VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

NÁZEV PRÁCE

THESIS TITLE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **JMÉNO PŘÍJMENÍ**
AUTHOR

VEDOUCÍ PRÁCE **Doc. RNDr. JMÉNO PŘÍJMENÍ, Ph.D.**
SUPERVISOR

BRNO 2018

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Citace

PŘÍJMENÍ, Jméno. *Název práce*. Brno, 2018. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Doc. RNDr. Jméno Příjmení, Ph.D.

Název práce

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jméno Příjmení
23. ledna 2018

Poděkování

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant, apod.).

Obsah

Kapitola 1

Úvod

Abychom mohli napsat odborný text jasně a srozumitelně, musíme splnit několik základních předpokladů:

- Musíme mít co říci,
- musíme vědět, komu to chceme říci,
- musíme si dokonale promyslet obsah,
- musíme psát strukturovaně.

Tyto a další pokyny jsou dostupné též na školních internetových stránkách [12].

Přehled základů typografie a tvorby dokumentů s využitím systému LaTeXje uveden v [11].

1.1 Musíme mít co říci

Dalším důležitým předpokladem dobrého psaní je *psát pro někoho*. Píšeme-li si poznámky sami pro sebe, píšeme je jinak než výzkumnou zprávu, článek, diplomovou práci, knihu nebo dopis. Podle předpokládaného čtenáře se rozhodneme pro způsob psaní, rozsah informace a míru detailů.

1.2 Musíme vědět, komu to chceme říci

Dalším důležitým předpokladem dobrého psaní je psát pro někoho. Píšeme-li si poznámky sami pro sebe, píšeme je jinak než výzkumnou zprávu, článek, diplomovou práci, knihu nebo dopis. Podle předpokládaného čtenáře se rozhodneme pro způsob psaní, rozsah informace a míru detailů.

1.3 Musíme si dokonale promyslet obsah

Musíme si dokonale promyslet a sestavit obsah sdělení a vytvořit pořadí, v jakém chceme čtenáři své myšlenky prezentovat. Jakmile víme, co chceme říci a komu, musíme si rozvrhnout látku. Ideální je takové rozvržení, které tvoří logicky přesný a psychologicky stravitelný celek, ve kterém je pro všechno místo a jehož jednotlivé části do sebe přesně zapadají. Jsou jasné všechny souvislosti a je zřejmé, co kam patří.

Abychom tohoto cíle dosáhli, musíme pečlivě organizovat látku. Rozhodneme, co budou hlavní kapitoly, co podkapitoly a jaké jsou mezi nimi vztahy. Diagramem takové organizace je graf, který je velmi podobný stromu, ale ne řetězci. Při organizaci látky je stejně důležitá otázka, co do osnovy zahrnout, jako otázka, co z ní vypustit. Příliš mnoho podrobností může čtenáře právě tak odradit jako žádné detaily.

Výsledkem této etapy je osnova textu, kterou tvoří sled hlavních myšlenek a mezi ně zařazené detaily.

1.4 Musíme psát strukturovaně

Musíme začít psát strukturovaně a současně pracujeme na co nejsrozumitelnější formě, včetně dobrého slohu a dokonalého značení. Máme-li tedy myšlenku, představu o budoucím čtenáři, cíl a osnovu textu, můžeme začít psát. Při psaní prvního konceptu se snažíme zaznamenat všechny své myšlenky a názory vztahující se k jednotlivým kapitolám a podkapitolám. Každou myšlenku musíme vysvětlit, popsat a prokázat. Hlavní myšlenku má vždy vyjadřovat hlavní věta a nikoliv věta vedlejší.

I k procesu psaní textu přistupujeme strukturovaně. Současně s tím, jak si ujasňujeme strukturu písemné práce, vytváříme kostru textu, kterou postupně doplňujeme. Využíváme ty prostředky DTP programu, které podporují strukturovanou stavbu textu (předdefinované typy pro nadpisy a bloky textu).

Kapitola 2

Proteíny

Proteíny (bielkoviny) môžeme charakterizovať ako základné stavebné prvky všetkých živých organizmov. Nie sú však iba stavebnými prvkami bunky, zabezpečujú väčšinu bunečných funkcií. Pochopenie procesu vzniku proteínov a ich funkcie nachádza široké uplatnenie v odvetviach ako medicína, poľnohospodárstvo, priemysel a mnohé ďalšie. V tejto kapitole sa budem zaoberať základným rozdelením proteínov, procesom ich vzniku z DNA a štruktúrou.

2.1 Základné rozdelenie proteínov

Proteíny sú biopolyméry tvorené z jedného alebo viacerých polypeptidových reťazcov, ktoré je možné chápať ako sekvenciu polymérov aminokyselín navzájom spojených kovalentnou peptidovou väzbou. Proteíny sa skladajú do množstva komplikovaných tvarov a ich funkcie súvisia s konkrétnym priestorovým usporiadaním (konformáciou), pričom sa snažia zaujať čo najlepšiu konformáciu z energetického hľadiska. Konformácia vychádza z primárnej štruktúry, ktorú môžeme chápať ako reťazec aminokyselín v danom poradí [1]. Podľa funkcie môžeme proteíny rozdeliť do niekoľkých skupín [1]:

