúvať algoritmus v smere, v ktorom išiel predtým a možno sa stane, že prekoná lokálne maximum a opäť pôjde smerom dole (ukážku vidieť na obrázku 2.11). Ak sa nedostane cez lokálne maximum, tak sa eventuálne uspokojí a skončí na rovnakom mieste, ako by skončil *Gradient Descent*.

Jedným z problémov neurónových sietí je to, že dosiahnuť ani len lokálne minimum nemusí byť požadujúce. Ideálne požadujeme, aby sieť správne generalizovala, a teda aby jej výstup bol správny aj pre dáta, na ktorých sa neučila. Aj keď trénovacia vzorka reprezentatuje realitu, môže sa stať, že sieť bude mať nízku hodnotu trénovacej chyby (stratovej funkcie), ale relatívne vysokú hodnotu testovacej chyby (príklad možno vidieť na obrázku 2.12). Testovacia chyba sa vypočítava ako hodnota používanej stratovej funkcie na dátach, ktoré neboli obsiahnuté v trénovacej vzorke. Tomuto fenoménu sa hovorí pretrénovanie (*overfitting*) (Brownlee, 2018).

Jedným zo spôsobov boja proti pretrénovaniu siete je použitie nejakej regularizačnej techniky. Regularizácia pomáha kontrolovať pretrénovanie siete (Brownlee, 2018). Jednou z takýchto techník je napríklad technika zvaná *dropout*. *Dropout* dostane spojenia od všetkých neurónov a náhodne zvolí, ktoré neuróny nedostanú svoje spojenie ďalej (ktorým neurónom nastaví váhy spojení na 0). Pre každú trénovaciu podmnožinu dát (*batch*) to volí náhodne, takže na konci je veľmi veľká pravdepodobnosť, že každý neurón sa dostane ďalej a bude môcť byť jeho prísun evaluovaný a váha jeho spojenia pozmenená.



Obr. 2.11: Ukážka rozdielu medzi algoritmami *Gradient Descent* (červenou) a *Adam* (zelenou). Čiernou farbou je znázornený graf hodnôt nejakej stratovej funkcie, A je bod, na ktorom obe algoritmy začínajú, B je lokálne minimum, C lokálne maximum a D globálne minimum. *Adam* používa aj gradienty z minulých iterácií a tak je v tomto prípade schopný preskočiť lokálne maximum a pokračovať ďalej.



Obr. 2.12: Ukážka troch stavov, v ktorých sa môže sieť nachádzať: môže byť podtrénovaná (underfitted), správne natrénovaná (good fit) a pretrénovaná (overfitted) (Gandhi, 2018).

3. Datasety

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

3.1 Futbal

Pre futbal získame 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 dôvodu udelený kontumačne jednému z tímov (ako sa napríklad stalo vo francúzskej lige v roku 2017 pre problémy s divákmi (Ligue 1.com, 2017)) 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.

Dáta pre každú ligu predstavujú výsledky všetkých zápasov odohraných len vrámci ligy za pár posledných sezón. Nebudeme používať žiadne dáta informujúce o hráčoch, ktorí sú na oficiálnej súpiske na zápas ani dáta o základnej zostave na daný zápas a ani ďalšie informácie o priebehu zápasu ako percentuálne držanie lopty alebo počet striel, či rohových kopov. Taktiež vzhľadom na to, že tímy v jednotlivých ligách hrajú zápasy aj mimo ligy, prinajmenšom zápasy v ligovom pohári, nebudú použité ani informácie o oddychu pred daným zápasom, teda koľko dní pred zápasom mali zúčastnené tímy voľno.

Dátaset 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 dátasete konzistentný),
2. Jednoznačný názov hosťujúceho tímu,
3. Identifikátor zápasu,
4. Ligové kolo, v ktorom sa zápas odohral (0, ak sa nevie),
5. Identifikátor domáceho tímu,
6. Identifikátor 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.

Posledné 3 stĺpce teda predstavujú kurzy na dané výsledky. Tieto ale nie sú pri trénovaní siete využívané, a teda pre dáta, ktoré sú vždy použité len pre trénovanie, nie sú nevyhnutné.

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

Dáta v trénovacom súbore obsahujú prvú polovicu sezóny 2018/2019 a 7 jej celých predchádzajúcich sezón. Prvá polovica sezóny 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 7 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).

3.1.1 Motivácia pre výber daných príznakov

Vybrané príznaky presne popísané v sekcii Prílohy (Príloha A.1) sa dajú rozdeliť viacerými spôsobmi do skupín. Z daných príznakmi sa ešte budú selektovať tie najdôležitejšie v jednej z nasledujúcich sekcii (konkrétne sekcia 4).

Prvým spôsobom je rozdeliť tieto príznaky tak, ako za sebou nasledujú do skupiny po desiatich. To nám vytvorí 5 skupín, môžeme ich po poradí nazvať príznaky sezóny, roly, formy, vzájomných zápasov a doplnkové.

Skupina príznakov sezóny obsahuje dáta o celom doterajšom priebehu sezóny pre obe tímy. Mohlo by byť dôležité vedieť, ako daný tím vystupuje v celej sezóne.

Príznaky roly obsahujú dáta o výsledkoch daných tímov v role, v akej sa predstavia v predikovanom zápase (domáci alebo hostia) počas celej sezóny. To by mohlo byť dôležité, pretože je rozdiel v zápasoch, kde tímy hrajú doma a v zápasoch, kde hrajú vonku. Tento rozdiel je individuálny. Tím môže vyhrávať napríklad len na domácom ihrisku a mimo neho sa im nemusí až tak dariť, v tom prípade by v role hostí nemuseli byť favoritom na víťazstvo, aj keď by to mohla predchádzajúca skupina očakávať.

Tretia skupina je forma. Hodnotených je posledných 5 zápasov, tradične to v predikovaných ligách predstavuje obdobie 4 – 5 týždňov. Forma by mala byť

dôležitá, pretože určuje, ako sa tímu darilo v lige v posledných zápasoch, teda predstavuje niečo ako psychickú pohodu tímu, s ktorou vstupuje do zápasu.

Príznaky vzájomných zápasov určujú posledných 5 vzájomných zápasov, ktoré dané tímy odohrali pred predikovaným zápasom. Prvých 5 príznakov sa týka celkových 5 vzájomných zápasov, ďalších 5 sa týka vzájomných zápasov odohraných na ihrisku domáceho v predikovanom zápase. Každý tím hrá iný štýl hry a každý štýl funguje lepšie proti nejakému štýlu a horšie zas proti inému štýlu (to bol jeden z výsledkov vo vyššie spomínanom článku (Shin a Gasparyan, 2014)). Tímy svoje štýly nezvyknú meniť veľmi často, pretože obvykle by na to potrebovali aj výmenu hráčov. Cieľom týchto príznakov je ohodnotiť, ako sa obvykle darí daným tímom, keď sa stretnú medzi sebou.

Poslednou skupinou sú zvyšné príznaky určené na trénovanie, a to je dlhodobá sila domáceho a hosťujúceho mužstva a skóre. Niekedy sa môže stať, že tím mal ťažký úvod do sezóny, ale z minulých sezón vieme, že sa im v lige darí obvykle oveľa lepšie a môžeme očakávať, že v nasledujúcich zápasoch začnú uhrávať lepšie výsledky. To je dôvodom výberu dlhodobej sily mužstva do skupiny príznakov. Je to najbližšie ako sa môžeme dostať ku abstraktnému ohodnoteniu kvality mužstva, ktoré používali iní autori (ako napríklad Shin a Gasparyan v (Shin a Gasparyan, 2014)) z reálnych dát.

Skóre je pokus ohodnotiť formu tímu čo najlepšie jedným údajom. Čím väčší počet vstupných neurónov, tým viac dimenzií dodávame neurónovej sieti. S väčším počtom dimenzií rastie objem celkového priestoru, a to znamená, že sa jednotlivé body dát od seba oddeľujú a samotné dáta sa stávajú redšie. Tomuto fenoménu hovoria vedci kliatba dimenzionality (Domingos, 2012). Práve kvôli tomu vznikol pokús vytvoriť umelé príznaky, ktoré by mohli ohodnotiť rozdiel formy oboch tímov v jednom údaji, na rozdiel od desiatich (popísané sú v Prílohe A.1 44 a 45, detailnejšie pod zoznamom).

Každá z prvých 4 skupín obsahuje dvakrát 5 rôznych údajov, z ktorých môžeme opäť spraviť päť skupín príznakov, a to víťazstvá, remízy, prehry, priemerný počet strelených gólov a priemerný počet inkasovaných gólov.

Prvé tri tieto skupiny hovoria samé za seba, snažíme sa predikovať výsledok zápasu, ktorý je buď výhra, remíza alebo prehra z hľadiska oboch tímov. Zvyšné dve skupiny sa snažia z daných dát simulovať niečo ako ofenzívnu a defenzívnu silu mužstva, podobne ako robili iní autori (napríklad (Igiri a Nwachukwu, 2014)).

3.2 Tenis

Pre tenis získame všetky zápasy každého turnaja ATP typu 500, 1000 a Grand Slam, kde hraá 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. Fakt, že množstvo hráčov z Top 100 sa pravidelne zúčastňuje turnajov typu ATP 250, určuje, že niekedy nastúpia dvaja takíto hráči aj proti sebe na takomto turnaji. Vďaka tomu a aj relatívnej kvalite týchto turnajov sme sa rozhodli, že tieto turnaje zoberieme do úvahy (vzájomné zápasy medzi jednotlivými hráčmi na takto ohodnotených turnajoch tiež patria medzi údaje, z ktorých sa stávajú vstupné neuróny (ako vidieť v Prílohe 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 dátach. 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 totiž nepatrí medzi ciele tejto práce. Ďalšou výhodou je spravodlivosť oblastí vstupu ako forma a vzájomné zápasy, pretože sa tam berú len zápasy, ktoré sa dohrali do konca, takže tieto čísla nie sú v žiadnom okamihu nadhodnotené. Ak by napríklad hráč natrafil 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).

Dátaset je tabuľka, v ktorej 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 dáta, pretože nie je možné hrať na dvoch turnajoch takéhoto typu zároveň). Zápasy v turnajoch sú zoradené od finále po prvé kolo, teda intuitívne opačne. Program v transformačnej časti si už to poradie upraví, aby zápasy nasledovali chronologicky. 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*,
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.