- **Enzýmy:** ich funkciou je katalýza rozpadu a tvorba kovalentných väzieb. Príkladom môže byť napríklad pepsín, ktorý sa podieľa na odbúraní bielkovín pri trávení.
- **Štruktúrne proteíny:** tvoria základné stavebné jednotky buniek a tkanív, poskytujú im mechanickú oporu. Príkladom je keratín tvoriaci základnú zložku vlasov a nechtov.
- **Transportné proteíny:** prenášajú malé molekuly a ionty v organizme. Príkladom sú proteín hemoglobín ako nosič kyslíka v krvnom obehu a proteín transferrin prenášajúci železo.
- **Pohybové proteíny:** sú pôvodcami pohybu buniek a tkanív. Príkladom takýchto proteínov sú kinesin a myosin.
- **Zásobné proteíny:** slúžia na skladovanie malých molekúl a iontov. Kasein v mlieku poskytuje zdroj aminokyselín pre novonarodené živočíchy.
- **Signálne proteíny:** ich funkcoiu je prenos informačných signálov medzi bunkami. Do tejto skupiny patrí známy proteín inzulín regulujúci hladinu cukru v krvi.
- a ďalšie.

2.2 Aminokyseliny

Aminokyseliny sú odvodené od organických kyselín a predstavujú rôzne triedy molekúl s jednou spoločnou vlastnosťou, všetky vlastnia karboxylovú (COOH) a aminovú (NH_2) skupinu. Tieto skupiny sú naviazané k jednému uhlíkovému atómu, ktorý je označovaný ako α uhlík. Rôznorodosť jednotlivých aminokyselín spočíva v postrannom reťazci (R) určujúcom chemické vlastnosti aminokyselín, resp. proteínov. Jednotlivé aminokyseliny sú v proteínovej molekule vzájomne spojené peptidovou väzbou, ktorá prepojuje karboxylovú skupinu jednej aminokyseliny s amino skupinou druhej. Reťazec viacerých aminokyselín je označovaný ako peptidový reťazec (polypeptid). Celkovo existuje 20 rôznych aminokyselín, ktoré môžeme na základe chemických vlastností postranných reťazcov rozdeliť na šesť základných skupín [15]:

- **Aminokyseliny s alifatickým postranným reťazcom:** alanin (Ala), valin (Val), leucin (Leu), isoleucin (Ile), glycin (Gly)
- **Bazické skupiny s aminovou skupinou na postrannom reťazci:** arginin (Arg), lysin (Lys)
- **S aromatickým jadrom alebo hydroxylovou skupinou na postrannom reťazci:** histidin (His), fenylalanin (Phe), serin (Ser), threonin (Thr), tyrosin (Tyr), tryptofan (Trp)
- **Kyslé skupiny s karboxylovou alebo aminovou skupinou na postrannom reťazci:** kyselina asparagová (Asp), asparagin (Asn), kyselina glutamová (Glu), glutamin (Gln)
- **So sírou v postrannom reťazci:** methionin (Met), cystein (Cys)
- **Obsahujúce sekundárny amin:** prolin (Pro)

2.3 Syntéza proteínov

Proteíny vznikajú z DNA v procese nazývanom proteosyntéza. Tento proces sa skladá z 2 hlavných častí, ktorými sú transkripcia a translácia.

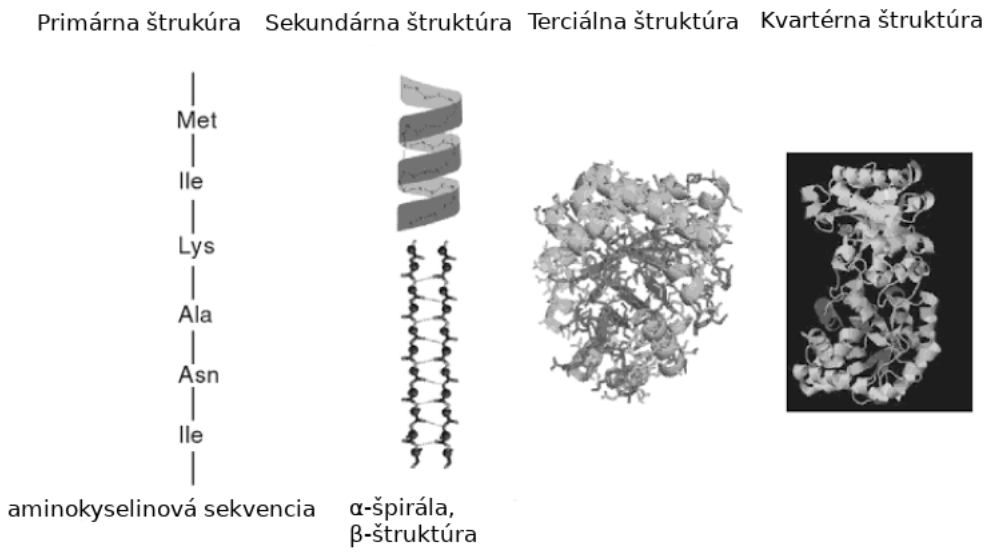
- **Transkripcia:** pri procese transkripcie dochádza k prepisu časti nukleotidovaj sekvencie DNA (génu) do molekuly RNA. Dôležitú úlohu zohráva enzým RNA-polymeráza, ktorá musí pred začiatkom traskripcie nájsť oblasť tzv. promotoru obsahujúcu informáciu o začiatku transkripcie a následne sa na túto oblasť naviazať. Proces prepisu končí keď RNA-polymeráza narazí na sekvenciu tzv. terminátoru. Výsledná molekula RNA sa označuje ako mediátorová RNA (mRNA).
- **Translácia:** pri procese translácie dochádza k prenosu informácie z mRNA do polypeptidového reťazca aminokyselín. Sekvencia nukleotidov RNA sa postupne číta po trojiciach (tzv. kodónoch), pričom každý kodón je preložený na jednu z dvadsiatich aminokyselín. Trojica nukleotidov umožňuje vytvoriť 64 možných kombinácií, takže jedna aminokyselina môže byť reprezentovaná viacerými kodónmi. Výsledkom translácie je proteín.

2.4 Štruktúra proteínov

Popis trojrozmernej štruktúry proteínov môžeme podľa [15] rozdeliť do štyroch úrovní organizácie:

- **Primárna štruktúra:** sekvencia aminokyselín v polypeptidovom reťazci
- **Sekundárna štruktúra:** zachytáva elementy, ktoré na krátkych úsekoch v sekvencií proteínu zaujímajú podobnú konformáciu. Ide najmä o α-špirálu (α-helix) a β- štruktúra alebo (β-skladaný list). α-špirála je také priestorové usporiadanie, kedy reťazec vytvára špirálu. Táto konformácia je stabilizovaná vodíkovými mostíkmi medzi peptidovými väzbami ležiacimi nad sebou [**?**]. V prípade β-štruktúry prebiehajú úseky reťazca paralelne vedľa seba a sú stabilizované vodíkovými mostíkmi medzi susediacimi úsekmi.
- **Terciálna štruktúra:** reprezentuje trojrozmerné priestorové usporiadanie zloženého polypeptidového reťazca [15]. Na podobe výslednej terciálnej štruktúry majú vplyv chemické vlastnosti aminokyselín a ich usporiadanie v reťazci.
- **Kvartérna štruktúra:** popisuje usporiadanie jednotlivých polypeptidových reťazcov v molekule proteínu. Týka sa to však iba tzv. oligomerných proteínov, ktoré sú tvorené z viac ako jedného polypetidového reťazca.

  Primárnu, sekundárnu, terciálnu a kvartérnu štruktúru je možné vidieť na obrázku 2.1:



Obrázek 2.1: Primárna, sekundárna, terciálna a kvartérna štruktúra. Prevzaté a upravené z [6].

Kapitola 3

Vplyv aminokyselinových substitúcií na stabilitu proteínu

Stabilita je jednou z najdôležitejších vlastností proteínu. Motivácia skúmania stability je dnes veľká, pretože táto vlastnosť proteínov je dôležitá v mnohých oblastich ako je medicína, kde chceme dosiahnuť výrobu účinnejších liečiv, v oblasti priemyslu a poľnohospodárstva. Stabilný proteín dokáže lepšie znášať nepriaznivé podmienky okolitého prostredia, akými sú vyššie teploty alebo chemické vlastnosti okolia v ktorom sa proteín nachádza. Na stabilitu proteínu však vplývajú aminokyselinové substitúcie, ktoré môžu proteín stabilizovať, ale aj destabilizovať. Preto vzniká potreba skúmať vplyv substitúcií na stabilitu. V tejto kapitole sa zmienim o stabilite proteínu, čo stabilitu určuje a o dôvodoch vzniku a vplyve mutácií.

3.1 Stabilita proteínu

Stabilita proteínu je určená množinou vzájomne pôsobiacich a ovplyvňujúcich sa síl. Tieto sily určujú, či sa proteín nachádza vo svojom pôvodnom zloženom alebo rozloženom (denaturovanom) stave. Stabilita je úzko prepojená so stavom, v ktorom sa proteín nachádza. Stabilný proteín sa nachádza v zloženom stave, ktorý je stabilizovaný rôznymi vzájomnými interakciami, kde patria hydrofóbne, elektrostatické, vodíkové väzby alebo van der Waalsove sily. Naopak, nestabilný proteín sa nachádza v denaturovanom stave, kde dominuje entropická a neentropická voľná energia. [6]

Stabilitu proteínu je možné reprezentovať ako zmenu tzv. Gibbsovej (voľnej) energie (ΔG) potrebnej na prechod proteínu zo zloženého do denaturovaného stavu alebo naopak. Gibbsova voľná energia je termodynamický potenciál vyjadrujúci maximálne množstvo reverzibilnej práce, ktorá môže byť uskutočnená termodynamickým systémom pri konštantnej teplote a tlaku. Gibbsova voľná energia je definovaná nasledujúcim vzťahom [14]:

$$G = H - TS, \tag{3.1}$$

kde H predstavuje entalpiu, T teplotu a S entropiu.

Existuje niekoľko laboratórnych metód na určenie stability ako napríklad cirkulárny dichroizmus (CD), diferencilna skenovacia kalorimetria (DSC), fluorescencia (Fl), absorpcia svetla (Abs), nukleárna magnetická rezonancia (NMR) [6].