* - ak je ID hráča NULL, znamená to, že hráč sa doposiaľ ani raz neumiestnil v Top 100 rebríčka ATP

Postavenia hráčov v rebríčku ATP sú získané z pomocného súboru (ukážku z tohto súboru je možné vidieť na obrázku 3.3), v ktorom sú uchovávané. Súbor v transformačnej vrstve práce si pospája hráčov s ich ID a ich umiestnením v práve vyhodnocovanom roku.

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.

ID	Rank	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	Novak Djokovic	NULL	NULL	NULL	NULL	NULL	NULL	78	16	3	3	3	3	1	1	2	1	1	2	12	1
1	Rafael Nadal	NULL	NULL	NULL	NULL	49	51	2	2	2	1	2	1	2	4	1	3	5	9	1	2
2	Roger Federer	64	29	13	6	2	1	1	1	1	2	1	2	3	2	6	2	3	16	2	3
3	Alexander Zverev	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	83	24	4	4
4	Juan Martin del Potro	NULL	NULL	NULL	NULL	NULL	NULL	NULL	92	44	9	5	NULL	11	7	5	NULL	NULL	38	11	5
5	Kevin Anderson	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	61	32	37	20	16	12	67	14	6
6	Marin Cilic	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	71	23	14	14	21	15	37	9	13	6	6	7
7	Dominic Thiem	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	39	20	8	5	8
8	Kei Nishikori	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	63	NULL	98	25	19	17	5	8	5	22	9
9	John Isner	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	34	19	18	14	14	19	11	19	17	10

Obr. 3.3: Ukážka prvých 10 riadkov spolu aj s hlavičkou z pomocného súboru udržiavajúceho postavenie hráčov v rebríčku ATP. Hodnota *NULL* znamená, že sa hráč na konci daného roku neumiestnil v prvej 100 rebríčka.

3.2.1 Motivácia pre výber daných príznakov

Podobne ako pri futbale si môžeme jednotlivé príznaky zhrnúť do skupín a jednotlivé skupiny predstaviť. Význam jednotlivých príznakov je vypísaný v prílohe (Príloha A.2).

Najprv si môžeme tieto príznaky rozdeliť na skupinové (prvých 30) a jednotlivé. Na skupinové sa môžeme najprv pozrieť ako na skupiny po 6, a to príznaky roku, formy, povrchu, formy na povrchu a vzájomných zápasov.

Príznaky roku určujú, ako sa danému hráčovi darilo v danom kalendárnom roku. Táto kategória má zväčša najvyššie hodnoty zo všetkých skupín. Tieto údaje

sú dôležitejšie v neskorších fázach roka, pretože ukazujú všeobecný obraz o celom ročníku.

Príznaky formy ukazujú, ako sa hráčovi darilo v posledných 10 zápasoch. Na rozdiel od futbalu sa tieto dáta prenášajú z roka na rok, takže ak hráč ukončil rok dvakrát po sebe v Top 100 rebríčka ATP počas sledovaných sezón, tak na začiatku druhého roku sa mu ukáže forma aj z minulého roku. Je to tak vyriešené preto, lebo tenis sa hrá v podstate celý rok narozdiel od futbalu, kde liga zvykne končiť v máji a začínať v auguste a počas tejto doby sa môžu udiať zmeny v tíme. Dôležité sú z podobného dôvodu ako pri futbale, ukazujú momentálnu výkonnosť a psychickú pohodu hráča, s ktorou prišiel do zápasu.

Príznaky povrchu ukazujú, ako hral hráč na povrchu, na ktorom odohrá predikovaný zápas, počas roka. V tenise majú rôzne povrchy rôzne vlastnosti. Každý hráč má svoj preferovaný povrch, kde sa mu hrá najlepšie alebo dosahuje najlepšie výsledky. Tieto príznaky ukazujú celkovú výkonnosť daného hráča na tomto povrchu v doterajšom priebehu sezóny.

Príznaky formy na povrchu predstavujú kombináciu oboch predošlých skupín, formy a povrchu. Uchováva údaje o posledných 10 zápasov odohraných na danom povrchu, dáta sa prenášajú cez rok. Dôvody sú dva: prvým je, že obvykle sa rok začína a aj končí na tvrdom povrchu a druhým dôvodom je psychika hráča. Aj keď hráč nehral skoro celý rok na danom povrchu, jeho podvedomie si určite pamätá, ako sa mu tam darilo naposledy, aj na základe toho sa môže tešiť alebo netešiť na zápasy na danom povrchu.

Poslednú skupinu v tomto pohľade predstavujú príznaky vzájomných zápasov, tie sa ešte delia na dve skupiny, všetky vzájomné zápasy a vzájomné zápasy odohrané na povrchu, na ktorom sa bude hrať predikovaný zápas. Tieto príznaky by mohli patriť medzi tie najdôležitejšie, ak majú hráči dostatočnú históriu vzájomných zápasov. Každý hráč má vlastný štýl a proti niektorým štýlom sa mu hrá lepšie ako proti iným. Naviac, narozdiel od futbalu, tenisti za celú svoju kariéru nezvyknú radikálne meniť svoje štýly hry, takže tieto údaje môžu byť relevantné aj po rokoch. Tieto údaje sú uvedené z pohľadu hráča 1.

Každá táto skupina má údaje o hráčovi 1 a hráčovi 2 v zápase a obsahuje 3 údaje pre každého hráča: počet výhier, prehier a priemerný rozdiel v počte vyhraných a prehraných hier za set. Prítomnosť prvých dvoch je zjavná, snažíme sa predikovať buď výhru alebo prehru. Prítomnosť poslednej je tu ako pokus o ukážku sily hráča. Táto skupina je postavená na teórii, že sa môže stať, že hráč narážal počas turnaja na hráčov, s ktorými v zápase vyhrával jednoznačne a potom prišiel zápas, kde prehral, ale bol to vyrovnaný zápas. Takýto hráč sa potom môže stretnúť s hráčom, ktorý má podobné výsledky, ale ak vyhral, tak to bol tesný zápas a ak prehral, tak prehral jednoznačne. Pri predikcii výsledku tohto zápasu by bol pravdepodobne favorizovaný prvý hráč, ale ak by sme tento údaj nemali a ostatné údaje by boli dostatočne podobné, tak by mohla mať sieť problémy s rozhodovaním.

Do kategórie jednotlivých príznakov patria umiestnenia oboch hráčov v poslednom koncoročnom rebríčku ATP, kategorizácia povrchu a skóre. Postavenie hráča v poslednom koncoročnom rebríčku ATP ukazuje, ako sa darilo hráčovi v poslednom roku (keďže rebríček uchováva body z posledného roka). Na tento údaj sa môžeme odvolávať hlavne na začiatku ročníka, keď je ešte málo údajov z tohto roka.

Kategorizácia povrchu je skupina príznakov, ktorá je prítomná hlavne z dôvodu, že pre rôzne povrchy môže fungovať iné ohodnotenie. Napríklad na tvrdom povrchu by sa kládol dôraz na iné aspekty ako na trávnatom povrchu.

Skóre je pokus ohodnotiť formu hráča výraznejšie ako len počtom výhier a prehier. Pri pokusoch a vylaďovaní siete budeme v kapitole Stavba siete (kapitola 4) selektovať dané vstupné neuróny podľa rôznych kritérií a vyskúšame tiež aj ako sa bude sieť správať, ak nahradíme všetky stĺpce obsahujúce dáta o forme rozdielom v skóre. Teoreticky by sme tým mohli ušetriť 4 vstupné neuróny (forma obsahuje 6 príznakov, ale skóre sa delí na dva príznaky, obe presne popísané v Prílohe A.2).

4. Stavba siete

Všetky siete boli napísané v programovacom jazyku *Python* s použitím knižníc *numpy* a *tensorflow*. Na vylepšovanie siete sme, ako je napísané v kapitole 5, použili údaje, ktoré sa napokon budú pri vyhodnocovaní nachádzať medzi trénovacími dátami.

Konkrétne, pre futbal dáta predstavovali 7 celých sezón a prvú polovicu ďalšej sezóny (ako popísané v 3.1), vyhodnocovacie dáta predstavujú druhú polovicu tejto sezóny. Takže sme trénovacie dáta rozdelili na tri časti, 6 celých sezón a polovicu ďalšej (trénovacie dáta), druhú polovicu siedmej sezóny (testovacie dáta) a zvyšnú prvú polovicu ôsmej sezóny (nepoužité dáta).

V prípade tenisu prišli dáta už priamo z transformačnej časti v troch súboroch, trénovacie, testovacie a vyhodnocovacie dáta.

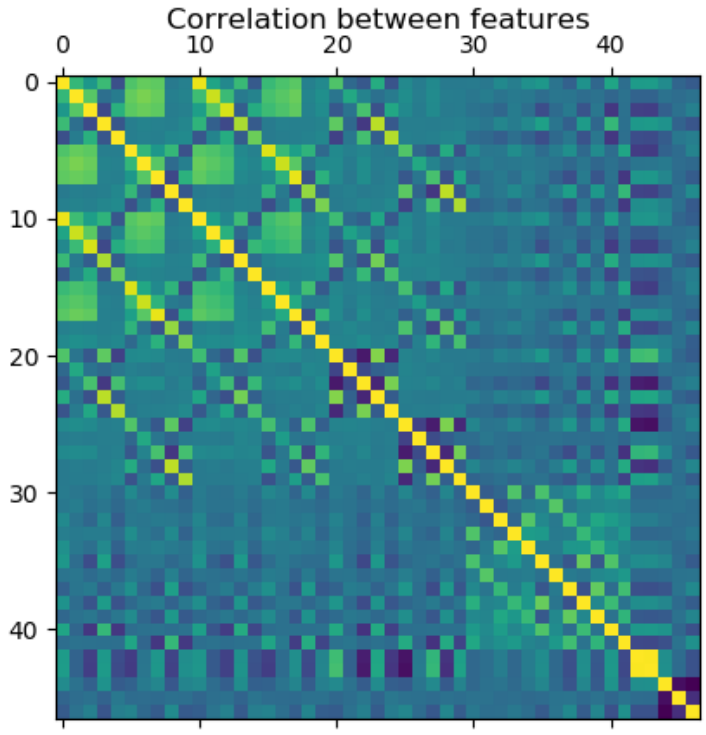
Každá sieť mala svoje nedostatky v celkovej úspešnosti, ale doposiaľ neexistuje efektívne nastavenie neurónových sietí pre každú situáciu (Gandhi, 2018), takže každá sieť sa musela vylepšovať osobitne a manuálne vzhľadom na rozdiely v prístupoch.