Určenie stability bez použitia niektorej z laboratórnych metód je možné uskutočniť výpočtom jedného z existujúcich silových polí (Talaris, Score12, ...). Výpočet takéhoto poľa ukazuje nasledujúci jednoduchý príklad [10] [6]:

Voľná energia v zloženom stave je daná vzťahom

$$G_F = G_{hy} + G_{el} + G_{hb} + G_{vw} + G_{ss}, \tag{3.2}$$

kde $G_{hy}, G_{el}, G_{hb}, G_{ss}, G_{vw}$ sú hydrofóbne, elektrostatické, vodíkové, disulfidické a van der Waalsove voľné energie.

Hydrofóbnu voľnú energiu je možné odhadnúť z dostupnej rozpustnosti. Vypočítaná je podľa vzťahu

$$G_{hy} = \Sigma\Delta\sigma_i[A_i(folded) - A_i(unfolded)], \tag{3.3}$$

kde i predstavuje rôzne typy atómov, $A_i(folded)$ a $A_i(unfolded)$ sú dostupné povrchové plochy (ASA) jednotlivých typov atómov v zloženom a nezloženom stave a σ_i reprezentuje atómové solvatačné parametre.

Elektrostatickými interakciami najviac prispievajú nabité postranné reťazce reziduí Lysine, Hystodine, Arginine a kyselina asparagová a glutamová.

Vodíkové väzby sú jednými z hlavných zúčastnených pri tvorbe sekundárnej štruktúry proteínu. Výpočet ich príspevku je založený najmä na ich geometrických informáciách.

Van der Waalsove energie je možné vypočítať zjednotením Lennard-Jonesovho potenciálu ako

$$G_{vw} = (A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}), \tag{3.4}$$

kde $A_{ij} = \varepsilon_{ij}^{*}(R_{ij}^{*})^{12}$, $B = 2\varepsilon^{*}(R^{*})^{6}$, $R^{*} = (R^{*} + R^{*})$, $\varepsilon^{*} = (\varepsilon^{*}\varepsilon^{*})$. R^{*} a ε^{*} sú van der Waalsov rádius a hĺbka TODO.

Voľná energia v nezloženom stave je daná vzťahom

$$G_U = G_{en} + G_{ne}, \tag{3.5}$$

kde G_{en} a G_{ne} sú entropické a neentropické voľné energie.

3.2 Mutácie

Na stabilitu proteínu vplývajú aminokyselinové mútacie, ktoré môžu spôsobiť to, že proteín sa stane nestabilným. Preto do hlavnej oblasti skúmania stability patrí predikcia zmeny stability na základe aminokyselinovej mutácie. Jedná sa o predikciu zmeny Gibbsovej voľnej energie ($\Delta\Delta$G) medzi voľnou energiou pôvodného a zmutovaného proteínu. Môžeme ju definovať nasledujúcim vzťahom:

$$G_U = G_{mutant} - G_{original}, \tag{3.6}$$

Podľa tejto hodnoty je možné rozdeliť mutácie na stabilizujúce, neutrálne a destabilizujúce. Väčšia snaha pri predikcií môže viesť k zlepšeniu návrhu nových odolnejších proteínov alebo pri štúdií rozličných chorôb.

Mutácie sú náhodné alebo cielené zmeny v DNA. Sú naprosto nevyhnutné pre biologickú evolúciu, bez nich by sa skôr či neskôr zastavila. Ak by sa výraznejšie zmenili podmienky vonkajšieho prostredia, organizmy by bez mutácií nemuseli na zmeny zareagovať a pravdepodobne by vyhynuli. Mutáciami sú označované všetky také zmeny genetickej informácie, ktoré nie sú výsledkom segregácií a rekombinácií už existujúcich častí genotypov [16]. Podľa úrovne, na ktorej sa mutácia vyskytla, môžeme rozlišovať [5]:

- **Génové mutácie:** zmena v stavbe DNA, ktorá je reprezentovaná zmenou nukleotidovej sekvencie na určitom mieste [16]. Nazývajú sa tiež bodovými mutáciami a z hľadiska predikcie sú najzásadnejšie.
- **Chromozómové mutácie:** mení sa štruktúra chromozómu.
- **Genónové mutácie:** mení sa počet chromozómov.

3.2.1 Vznik mutácií

Mutácie nevznikajú náhodne, každá mutácia má svoju príčinu za ktoru stojí pôsobenie tzv. mutagénnych faktorov. Medzi najdôležitejšie patria chemické a fyzikálne faktory.

Medzi fyzikálne faktory patria rôzne zdroje žiarenia, najmä ionizujúce a ultrafialové. Poškodenie štruktúry DNA je priamo úmerné množstvu absorbovaného žiarenia.

Medzi chemické faktory môžeme zaradiť genotoxické látky, tzv. genotoxíny. Takýchto látok je veľké množstvo a patria medzi ne napríklad pesticídy, herbicídy, niektoré farbivá, konzervačné a dezinfekčné látky [16].

3.2.2 Typy mutácií

Podľa [5] rozlišujeme tri základné typy génových mutácií:

- **Substitúcia:** jedná sa o zámenu jedného alebo viacerých párov po sebe nasledujúcich bází inými [16]. V tomto prípade sa nemení dĺžka pôvodného proteínu. Novovziknutý proteín sa obvykle líši v jednej aminokyseline oproti pôvodnému.
- **Vloženie:** jedná sa o vloženie jedného alebo viacerých nových párov bází do pôvodnej sekvencie, spôsobuje zväčšenie dĺžky sekvencie.
- **Odstránenie:** odstránenie jedného alebo viacerých po sebe nasledujúcich párov bází, mení dĺžku sekvencie rovnako ako inzercia.

Obrázek 3.1: Jednotlivé typy mutácií[1].

V prípade, že k mutácií dôjde v kódujúcej oblasti, môžeme mutácie rozlíšiť na [5]:

- **Synonymné:** vychádzajú z tzv. degenerovanosti genetického kódu. Zámena nukleotidu v kodóne sa tak na štruktúre proteínu nemusí vôbec prejaviť a vyzerá to tak, ako keby k mutácií vôbec nedošlo.
- **Nesynonymné:** pri zmene nukleotidu v kodóne dochádza k zmene aminokyseliny a rovnako aj k zmene konformácie proteínu.
- **Posunové:** spôsobujú zmenu čítacieho rámca a často vedú k predčasnému ukončeniu prekladu proteínu.
- **Nezmyselné:** vytvárajú STOP kodón a tým spôsobujú predčasné ukončenie prekladu proteínu.

[1] Zdroj: https://www.bbc.co.uk/education/guides/zc499j6/revision/2

Kapitola 4

Strojové učenie

Strojové učenie je v súčasnej dobe chápané ako disciplína umelej inteligencie. Základnou technikou strojového učenia je prehľadávanie stavového priestoru. K charakteristickým vlastnostiam patrí využívanie znalostí, práca so symbolickými či štruktúrovanými premennými [8]. Pojem strojové učenie takisto označuje počítačové metódy pracujúce s obrovským množstvom dát, medzi ktorými je snahou nájsť vzťahy. Takéto metódy nachádzajú svoje uplatnenie pri hľadaní riešení v mnohých odvetviach akými sú počítačové videnie, rozpoznávanie reči a takisto bioinformatika. Keďže strojové učenie nachádza v mnohých odvetviach čoraz väčšie uplatnenie, je potrebné brať do úvahy jeho výhody a rovnako aj nevýhody. Medzi výhody patrí automatické hľadanie vzťahov vo veľkom množstve dát, čo by bolo pri mechanickom hľadaní takmer nemožné. Medzi hlavné nevýhody metód patrí neschopnosť správnej analýzy dát pri nevyváženosti predložených dát, nesprávne výsledky pri malom množstve tréningových dát pre metódu alebo nemožnosť práce s dátami obsahujúcimi veľké množstvo parametrov.

V tejto kapitole sa budem venovať základným technikám strojového učenia a popisom používaných algoritmov.

4.1 Úvod do strojového učenia

Podľa [2] je možné algoritmy strojového učenia rozdeliť do 3 základných skupín:

- klasifikácia
- regresia
- hľadanie asociácií

Klasifikácia rieši problém priradenia výstupnej triedy vstupným dátam, ktoré môžu byť reprezentované vektorom hodnôt. Ako príklad si je možné predstaviť zatriedenie žiadateľov o pôžičku do tried s vysokým alebo nízkym rizikom toho, že pôžičku nebudú schopní splácať na základe rôznych údajov o žiadateľoch.

Regresné metódy narozdiel od klasifikačných nepriraďujú vstupom výstupnú triedu, ale snažia sa určiť priamo číselnú hodnotu výstupu. Príkladom môže byť určenie ceny ojazdeného auta na základe parametrov ako počet najazdených kilometrov, značka, rok výroby.

Asociačné pravidlá slúžia na hľadanie zaujímavých asociácií vo veľkom objeme dát. Pri ich hľadaní nás zaujíma podmienená pravdepodobnosť, ktorá sa uvádza vo forme $P(X|Y)$, Y je produkt podmienený výskytom produktu X.

Metódy strojového učenie môžeme ďalej rozdeliť na základe spôsobu, akým sa učia. Podľa [2] ich rozdeľujeme na:

- **Učenie s učiteľom:** Pri tomto type učenia je nutné mať k dispozícií vstupné aj výstupné dáta. Cieľom je nájsť vzťahy medzi vstupom a výstupom, ktoré slúžia na naučenie metódy. Medzi algoritmy patriace do tejto skupiny radíme regresiu aj klasifikáciu.
- **Učenie bez učiteľa:** V tomto type učenie nie sú k dispozícií referenčné výstupné dáta, ale len vstupné. Snahou je nachádzať pravidelnosti vo vstupných dátach. Medzi takéto metódy patria rôzne typy zhlukovania.

4.2 Rozhodovacie stromy

Rozhodovací strom je hierarchický model so stromovou štruktúrou. Metódy tohto typu používajú učenie s učiteľom a môžeme ich použiť na klasifikáciu aj regresiu. Štruktúra stromu je tvorená z dvoch typov uzlov, vnútorných (nelistových) a listových uzlov. Každý z nelistových uzlov obsahuje testovaciu funkciu. Po vyhodnotení tejto funkcie sa vyberie nasledujúci uzol, v ktorom sa bude pokračovať. Tento proces začína v koreňovom uzle a pokračuje rekurzívne až do dosiahnutia listového uzlu. Listový uzol obsahuje označenie triedy do ktorej bude zaradený vstupný vektor alebo číselnú hodnotu.