4.1 Selekcia príznakov

Kliatba dimenzionality nám hovorí, že čím viac vstupných príznakov zadáme neurónovej sieti, tým sú dáta redšie a teda je ich potreba získať viac, aby sa sieť správne učila. Viac dát získať nevieme, takže sa pokúsime znížiť počet dimenzií a pozrieť sa na to, ako sa to prejaví na trénovacích dátach. Na začiatok spravíme korelačný test všetkých príznakov testovacích a trénovacích dát (obrázky 4.1 a 4.2). Teória hovorí, že by nám mohla napovedať, aké hodnoty sú dôležité. Obrázok 4.1 týkajúci sa futbalu hovorí, že najvyššiu koreláciu s výsledkom zápasu majú príznaky 41 a 42 určujúce dlhodobú silu tímu. Skóre dosahuje tiež celkom vysokú koreláciu (okolo 0,2) s finálnym výsledkom.

V prípade tenisu obrázok 4.2 naznačuje najvyššiu koreláciu medzi výsledkom a oboma druhmi skóre, veľkú rolu zohráva postavenie v rebríčku ATP a vzájomné zápasy (konkrétne priemerný rozdiel v počte vyhraných gemov za set).

Tieto tabuľky boli pre nás viac-menej informačné. Korelovanosť jedného príznaku s výsledkom nám nemusí hovoriť nič. Stačí sa pozrieť na funkciu XOR (tabuľka 2.1). Ak by sme na hodnoty v tejto tabuľke aplikovali koreláciu, tak by sme sa dozvedeli, že x, y a ani hodnota x XOR y nemajú po dvoch žiadnu koreláciu (ak vieme, akú hodnotu má x, nijak to pre nás nemení pravdepodobnosť, že x XOR y bude napríklad 0). Až kombinácia hodnôt x a y priamo určuje hodnoty funkcie XOR. Selekcia príznakov, s ktorými dosahovala sieť najlepšie trénovacie výsledky a ktoré budú použité na získanie výsledkov v kapitole 6, bol uskutočnený spôsobom pokus-omyl, keďže nič lepšie nevieme, ako už bolo spomínané na začiatku kapitoly.

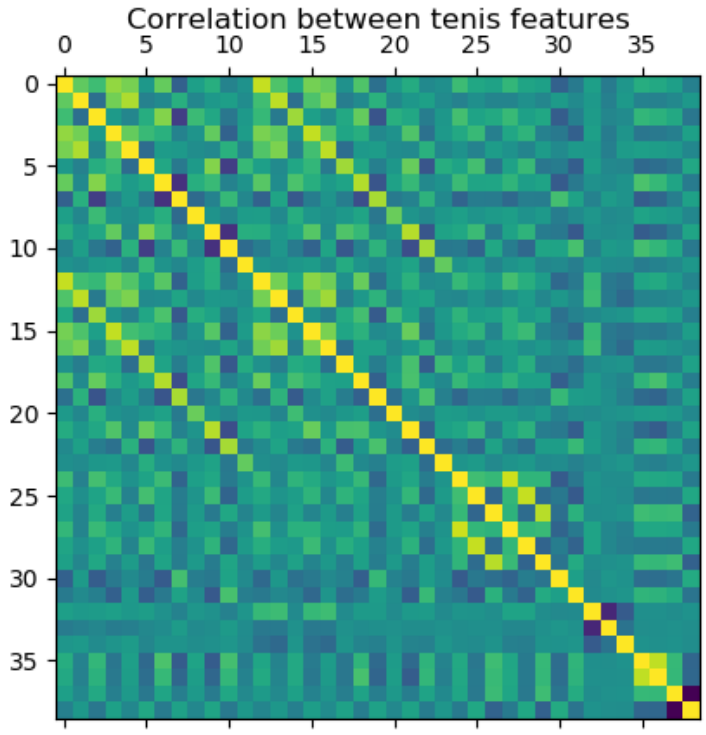


Obr. 4.1: Korelačná tabuľka všetkých príznakov pre anglickú Premier League, žltá predstavuje kladnú koreláciu, modrá zápornú. Nás hlavne zaujímajú posledné 3 riadky určujúce koreláciu príznaku s výsledkom zápasu (príznaky sú v poradí ako v Prílohe A.1).

4.2 Proces tréningu

Proces tréningu siete začínal základným nastavením siete, ktoré je presnejšie popísané v nasledujúcich kapitolách, pretože pre každú sieť bolo unikátne. Po trénovaní siete v základnom nastavení nasledovalo spustenie siete na testovacích dátach. Nasledoval pokus nastaviť parametre (tie budú tiež popísané v nasledujúcich sekciách, líšili sa pre jednotlivé druhy neurónových sietí) jednotlivých sietí pre každý šport a ligu tak, aby dosahovali čo najvyššie sledované hodnoty. Tieto hodnoty predstavovali referenčnú hodnotu, ktorú sme chceli selektovaním jednotlivých parametrov vylepšiť. Následne sme vybrali jednu z podmnožín príznakov (napríklad odstránili údaje o forme) a opakovali proces vyladovania parametrov. Na konci pre každú sieť ostala jedna podmnožina príznakov a nastavenie parametrov siete, ktoré v kombinácii s ňou vydávalo najlepšie výsledky v sledovaných oblastiach.

Sledované oblasti zo začiatku predstavovali trénovaciu a testovaciu úspešnosť, úspešnosť v zápasoch bez favorita a eventuálny zisk, ktorý by sme dosiahli, ak by sme v týchto zápasoch uzatvárali stávky podľa nápovedy danej siete. Všetky modely ale ukazovali veľmi podobnú úspešnosť v tipovaní víťazov zápasov bez favorita (napríklad pri vytváraní siete pre tipovanie anglickej futbalovej Premier League vydávali všetky siete trénovaciu úspešnosť 36–39%). Eventuálny zisk osciloval bez ohľadu na túto úspešnosť. Neobjavil som žiadnu koreláciu medzi na-



Obr. 4.2: Korelačná tabuľka všetkých príznakov pre tenisové zápasy, žltá predstavuje kladnú koreláciu, modrá zápornú. Nás hlavne zaujímajú posledné 2 riadky určujúce koreláciu príznaku s výsledkom zápasu (príznaky sú v poradí ako v Prílohe A.1).

stavenými parametrami, selektovanými príznakmi a týmto ziskom, takže predpokladám, že na hodnote tohto údaja pri trénovaní až tak nezáleží, pretože vo vyhodnocovacom procese môže dosiahnuť úplne iné výsledky. Nakoniec teda som sledované oblasti zúžil na trénovaciu a testovaciu úspešnosť, kde som sa pokúšal maximalizovať testovaciu úspešnosť a trénovacia slúžila hlavne ako referenčný údaj pre pretrénovanie siete (vysvetlené v sekcii 2.3), aby som vedel, kedy už netreba pridávať ďalšie trénovacie iterácie vybranej podmnožiny trénovacích dát (*epoch*), pretože úspešnosť sa už ďalej bude len znižovať.

Je potrebné dodať, že celý trénovací proces prebiehal v iteráciách. Každý model siete bol trénovaný a vyhodnotený 40-krát (s rôznym náhodným nastavením váh spojení) a výsledky v sledovaných oblastiach boli aritmetickým priemerom cez týchto 40 modelov. V prípade tenisu som do tohto aritmetického priemeru nepočítal dáta, ktoré sa od začiatku „zasekli“, čo bolo jednoznačne viditeľné už z trénovacej úspešnosti a trénovacej chyby. Zaseknutie vyzeralo asi tak, že sieť zistila, na ktorej strane vyhráva viac hráčov (napríklad hráč 1 vyhráva v 50,2% prípadov) a ani ďalšie učenie nepomohlo zmeniť názor siete. Fakt, že to bolo viditeľné už z trénovacích dát (obvykle pri tenise obe druhy neurónových sietí dosiahli viac ako 70% trénovaciu úspešnosť), tak sme mohli tieto pokusy odstrániť a vylepšiť tým úspešnosť siete bez toho, aby sme nechali sieť vyhodnocovať testovacie dáta. Ak by sme chceli prakticky použiť neurónové siete na predikciu výsledkov v tenise, tak by sme si tohto faktu všimli ešte predtým, ako by sme stavili na prvý

zápas.

4.3 Dopredné neurónové siete

V prípade futbalu bola na začiatku sieť skonštruovaná so všetkými 44 príznakmi, ktoré sme dostali z transformačnej časti práce (príloha A.1). Sieť teda obsahovala vstup o veľkosti 44 príznakov, vrstva regularizačnej techniky zvanej dropout (s jeho nastavením na 50%), 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.

V prípade tenisu bola sieť skonštruovaná so všetkými 37 príznakmi (ich popis je príloha A.2) a taktiež obsahovala vrstvu *dropout* regularizácie, 2 skryté vrstvy neurónov s rovnakým počtom neurónov na nich ako v prípade futbalu. Výstup predstavoval dva neuróny, na ktoré bol opäť použitý softmax, ktorý vyberie najvyššiu hodnotu.

Parametre siete, ktoré boli počas tréningu prenastavované a vylaďované sú:

1. Spôsob, ktorým sa vylaďujú jednotlivé váhy siete (*optimizer*),
2. Počet iterácií trénovacieho procesu (*počet trénovacích epoch*),
3. Veľkosť jedného kroku trénovania (*batch size*),
4. Veľkosť modifikácie pri učení (*learning rate*),
5. Hodnota náhodne vypustených spojení po prvej vrstve (*dropout*),
6. Počet neurónov v prvej skrytej vrstve,
7. Počet neurónov v druhej skrytej vrstve.

Nastavenia jednotlivých parametrov, ktoré prinášali najlepšie výsledky pre každú sieť, sa nachádzajú v tabuľke 4.2.