4.2.1 Algoritmus J48

Algoritmus J48 patrí k metódam rozhodovacích stromov. Algoritmus produkuje klasifikačno-rozhodovací strom pre poskytnuté dáta rekurzívnym rozdeľovaním dát. Pri rozhodovaní je využitá stratégia depth-first. Algortimus berie do úvahy všetky možné testy, ktoré môžu rozdeliť dáta a vyberá test udávajúci najlepšiu informačnú hodnotu. Pre každý diskrétny atribút je zvážený jeden test s počtom výsledkov, ktorý zodpovedá počtu rôznych hodnôt atribútov.

4.2.2 Algoritmus Náhodný strom (Random Tree)

Pri tomto algoritme je strom náhodným stromom vytvoreným náhodne z množstva všetkých možných stromov. Každý list obsahuje k náhodných parametrov. Náhodné vytvorenie stromu v tomto kontexte znamená, že každý strom v množine stromov má rovankú šancu výberu. Kombinácia veľkého počtu náhodných stromov obvykle vedie k správnemu modelu.

4.2.3 Algoritmus Náhodný les (Random Forest)

Random Forest [4] je metóda založená na kombinácií viacerých rozhodovacích stromov. Každý strom závisí na hodnotách náhodného vektora hodnôt navzorkovaného nezávisle a s rovnakým rozložením pre všetky stromy v tzv. lese stromov. Obecne môžeme metódu popísať nasledujúcou definíciou:

Náhodný les (random forest) je klasifikátor tvorený kolekciou klasifikátorov so stromovou štruktúrou $\{h(x, \Theta_k), k = 1, ...\}$, kde $\{k\}$ sú nezávisle identicky rozdelené náhodné vektory a každý strom hlasuje jednotlivo o najpopulárnejšej triede vo vstupe x.

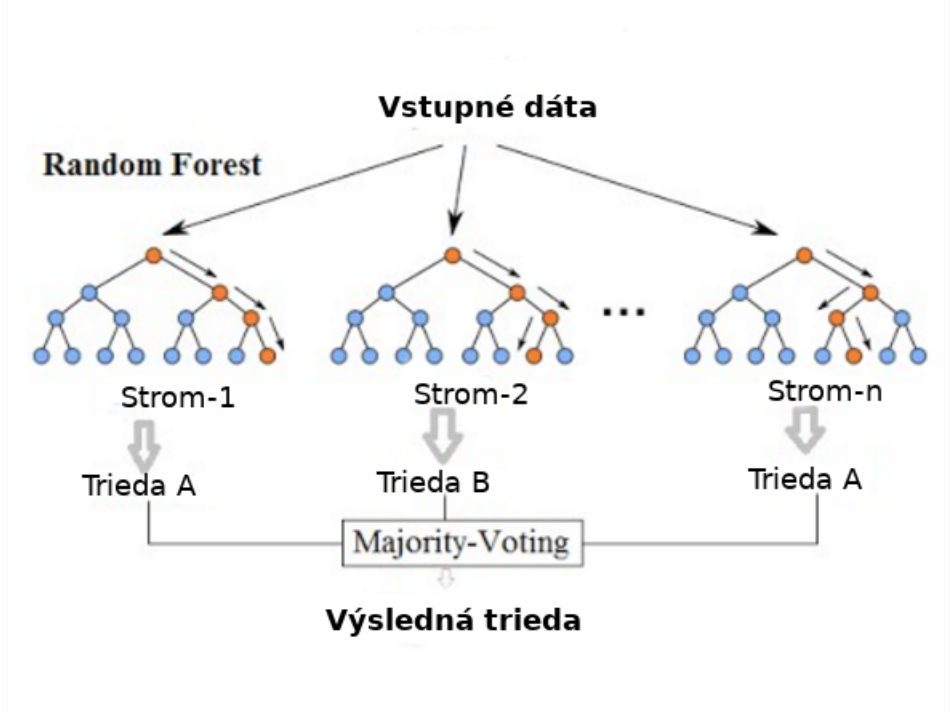
Najväčšiu pozornosť tejto metódy tvoria 3 vlastnosti:

- poskytuje presnú predikciu pre mnohé typy aplikácií
- je schopná merať dôležitosť jednotlivých parametrov pri trénovaní modelu
- blízkosť medzi vzorkami môže byť meraná trénovaným modelom

Algoritmus náhodný les pre klasifikáciu aj regresiu môžeme zjednodušene popísať nasledovne, pričom uvažujeme M rozhodovacích stromov. Schéma metódy je znázornená na obrázku 4.1:

- Pre každý z M rozhodovacích stromov vytvoríme sadu tréningových dát z originálnych dát. Na ich výber slúži metóda tzv. bagging, ktorá náhodne vyberie zadaný počet trénovacích dát.
- Pre každú množinu tréningových dát vytvoríme klasifikačný alebo regresný strom, ktorý je následne natrénovaný na M-tej náhodnej množine tréningových dát. V tejto metóde je každý uzol rozdelený najlepším rozdelením spomedzi podmnožiny prediktorov náhodne vybraných v tomto uzle. Naopak, pri klasických stromoch je uzol rozdelený na základe najlepšieho rozdelenia medzi všetkými premennými.
- Predikcia nových dát spojením výsledkov predikcie M stromov, napríklad hlasovaním väčšiny (tzv. majority voting) pri klasifikácií alebo priemerom hodnôt pri regresí.