Predpovedaný šport (liga)	Parametre siete							Trénovacie výsledky	
	O	E	B	LR	D	H_1	H_2	TrA%	TeA%
Futbal (ENG)	GD	100	64	0,005	0,5	15	10	52,9	50,9
Futbal (GER)	GD	75	16	0,005	0,5	15	10	50,01	50,68
Futbal (SPA)	GD	100	64	0,005	0,5	15	10	55,36	52,78
Tenis	A	30	256	0,01	0,5	15	10	73,98	65,8

Tabuľka 4.1: Tabuľka nastavenia parametrov doprednej neurónovej siete, pri ktorých dávala sieť najlepšie trénovacie výsledky a aj hodnoty, ktoré dosiahli v sledovaných oblastiach. Skratky v hlavičke tabuľky pod parametrami siete sú skratky sledovaných parametrov zo zoznamu (v tom istom poradí, v akom sú v zozname uvedené), trénovacie výsledky sú trénovacia úspešnosť a testovacia úspešnosť v tomto poradí. V stĺpci O (*optimizer*) skratka GD znamená metódu *Gradient Descent* a skratka A metódu *Adam*.

Už z týchto trénovacích výsledkov môžeme vidieť, že model, ktorý dosahoval najlepšie výsledky sa dosť líši pri predikcii rôznych športoch. Celkovo pre doprednú neurónovú sieť bolo vytvorených viac ako 200 rôznych modelov, aj preto je zvláštnosťou, že *Gradient Descent* algoritmus vydával lepšie výsledky pri predikcii futbalu ako jeho vylepšená verzia *Adam*.

Rôzne modely sa líšili aj selekciami rôznych príznakov. V prípade futbalu sa dokonca vybrané príznaky líšili aj na predikovanej lige. Pre anglickú (ENG) a španielsku ligu (SPA) dosahovala najlepšie výsledky sieť, ktorá používala 21 z 44 príznakov. Použité boli všetky príznaky okrem príznakov určujúcich formu, priemerný počet gólov (strelených aj inkasovaných) na zápas a momentálne skóre (konkrétne boli použité príznaky 1, 2, 3, 6, 7, 8, 11, 12, 13, 16, 17, 18, 31, 32, 33, 36, 37, 38, 41, 42 a 43 z prílohy A.1). Tieto vedomosti nám poukazujú na fakt, že som bol úspešný pri nahradení formy jedným z druhov skóre a ušetril som tým 9 dimenzií. Rozdiel v úspešnosti, ak bolo použité skóre a ak nebolo použité skóre bol minimálny, čo poukazuje na fakt, že skóre bolo užitočné, ale nie perfektné a určite by sa dalo nejak vylepšiť.

Pre nemeckú ligu (GER) uvedená sieť používala 19 príznakov. Použité boli všetky príznaky okrem príznakov určujúcich počet remíz vo všetkých kategóriách, priemerný počet gólov na zápas a rozdiel v momentálnom skóre (konkrétne použité príznaky boli príznaky s číslami 1, 3, 6, 8, 11, 13, 16, 18, 21, 23, 26, 28, 31, 33, 36, 38, 41, 42 a 43 z prílohy A.1).

Pre tenis dosahovali najlepšie výsledky siete, ktoré používali 27 príznakov. Príznaky, ktoré neboli potrebné v procese učenia sú povrch, na ktorom sa zápas odohrá, údaje o forme a skóre povrchu (použité teda boli príznaky s číslom 1, 2, 3, 4, 5, 6, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32 a 36 z prílohy A.2). Opäť vidíme, že sa nám podarilo nahradiť údaje o forme umelo (ale automaticky) vytvoreným skóre, čo mierne zvýšilo úspešnosť siete. Pri forme na povrchu sa to ale nepodarilo a lepšie výsledky dosahovala sieť bez použitia skóre povrchu a s použitím tejto skupiny príznakov.

4.4 Rekurentné neurónové siete

Tréning rekurentnej neurónovej siete (*RNN*) je všeobecne časovo náročnejší ako trénovanie doprednej neurónovej siete, nakoľko sa do procesu tréningu zapojí aj stav siete. Sieť sa musí naučiť, ktoré údaje sú dôležité na uchovávanie v stave siete. Z tohto dôvodu som pri tréningu RNN vychádzal aj z údajov získaných v predošlej sekcii, pričom som ale celý proces tréningu (tak, ako je popísaný v sekcii 4.2) zachoval. Základné nastavenie siete používalo všetky príznaky z príloh. Pre futbal ich bolo 44, pre tenis 37 (Prílohy A.1 a A.2 v poradí). Tieto príznaky boli napojené do vrstvy 25 LSTM neurónov, potom nasledoval *dropout* a výstupná vrstva. Pár vyskúšaných nastavení siete obsahovalo pred výstupnou vrstvou ešte skrytú vrstvu neurónov.

Parametre siete, ktoré boli počas tréningu prenastavované a vylaďované sú:

1. Spôsob, ktorým sa vylaďujú jednotlivé váhy siete (*optimizer*),
2. Počet iterácií trénovacieho procesu (*počet trénovacích epoch*),
3. Veľkosť jedného kroku trénovania (*batch size*),

4. Veľkosť modifikácie pri učení (*learning rate*),
5. Hodnota náhodne vypustených spojení po prvej vrstve (*dropout*),
6. Počet LSTM neurónov,
7. Počet zápasov, pre ktoré si pri trénovaní sieť pamätá údaje (*LSTM Timestamp*),
8. Počet neurónov v skrytej vrstve (ak 0, tak model siete neobsahuje túto vrstvu).

Nastavenia jednotlivých parametrov, ktoré prinášali najlepšie výsledky pre každú sieť, sa nachádzajú v tabuľke 4.2.

Opäť môžeme z trénovacích výsledkov vidieť rozdiel medzi jednotlivými športmi. Tentokrát ale v nastaveniach parametrov siete nie je žiadny rozdiel, ale je rozdiel pri vybraných príznakoch, s ktorými dosahovali siete najlepšie výsledky. použitých bolo okolo 60 rôznych modelov a vedomosti z predchádzajúcej sekcie. Pri všetkých druhoch rekurentných neurónových sietí dosahoval najlepšie výsledky optimalizačný algoritmus *Adam*.

Jednotlivé vybrané príznaky pre anglickú aj nemeckú ligu boli rovnaké ako vybrané príznaky pre nemeckú ligu v prípade doprednej neurónovej siete. Použitých bolo 19 príznakov (konkrétne príznaky 1, 3, 6, 8, 11, 13, 16, 18, 21, 23, 26, 28, 31, 33, 36, 38, 41, 42 a 43 z prílohy A.1). Vyradené príznaky predstavovali údaje o počte remíz, priemernom počte gólov na zápas a rozdiele v momentálnom skóre.

V prípade španielskej ligy boli vybrané príznaky rovnaké ako pre dopredné neurónové siete. Použitých bolo 21 príznakov (konkrétne 1, 2, 3, 6, 7, 8, 11, 12, 13, 16, 17, 18, 31, 32, 33, 36, 37, 38, 41, 42 a 43 z prílohy A.1). Vyradené príznaky boli údaje o forme, priemernom počte gólov na zápas a rozdiele v momentálnom skóre.

V prípade tenisu sa vybrané príznaky líšili od doprednej neurónovej siete. Z 37 príznakov bolo vybraných 32, vyradené príznaky boli povrch, na ktorom sa zápas odohral (príznaky číslo 33, 34 a 35 z prílohy A.2) a obe druhy skóre (príznaky

Predpovedaný šport (liga)	Parametre siete								Trénovacie výsledky	
	O	E	B	LR	D	N	T	H	TrA%	TeA%
Futbal (ENG)	A	70	64	0,005	0,5	25	30	10	52,44	51,95
Futbal (GER)	A	100	64	0,005	0,5	25	30	10	49,66	51,21
Futbal (SPA)	A	100	64	0,005	0,5	25	30	10	54,6	53,24
Tenis	A	70	256	0,005	0,5	25	31	0	74,31	65,96

Tabuľka 4.2: Tabuľka nastavenia parametrov RNN, pri ktorých dávala sieť najlepšie trénovacie výsledky a hodnoty, ktoré dosiahli v sledovaných oblastiach. Skratky v hlavičke tabuľky pod parametrami siete sú skratky sledovaných parametrov zo zoznamu (v tom istom poradí, v akom sú v zozname uvedené), trénovacie výsledky sú trénovacia úspešnosť a testovacia úspešnosť v tomto poradí. V stĺpci O (*optimizer*) skratka A znamená metódu *Adam*.

číslo 36 a 37). V tomto prípade zlyhal koncept skóre a použitie žiadneho skóre na úkor skupiny príznakov, ktoré malo nahrádzať nepomohlo, dokonca sieť vydávala lepšie výsledky, ak skóre nevidela.

5. Dokumentácia

Z programátorského hľadiska je práca rozdelená na tri časti. Prvú časť predstavuje získavanie výsledkov a kurzov jednotlivých zápasov. Druhá časť práce je zameraná na transformáciu dát na údaje priamo použiteľné pri konštrukcii neurónových sietí. Poslednú časť tvorí stavba daného typu neurónovej siete pre daný šport.

Transformačná časť v prípade tenisu rozdelí dáta na 3 časti: trénovacie dáta, testovacie dáta (pre optimalizovanie siete) a vyhodnocovacie dáta. V prípade futbalu dáta rozdelí na 2 časti, trénovacie a vyhodnocovacie dáta. Pre optimalizáciu bol použitý program podobný programu v poslednej časti, ktorý si trénovacie dáta rozdelil podľa potreby (konkrétne na prvých 6 celých sezón a prvú polovicu ďalšej a druhú časť tvorí druhá polovicu tejto sezóny), až následne prebiehalo učenie siete. Je teda zaručené, že žiadna sieť neuvidí vyhodnocovacie dáta pred finálnym vyhodnocovaním. Vyhodnocovacie dáta sú použité až pre získavanie výsledkov použitých v tejto práci, a teda použili sa až v poslednej fáze.

5.1 Futbal

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*. Tento program si tiež vytvorí pomocný súbor *teamdatabase.csv*, ktorý slúži na udržiavanie názvov tímov a ich identifikátorov.

Súbor, ktorý vznikne z činnosti tohto programu predáme programu *DataMaker.exe*. To znamená, že ako prvý parameter programu *DataMaker.exe* (tento program sa nachádza v hlavnej zložke) je potrebné predať názov súboru bez prípony aj s jeho relatívnou cestou, 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 dáta na vstupné neuróny pre neurónovú sieť. Ak tento súbor neexistuje, je potrebné otvoriť súbor *DataMaker.sln* v prvej zložke *DataMaker* a zdrojové kódy skompilovať v režime *Debug*, aby tento súbor vznikol. Pre prípad, že pracujeme priamo v zložke, ktorá je elektronickou prílohou k práci, tak relatívna cesta nie je žiadna a stačí ako parameter predať skratku danej ligy (napríklad *DataMaker.exe SPA*). Všetkých vstupných neurónov je 44. Presné poradie aj popis je uvedený v sekcii Prílohy (Príloha A.1). Program taktiež vydá ako posledné tri stĺpce aj výsledok zápasu ako kategorické hodnoty v poradí domáci, remíza, hostia, kde výsledok, ktorý nastal je ohodnotený 1, zvyšné sú 0. (tiež popísané v Prílohe A.1). Tieto údaje boli vybrané špecificky aj s pomocou súvisiacich prác ako údaje, ktoré popisujú stav oboch tímov, ktoré hrajú proti

sebe zápas (viac v sekcii 3.1). Program tiež vytvorí ďalší súbor, ktorý obsahuje testovacie dáta, teda dáta, ktoré sa nevyužívajú pri trénovaní siete, ale len pri vyhodnocovaní výsledkov. Tieto dáta sú v rovnakom poradí a musia obsahovať kurzy na dané výsledky a aj výsledok zápasu vo forme troch stĺpcov. Je to potrebné pre vyhodnocovanie, pretože neurónová sieť bude mať 3 výstupné neuróny v rovnakom poradí a predikovaný výsledok ohodnotí na 1.

Pre upresnenie, súbor vytvorí dve tabuľky opäť s formátom *csv*, názov je zložený zo skratky pre názov danej ligy a slova *input* pre trénovacie dáta, pre testovacie je to skratka danej ligy a slovo *resinput*.

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 dopredná alebo rekurentná neurónová sieť.

Výpis na štandardný výstup sa nám pre učenie siete nepodarilo potlačiť, takže štandardný výstup bude plný týchto vecí, ale po konci každého cyklu učenia a vyhodnocovania vypíše trénovaciu chybu a trénovaciu úspešnosť, rovnako ako testovaciu úspešnosť, úspešnosť v zápasoch bez favorita a zisk v týchto zápasoch. Na konci vypíše priemernú trénovaciu a testovaciu úspešnosť, spolu aj s priemernou úspešnosťou v zápasoch bez jasného favorita a zisk pri uzatváraní stávok na tieto zápasy. Najdôležitejšie údaje pre nás sa zhrnú do záznamového súboru (*logu*). Log obsahuje rôzne údaje od názvu ligy, selektovaných príznakov, cez architektúru siete, až po trénovaciu chybu a trénovaciu a testovaciu úspešnosť, úspešnosť v zápasoch bez favorita a zisk v týchto zápasoch pre každý cyklus učenia a vyhodnocovania a na konci priemerné hodnoty týchto údajov. Log sa vytvorí v zložke *Logs* a jeho meno bude obsahovať skratku pre ligu, na ktorej prebiehalo učenie, časovú známku vo formáte rok, mesiac, deň, hodina, minúta a sekunda zapnutia finálneho programu a aktivačnú funkciu vo výstupnej vrstve. Na začiatku názov obsahuje skratku „RNN“, ak sa vytvárala rekurentná neurónová sieť. Tento súbor bude vo formáte *txt*. Napríklad, ak sieť vytvorí textový súbor s názvom *RNNger20190716212031softmax*, znamená to, že 16.7.2019 o 21:20:31 bola vytvorená neurónová sieť na predikciu výsledkov nemeckej Bundesligy. Aktivačná funkcia tejto siete na výstupnej vrstve bola softmax.

5.2 Tenis

Údaje o tenisových zápasoch sú predpripravené v súbore *atpresults.csv*. Súbor *atpranking.csv* je tiež predpripravený a obsahuje ID jednotlivých hráčov, ich mená a ich poradie v koncoročných rebríčkoch hodnotenia ATP za roky 1999–2018 (ukážku tejto tabuľky predstavuje obrázok 3.3). Poradie berieme, len ak sa hráč umiestnil na miestach 1–100. Poradie údajov v tabuľke *atpresults.csv* je popísané v sekcii 3.2. Na získanie kurzov do tejto tabuľky je ale potrebné spustiť program *atpodds.py* a v zložke, v ktorej sa práve nachádza, je potrebné mať súbory *atpresults.csv* a *atpvenues.csv*, ktorý obsahuje len rozličnosti v menách jednotlivých turnajov medzi *atpresults.csv* a stránkou OddsPortal.com, z ktorej sa online sťahujú tieto kurzy. Je podstatné upozorniť, že tento program nie vždy pracoval úplne bezchybne, kľúčové je mať kvalitné pripojenie, z tohto dôvodu som do elektronickej prílohy priložil súbor *atpresults_backup.csv*, ktorý je už vopred vyplnený a pre pokračovanie ho stačí premenovať na *atpresults.csv*. Program v

zložke, v ktorej sa akutálne nachádza prepíše súbor *atpresults.csv* tak, že už obsahuje dáta v správnom formáte, aby sme ich mohli predať transformačnej vrstve.

Predzápasové kurzy môžu byť prázdne (vyplnené kurzom 0.0), ale len, ak nás pre daný zápas nezaujímajú kurzy (zaujímajú nás len za posledné dva roky, prvý je na testovanie a druhý na vyhodnocovanie).

Dátasety *atpresults.csv* a *atpranking.csv* musia byť v zložke, kde sa nachádza aj program *ATPDataMaker.exe*. Ak sú všetky súbory v zložke rovnako, ako v elektronickej prílohe, tak stačí priamo spustiť program *ATPDataMaker.exe* (ktorý by sa tiež mal nachádzať v tejto zložke), ktorý ich pretransformuje na dáta pre vstupné neuróny neurónových sietí. Ak tento súbor neexistuje, je potrebné otvoriť súbor *ATPDataMaker.sln* v prvej zložke *ATPDataMaker* a zdrojové kódy skompilovať v režime *Debug*, aby tento súbor vznikol.

Tento súbor vytvorí to zložky, v ktorej sú súbory *atpresults.csv* a *atpranking.csv*, tri nové súbory, *atp.csv*, *atpres.csv* a *atpres_final.csv*. Všetky tieto súbory majú údaje v presnom poradí, ako je popísané v Prílohe A.2.

Z popisu dátasetu (v sekcii 3.2) je vidieť, že hráči v zápasoch sú zoradení tak, že najprv je napísaný víťaz a po ňom porazený. To by nám očividne zamiešalo výsledky 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.

Potom môžeme konečne spustiť programy, ktoré vytvárajú neurónové siete. V zložke, v ktorej sa nachádzajú tieto programy sa musia nachádzať aj tri zložky vytvorené v transformačnej vrstve programom *ATPDataMaker.exe*. Program textitffnntennis.py vytvorí doprednú neurónovú sieť aj s najlepším modelom získaným pre tenis v sekcii 4.3. Program *rnntennis.py* naopak vytvorí rekurentnú neurónovú sieť aj s najlepším modelom získaným v sekcii 4.4.

Výpis na štandardný výstup sa mi pre učenie sietí opäť nepodarilo potlačiť, takže štandardný výstup bude plný týchto výpisov z učenia, ale po konci každého cyklu učenia a vyhodnocovania vypíše trénovaciu chybu a trénovaciu úspešnosť, rovnako ako testovaciu úspešnosť, úspešnosť v zápasoch bez favorita a zisk v týchto zápasoch. Na konci vypíše priemernú trénovaciu a testovaciu úspešnosť, spolu aj s priemernou úspešnosťou v zápasoch bez jasného favorita a zisk pri uzatváraní stávok na tieto zápasy. Najdôležitejšie údaje pre nás sa opäť zhrnú do logu. Log obsahuje rôzne údaje od názvu ligy, selektovaných príznakov, cez architektúru siete, až po trénovaciu chybu a trénovaciu a testovaciu úspešnosť, úspešnosť v zápasoch bez favorita a zisk v týchto zápasoch pre každý cyklus učenia a vyhodnocovania a na konci priemerné hodnoty týchto údajov. Log sa vytvorí v zložke *Logs* a jeho meno bude obsahovať skratku „ATP“, a časovú známku vo formáte rok, mesiac, deň, hodina, minúta a sekunda zapnutia finálneho programu. Na začiatku názov obsahuje skratku „RNN“, ak sa vytvárala rekurentná neurónová sieť podobne ako v predchádzajúcej sekcii.

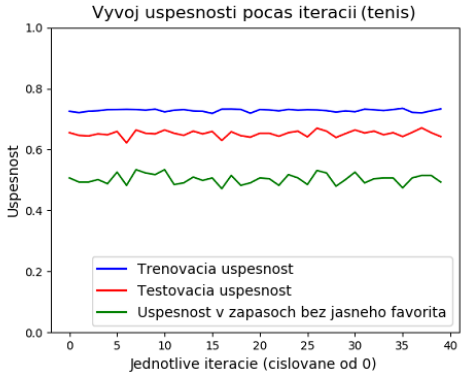
6. Výsledky

6.1 Dopredná neurónová sieť

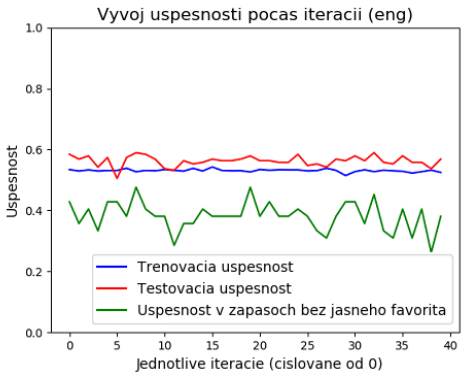
Výsledky vyhodnocovania môžeme nájsť v tabuľke 6.1.