Rozšírenie algoritmu náhodný les je momentálne veľmi aktívnou oblasťou vo výpočetnej biológií. Metóda nachádza veľké uplatnenie v bioinformatike, napríklad aj pri nástrojoch predikujúcich stabilitu proteínov.



Obrázek 4.1: Metóda Random Forest[1].

4.3 Support vector machines (SVM)

Support vector machines patria k najnovším metódam strojového učenia. Tieto metódy uskutočňujú klasifikáciu konštruovaním N-dimenzionálnej nadroviny, ktorá optimálne rozdeľuje dáta do dvoch kategórií. Cieľom je nájsť takú nadrovinu, ktorá rozdelí vstupné vektory tak, že jedna skupina vektorov je na jednej strane roviny a druhá na strane opačnej. Vektory nachádzajúce sa blízko nadroviny označujeme ako tzv. podporné vektory (support vectors).

Ak sú trénovacie dáta lineárne rozdeliteľné, potom pár $(\mathbf{w}, b)$ existuje ako
$\mathbf{w}^T\mathbf{x}_i + b \geq 1$, pre všetky $\mathbf{x}_i \in P$
$\mathbf{w}^T\mathbf{x}_i + b \leq -1$, pre všetky $\mathbf{x}_i \in N$
s rozhodovacím pravidlom daným vzťahom $f_{\mathbf{w},b}(x) = sgn(\mathbf{w}^T x + b)$, kde $\mathbf{w}$ je váhový vektor a b je odchýlka (tzv. bias).

V prípade, že dáta sú lineárne rozdeliteľné do dvoch tried, optimálnu nadrovinu je možné nájsť minimalizovaním štvorcovej normy rozdeľujúcej nadroviny. Jedná sa o konvexný kvadratický programovací problém. Pri možnosti lineárneho rozdelenia sa snaží SVM nájsť 1-dimenzionálnu nadrovinu (priamku), ktorou rozdelí skupiny vstupných vektorov. Po rozdelení dát priamkou metóda zistí vzdialenosť priamky od najbližších podporných vektorov. Táto vzdialenosť sa označuje ako tzv. krajná hranica (margin), pričom sa hľadá najväčšia vzdialenosť medzi podpornými vektormi.

Lineárne rozdeliteľné dáta sú však len ideálnym príkladom. Ak by analýza pozostávala len z premenných z dvoch kategórií, dvoch predikovaných premenných a množiny bodov rozdeliteľných priamkou, bolo by to veľmi jednoduché. V skutočnosti sa tieto metódy musia vysporiadať s viac ako dvomi predikovanými premennými, rozdelením dát nelineárnymi krivkami alebo množinami dát, ktoré nemôžeme úplne rozdeliť.

4.3.1 Jadrové funkcie

Pri väčšine reálnych problémov neexistuje lineárna nadrovina rozdeľujúca pozitívne a negatívne vzorky v tréningových dátach. Jedným z riešení je prenesenie dát do priestoru, ktorý má viac dimenzí a definovať rozdeľujúcu nadrovinu v tomto priestore. Takýto viacdimenzionálny priestor sa nazýva priestor transformovaných vlastností. S vhodne vybraným priestorom transformovaných vlastností dostatočnej dimenzie je možné rozdeliť ľubovoľnú tréningovú dátovú sadu. Mapovanie dát do iného (potencionálne nekonečného) Hilbertovho priestoru H je definované ako $\Phi : R^d \to H$. Trénovací algoritmus bude potom závisieť len na dátach skrz bodové produkty v H, napríklad na funkciách v tvare $\Phi(x_i) \cdot \Phi(x_j)$. Ak by existovala tzv. jadrová funkcia K taká, že $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, v algoritme by sme potrebovali iba K.

Jadrové funkcie sú špeciálnou triedou funkcií umožňujúce výpočet vnútorných produktov priamo v priestore vlastností bez nutnosti mapovania dát tak ako to bolo popísané vyššie. Akonáhle je vytvorená nadrovina, jadrová funkcia je použitá na mapovanie nových bodov do priestoru vlastností pre klasifikáciu.

[1] Zdroj: http://bit.ly/2rpjSDK

Výber vhodnej aproximačnej jadrovej funkcie je dôležitý, pretože funkcia definuje transformovaný priestor vlastností v ktorom budú klasifikované tréningové dáta. Medzi najpoužívanejšie jadrové funkcie patria

- $K(x, y) = (x \cdot y + 1)^P$
- $K(x, y) = e^{-\|x-y\|^2/2\sigma^2}$
- $K(x, y) = tanh(\kappa x \cdot y - \delta)^P$

4.3.2 Algoritmus SMO

Sekvenčná minimalizačná optimalizácia (SMO) je algoritmus na trénovanie SVM, ktorý jednoducho rieši kvadratický SVM problém. SMO uskutočňuje dekompozíciu celkového problému na podproblémy riešené analyticky. Metóda si v každom kroku vyberá na vyriešenie najmenší optimalizačný problém. Pre typický kvadratický SVM problém, najmenší možný optimalizačný problém zahŕňa dva Lagraengove násobitele. V každom kroku si metóda vyberia dva tieto násobitele na spoločnú optimalizáciu, nájde pre ne optimálne hodnoty a aktualizuje SVM. Výhodou tohto algoritmu je, že množstvo potrebnej pamäte pri trénovacej sade je lineárne, čo umožňuje algoritmu pracovať s veľkými tréningovými sadami.