Šport (liga)	Tr %	Te %	CG %	CGP	CGP/M
Futbal (ENG)	53,09	56,24	38,15	1,35	0,0321
Futbal (GER)	50,16	53,33	38,68	−0,32	−0,0089
Futbal (SPA)	54,24	50,76	26,86	−14,24	−0,2792
Tenis	72,8	65,2	50,82	−10,57	−0,0503

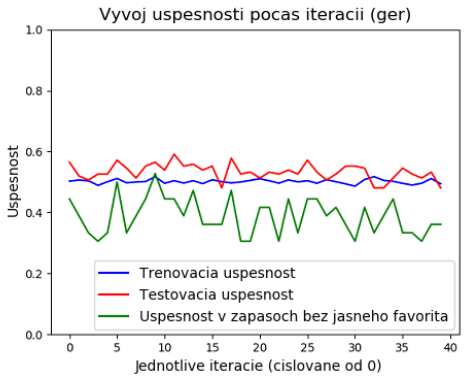
Tabuľka 6.1: Priemerné výsledky vyhodnocovania cez 40 iterácií. Skratka Tr % znamená trénovaciu úspešnosť, Te % testovaciu úspešnosť. CG značí úspešnosť pri tipovaní zápasov bez favorita, CGP zisk pri uzatváraní stávok na tieto zápasy a CGP/M značí priemerný takýto zisk na zápas bez jasného favorita..



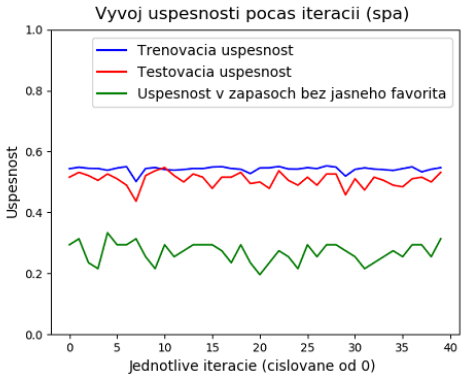
(a) Tenis



(b) Futbal (anglická *Premier league*)



(c) Futbal (nemecká *Bundesliga*)



(d) Futbal (španielska *La Liga*)

Obr. 6.1: Výsledné trénovacie, testovacie úspešnosti a úspešnosti pri predikovaní zápasov bez jasného favorita

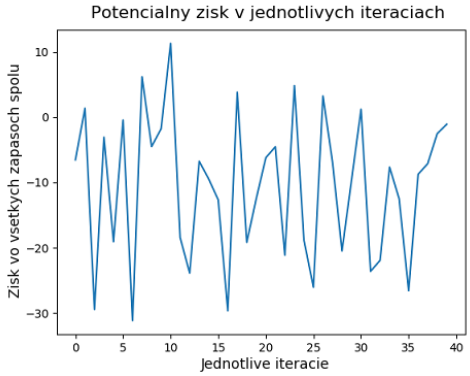
Ak výsledky porovnáme s výsledkami získanými z kapitoly 4, kde sme predikovali sezóny pred touto, tak vidíme, že nám sieť vydala iné výsledky. V prípade

tenisu sa to prejavilo mierne horšími hodnotami v sledovaných oblastiach, ale pri zápasoch bez jasného favorita došlo k výraznejšiemu poklesu.

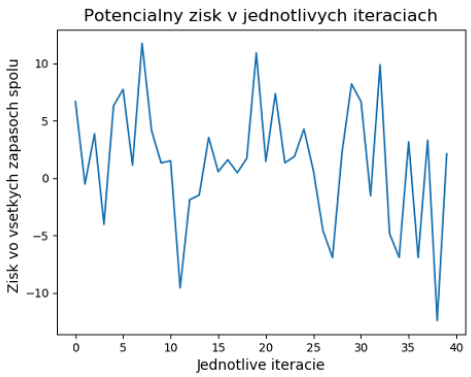
V prípade futbalu pri španielskej lige ako jedinej došlo k poklesu testovacej úspešnosti a obrovský neúspech pre zápasy bez favorita, kde sieť zvolila správny výsledok len pri 26,86 % zápasov, čo je dokonca menej ako priemerná úspešnosť pri náhodnom zvolení výsledkov alebo tipovaním len domáceho tímu (obe sú viac ako 33 %). V jednej sezóne sme dokonca dosiahli stratu skoro 25 jednotiek pri tipovaní 51 zápasov a úspešnosti pod 20 %, čo značí správne uhádnutých len 10 z 51 zápasov, ktoré v danej sezóne ligy nemali jasného favorita.

Na druhej strane nemecká aj anglická liga sa vyhodnocovali sieti o dosť lepšie ako pri trénovaní, v prípade nemeckej ligy bola testovacia úspešnosť vyššia o 2,5 % a v prípade anglickej dokonca o viac ako 5 %. Zaujímavosťou je, že úspešnosť v zápasoch bez favorita je na podobnej úrovni ako pri testovaní a celkový zisk sa pohybuje pri oboch ligách okolo 0.

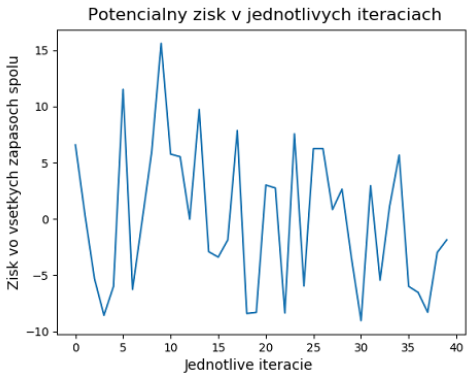
Na obrázkoch 6.1 a 6.2 môžeme vidieť, ako sa vyvíjala trénovacia a testovacia úspešnosť a úspešnosť v zápasoch bez favorita a aký to malo dopad na celkový zisk. Konkrétne na obrázku 6.1 a jeho častiach môžeme vidieť, že jednotlivé úspešnosti boli celkom konzistentné a zatiaľ čo celkové zisky zo zápasov bez favorita sa pohybovali chaotickejšie.



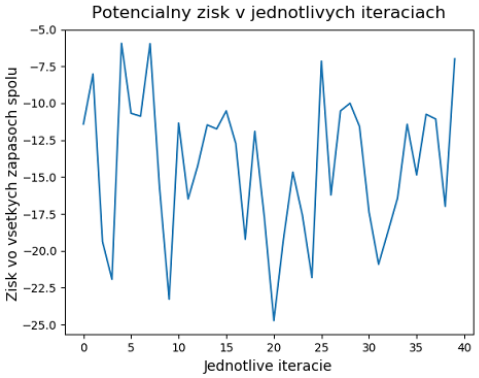
(a) Tenis



(b) Futbal (anglická *Premier league*)



(c) Futbal (nemecká *Bundesliga*)

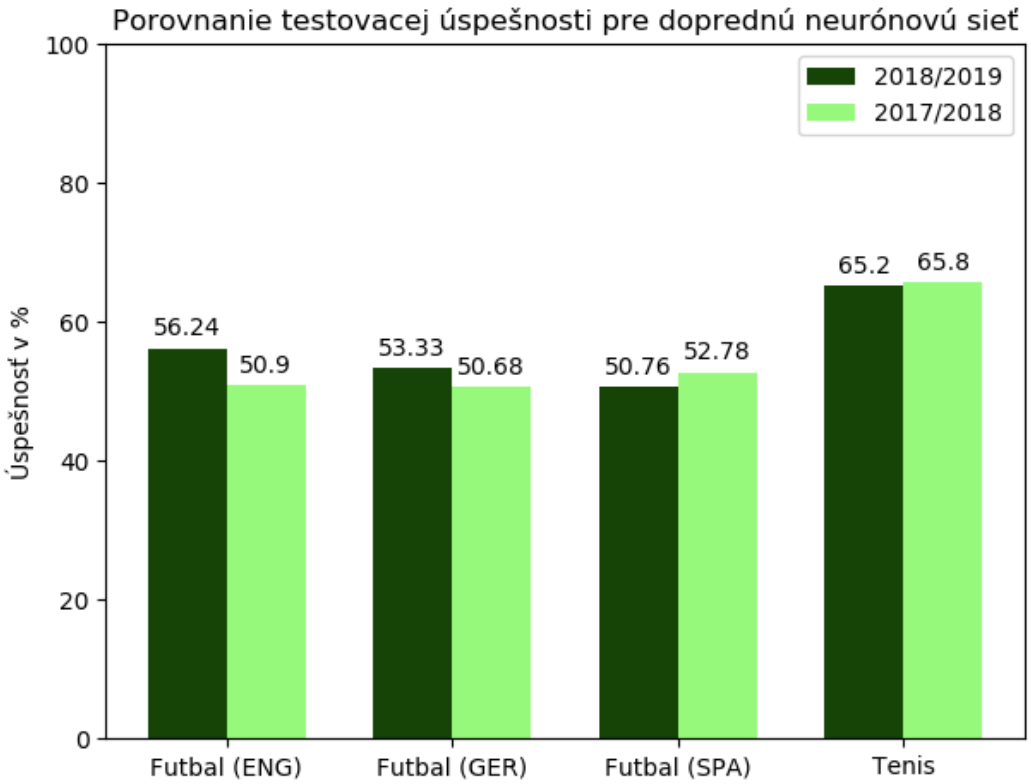


(d) Futbal (španielska *La Liga*)

Obr. 6.2: Výsledný zisk pri predikovaní zápasov bez jasného favorita (každý šport aj liga mali rôzny počet takýchto zápasov, tento obrázok ukazuje celkový zisk)

Na obrázku 6.3 môžeme vidieť testovacie úspešnosti pri predikovaní dvoch sezón, tmavšou farbou sú výsledky z tejto kapitoly, svetlejšou výsledky z kapitoly

4, teda výsledky z testovania a optimalizovania siete.



Obr. 6.3: Porovnanie testovacej úspešnosti predikovaní poslednej a predposlednej sezóny. Úspešnosti z predposlednej sezóny (svetlá farba) prebehli pokusom o optimalizáciu a dané dáta aj model použitý pri predikovaní poslednej sezóny (tmavá farba)

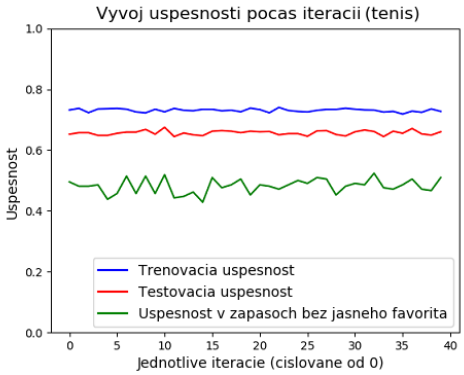
6.2 Rekurentná neurónová sieť

Výsledky vyhodnocovania môžeme nájsť v tabuľke 6.2.

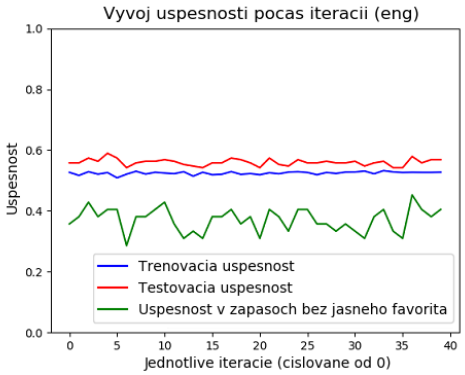
Šport (liga)	Tr %	Te %	CG %	CGP	CGP/M
Futbal (ENG)	52,44	56	37	−0,12	−0,0029
Futbal (GER)	49,88	50,54	38,68	−0,1	−0,0028
Futbal (SPA)	53,98	51,51	28,87	−12,44	−0,2439
Tenis	73,07	65,67	48,14	−22,34	−0,1064

Tabuľka 6.2: Priemerné výsledky vyhodnocovania cez 40 iterácií. Skratka Tr % znamená trénovaciu úspešnosť, Te % testovaciu úspešnosť. CG značí úspešnosť pri tipovaní zápasov bez favorita, CGP zisk pri uzatváraní stávok na tieto zápasy a CGP/M značí priemerný takýto zisk na zápas bez jasného favorita.

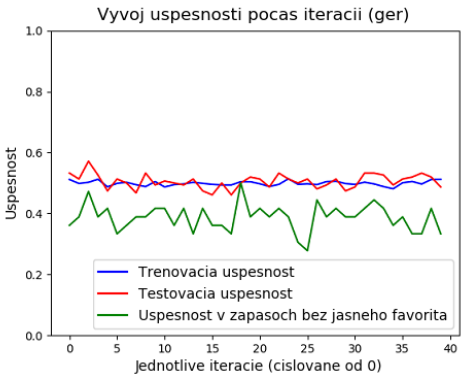
Ak tieto výsledky porovnáme s výsledkami získanými pri rovnakej architektúre siete, ale pri vynechaní predposlednej sezóny z trénovacích dát a predikovaní



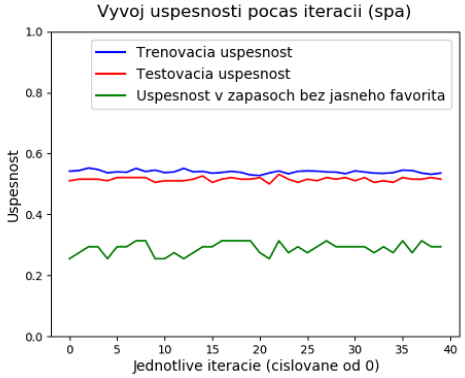
(a) Tenis



(b) Futbal (anglická *Premier league*)



(c) Futbal (nemecká *Bundesliga*)



(d) Futbal (španielska *La Liga*)

Obr. 6.4: Výsledné trénovacie, testovacie úspešnosti a úspešnosti pri predikovaní zápasov bez jasného favorita

tejto sezóny (sekcia 4.4), tak sa dozvieme, že nie sú až tak odlišné, s výnimkou anglickej futbalovej *Premier league*, kde nastalo zlepšenie z 51,95 % na 56 %. Toto zlepšenie ale nastalo na úkor úspešnosti v zápasoch bez jasného favorita, kde pri trénovaní sa dosiahla úspešnosť 39,2 % a celkový zisk 2,78 jednotky.

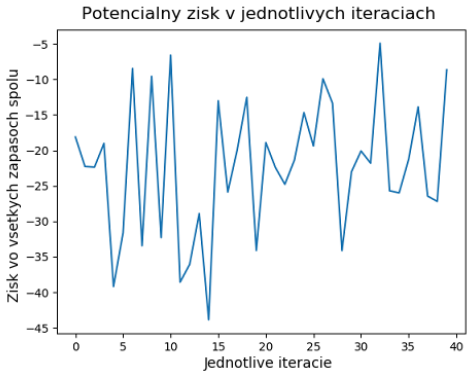
Najvýraznejší pokles oproti trénovaniu nastal v prípade španielskej *La Ligy*, kde pri optimalizovaní siete bola testovacia úspešnosť 53,24 %, pri dátach z poslednej ukončenej sezóny dosiahla v priemere sieť úspešnosť 51,51 % a pri celkovom zisku pri hypotetickom uzatváraní stávok na zápasy bez favorita dopadla najhoršie z futbalových líg. Horšie dopadol tenis, ten ale mal väčší počet zápasov bez jasného favorita. Opäť treba dodať, že úspešnosť španielskej ligy v zápasoch bez favorita bola nižšia ako priemerná náhodná úspešnosť.

V prípade nemeckej *Bundesligy* dopadla predikcia poslednej sezóny o niečo horšie ako v prípade tej predposlednej, všetky výsledky sú porovnateľné.

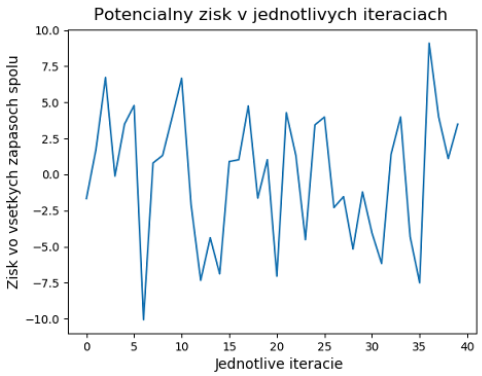
Pre tenis dopadla testovacia aj trénovacia úspešnosť mierne horšie, veľký pokles bol v úspešnosti predikcie zápasov bez favorita, kde prišiel pokles z 52 % na 48,14 % a to sa prejavilo aj na celkovom zisku v zápasoch bez favorita.

Na obrázkoch 6.4 a 6.5 môžeme vidieť, ako sa vyvíjala trénovacia a testovacia úspešnosť a úspešnosť v zápasoch bez favorita a aký to malo dopad na celkový zisk. Konkrétne na obrázku 6.4 a jeho podobrázkoch môžeme vidieť, že jednotlivé úspešnosti boli celkom konzistentné (s výnimkou úspešnosti v zápasoch bez favorita, kde sa to mierne kolísalo), zatiaľ čo celkové zisky zo zápasov bez favorita

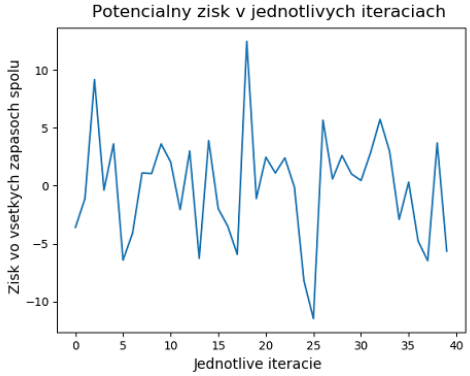
(obrázok 6.5) sa pohybovali chaotickejšie.



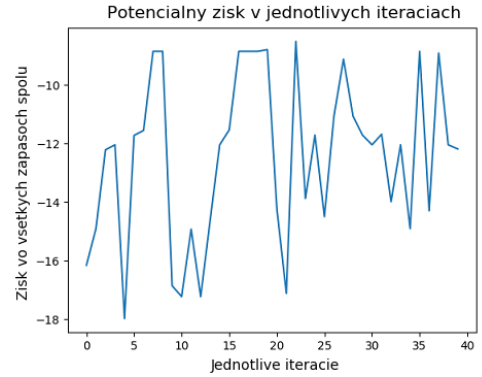
(a) Tenis



(b) Futbal (anglická *Premier league*)



(c) Futbal (nemecká *Bundesliga*)



(d) Futbal (španielska *La Liga*)

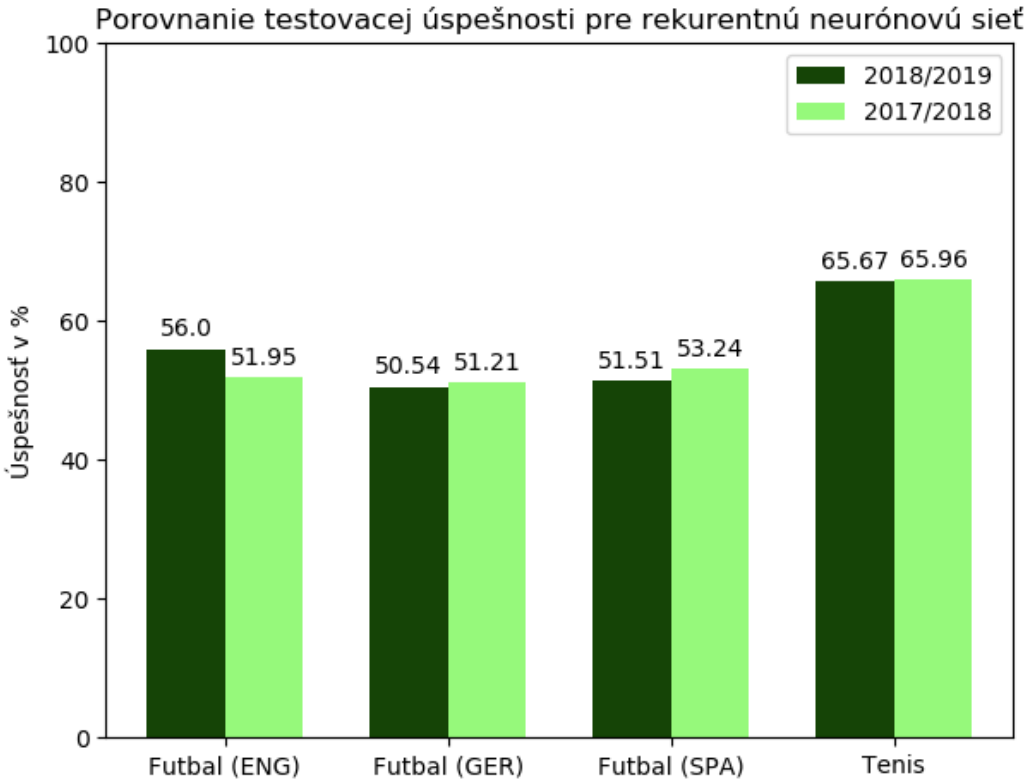
Obr. 6.5: Výsledný zisk pri predikovaní zápasov bez jasného favorita (každý šport aj liga mali rôzny počet takýchto zápasov, tento obrázok ukazuje celkový zisk)

Na obrázku 6.6 môžeme vidieť testovacie úspešnosti pri predikovaní dvoch sezón, tmavšou farbou sú výsledky z tejto kapitoly, svetlejšou výsledky z kapitoly 4, teda výsledky z testovania a optimalizovania siete.

6.3 Porovnanie

Výsledky dosiahnuté v tejto kapitole boli odlišné, ako výsledky dosiahnuté v kapitole 4. Videli sme, o koľko sa tieto výsledky líšili. Môže to značiť mnohé veci, najskôr to ale značí fakt, že každá sezóna má vlastnú predvídateľnosť a aj keď dokážeme sieť naučiť predpovedať jednu sezónu, neznamená to, že rovnaká architektúra bude úspešná aj pri inej sezóne, nie to ešte inej lige alebo športe. V našom prípade sa nám podarilo poraziť stávkové kancelárie len v prípade doprednej neurónovej siete na predikciu anglickej *Premier League*, aj to len o 1,35 jednotky, čo v preklade znamená, že ak by sme celú sezónu uzatvárali stávky na zápasy bez favorita tak, ako by nám to predpovedala naša sieť a na každý zápas by sme vsadili rovnakú čiastku, tak by sme na konci ostali v zisku, ktorý by sa rovnal 1,35-násobku jedného vkladu.

Na vychýlenie úspešnosti siete vo futbalovej lige zo sezóny na sezónu stačí, ak jeden z dvoch najlepších hráčov sveta prestúpi medzi sezónou do inej ligy a



Obr. 6.6: Porovnanie testovacej úspešnosti predikovaní poslednej a predposlednej sezóny. Úspešnosti z predposlednej sezóny (svetlá farba) prebehli pokusom o optimalizáciu a dané dáta aj model použitý pri predikovaní poslednej sezóny (tmavá farba)

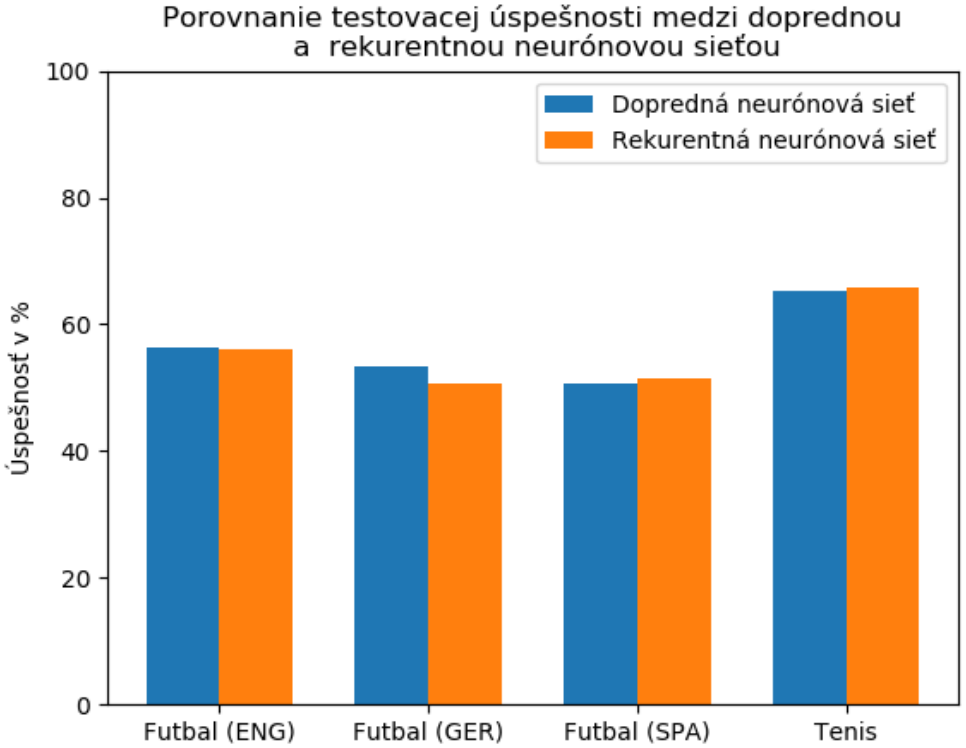
jeho tím nebude bez neho dosahovať rovnaké výsledky (Madu, 2018). Zdá sa, že obe siete sa v tomto prípade, minimálne zo začiatku sezóny opierali o údaje o dlhodobej sile tímov (príznaky 41 a 42 v Prílohe A.1), čo mohlo viesť k miernemu poklesu úspešnosti tak, ako sme videli. V anglickej lige sme zaznamenali vzostup aj v trénovacej a aj v testovacej úspešnosti, čo potvrdzuje teóriu o rozličnej predvídateľnosti rozličných sezón.

Zo získaných údajov sme zistili mimo iného aj to, ako sa zmenila predvídateľnosť jednotlivých športových výsledkov z roka na rok (sezóny na sezónu). Môžeme povedať, že predikovaná sezóna anglickej *Premier League* bola ľahšie predvídateľná ako tá minulá. Na druhej strane horšie sa obom typom sietí predikovala španielska *La Liga*. Tieto poznatky môžeme zformulovať do pár hypotéz o rozdiele medzi doprednou a rekurentnou neurónovou sieťou.

Ak je sezóna výrazne horšie predvídateľná ako sezóna, z ktorej sú data, tak oba modely prezentované v tejto práci vykazujú horšie výsledky. Rekurentná neurónová sieť vykazuje mierne lepšie výsledky, LSTM neuróny aj s ich implementáciou krátkodobej pamäte pomáhajú v tomto ohľade sieti.

Naopak, ak je sezóna lepšie predvídateľná ako predchádzajúce, tak obe modely dosahujú vyššej testovacej ako trénovacej úspešnosti. Dopredná neurónová sieť dosahuje o niečo lepšie výsledky, čo môže znamenať, že pamäť rekurentnej neurónovej siete núti túto sieť rozmýšľať konzervatívnejšie.

Celkovo ale sú výsledky veľmi podobné na to, aby sa dali s určitosťou vysloviť



Obr. 6.7: Porovnanie testovacej úspešnosti modelov doprednej (modrá farba) a rekurentnej (oranžová) neurónovej siete pri predikcii výsledkov jednotlivých líg

nejaké tvrdenia o prezentovaných typoch neurónových sietí.

Na obrázku 6.7 môžeme vidieť rozdiel v úspešnosti medzi najlepšími modelmi doprednej a rekurentnej neurónovej siete podľa kapitoly 4 pri predikcii vyhodnocovanej sezóny (sezóny 2018/2019).

Záver

Jedným z dôvodov, prečo ľudia sledujú šport je ten, že šport je nepredvídateľný. Na druhej strane, ľudstvo sa učí od nepamäti získavať kontrolu nad nepredvídateľným a ak sa to nepodarí, tak aspoň to s nejakou určitosťou predvídať (napríklad zatmenie slnka). Vedieť predvídať športové výsledky by nám zaručilo nielen slávu a pozornosť médií, ale mohlo by nám to pred týmto všetkým zaručiť aj bohatstvo, pretože stávkové kancelárie dávajú možnosť zarobiť na správnom tipovaní (v našom prípade predpovedaní) výsledkov.

Jedným z relatívne nových spôsobov, ktorý sa začína používať na predpovedanie športových výsledkov je strojové učenie. Jedným z najznámejších a momentálne aj naobľúbenejších typov strojového učenia sú neurónové siete. V tejto práci sme sa zamerali na dva druhy neurónových sietí, a to dopredné a rekurentné neurónové siete a ich úspešnosť pri predikcii výsledkov v troch futbalových ligách a na najvyšších tenisových turnajoch medzi najlepšími hráčmi.

Používali sme len dáta, ktoré vieme vyčítať z výsledkov (v prípade tenisu aj z koncoročných rebríčkov), ale nepodarilo sa nám ani priblížiť výsledkom iných autorov publikujúcich v tejto oblasti, ktorý používali aj fakticky nepodložené abstraktné dáta od expertov v oblasti. V prípade futbalu sme dosiahli úspešnosť od 50,54 – 56,24 %. V prípade tenisu sme dosiahli úspešnosť okolo 65,5 %.

Jednou z odlišností tejto práce od ostatných v oblasti bolo zameranie sa aj na zápasy bez jasného favorita stávkových kancelárií. Aj napriek tomu, že to bol celkom nový prístup, nepodarilo sa nám dosiahnúť požadované výsledky, ktoré by mohli podnietiť väčší výskum týmto smerom. To ale neznamená, že sa lepšie výsledky dosiahnuť nedajú. Znamená to iba, že sa zo získaných dát ukázaným spôsobom nebude dať vytvoriť neurónová ani rekurentná neurónová sieť, ktorá by dosahovala stabilné výsledky a dokázala by pravidelne a dlhodobo poraziť stávkové kancelárie a vykázať teda nejaký zisk.

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.

Bengio, Y., Simard, P., Frasconi, P. a kol. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, **5**(2), 157–166.

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].

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Brownlee, J. (2018). How to avoid overfitting in deep learning neural networks. URL `https://machinelearningmastery.com/introduction-to-regularization-to-reduce-overfitting-and-improve-generalizat` [cit. 2019-07-13].

Domingos, P. M. (2012). A few useful things to know about machine learning. *Commun. acm*, **55**(10), 78–87.

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].

Hochreiter, S. a Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

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.

Kingma, D. P. a Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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 (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].

Madu, Z. (2018). The calculus of cristiano ronaldo. URL `https://www.sbnation.com/soccer/2018/9/27/17861974/cristiano-ronaldo-juventus-real-madrid`. [cit. 2019-07-17].

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].

McCulloch, W. S. a Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**(4), 115–133.

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].

Olah, C. (2015). Understanding LSTM Networks. URL `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`. [cit. 2019-05-20].

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.

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M. a Verri, A. (2004). Are loss functions all the same? *Neural Computation*, **16**(5), 1063–1076.

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.

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 dátasete),

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 dátasete),

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. H - home - 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. A - away - 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 dáta; tie sú od sezóny 2003, najstaršie trénovacie dáta 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.