4.4 Algoritmus Naive Bayes

Naive Bayes je klasifikačným algoritmom pre klasifikačné problémy, kde sa vyskytujú dve alebo viac tried. Je založený na Bayesovom teoréme s nezávislými predpokladmi medzi prediktormi. Klasifikátor predpokladá, že efekt hodnoty x prediktoru na danú triedu c je nezávislý na hodnotách ďalších prediktorov. Predpoklad sa nazýva triedna podmienená nezávislosť.

Tento model je jednoduchý na vytvorenie, no napriek svojej jednoduchosti klasifikátor často poskytuje dobré výsledky a je široko používaný, pretože v mnohých prípadoch prekonáva komplikovanejšie klasifikačné metódy.

Kapitola 5

Ensemble metódy

V posledných rokoch sa metódy strojového učenia využívajú vo veľkej miere v oblasti bioinformatiky. Napriek ich výhodám narážame na rôzne problémy spojené najmä s nedostatkom dát, veľkou rôznorodosťou dát alebo tzv. preučením metód. Jedným z riešení, ktoré vykazujú dobré výsledky, je využitie výstupov z viacerých klasifikátorov namiesto použitia jedného samostatného klasifikátora. Takáto kombinácia môže niekedy poskytnúť protichodné výsledky a tým pomôže zvýšiť presnosť a robustnosť predikcie. Prístup využitia viacerých klasifikátorov označujeme ako tzv. ensemble.

Podľa [9] existuje mnoho teoretických a praktických dôvodov na použitie ensemble systémov:

- **Štatistické dôvody:** Dobrý výkon metódy na trénovacích dátach nemusí predpovedať dobrú výkonnosť zovšeobecňovania. Množina klasifikátorov s podobným výkonom na trénovacích dátach môže mať rôznu zovšeobecňovaciu výkonnosť. Tento poznatok je možné vidieť najmä ak testovacie dáta určené na rozlíšenie schopnosti zovšeobecňovať nie sú dostatočne reprezentatívne. V takýchto prípadoch môže spriemerovanie výstupov niekoľkých klasifikátorov znížiť riziko zlého výberu alebo slabého klasifikátoru.

- **Veľký objem dát:** V určitých oblastiach môže nastať problém príliš veľkého objemu dát, ktorý má byť spracovaný. To nie je niekedy možné efektívne uskutočniť len pomocou jedného klasifikátora. V takýchto situáciách je vhodné zvážiť rozdelenie dát na menšie celky, trénovať klasifikátory na menších celkoch a skombinovať ich výstupy pomocou vhodného kombinačného pravidla.

- **Malý objem dát:** Dostupnosť dostatočného a reprezentatívneho mmnožstva tréningových dát je dôležitá pre klasifikačný algoritmus, aby bol schopný dosiahnuť úspešného naučenia rozdelenia dát. Pri nedostatku tréningových dát sa osvedčili techniky prevzorkovania, ktoré sú vhodné na vytvorenie náhodných prekrývajúcich sa podmnožín dostupných dát a každú je možné použiť na učenie iného klasifikátora.

- **Rozdeľuj a panuj:** Napriek dostatku dát sú niektoré problémy pre klasifikátory príliš zložité, napríklad rozhodovacia hranica rozdeľujúca dáta môže byť veľmi komplexná alebo leží mimo oblasti funkcií dostupných pre vybraný klasifikačný model. Ako príklad je možné uviesť dáta, ktoré nie je možné lineárne rozdeliť. Žiadny lineárny klasifikátor nie je schopný dáta rozdeliť, avšak vhodná kombinácia lineárnych klasifikátorov tzv. ensemble systému by bola schopná naučiť sa túto nelineárnu hra-

nicu. Dáta sú v tomto prípade rozdelené na menšie celky, pričom každý klasifikátor sa učí jednu z častí.

5.1 Tvorba ensemble systémov

Existujú dva základné spôsoby, ako vybrať členov ensemble systému: bagging a boosting.

5.1.1 Bagging

Bagging [3] alebo aj tzv. bootstrap aggregation je jednou z najstarších, ľahko implementovateľných ensemble stratégií. Bola navrhnutá na zlepšenie presnosti algoritmov strojového učenia používaných pri štatistickej klasifikácií a regresií. Rôznorodosť je dosiahnutá využitím bootstrapovaných (TODO ako to prelozit) podmnožín tréningových dát, pričom rôzne tréningové množiny sú vybrané náhodne s náhradou z celého množstva tréningových dát. Každá takáto podmnožina je určená na trénovanie iného klasifikátora rovnakého typu. Nakoniec sa využije hlasovania väčšiny na výsledkoch jednotlivých klasifikátorov. Trieda vybraná najväčším počtom klasifikátorov sa stáva konečným rozhodnutím ensemble systému.

Bagging je výkonným mechanizmom najmä pri obmedzenom množstve spoľahlivých dát. Na zaisteneie dostatočného množstva tréningových vzorkov v podmnožinách je približne 75 až 100% vzorkov prítomných v každej podmnožine. Nastáva tak veľký prekryv medzi tréningovými podmnožinami a mnoho vzorkov sa nachádza viackrát v danej podmnožine. V takomto prípade sa rôznorodosť zaistí použitím nestabilného klasifikačného modelu pre ktorý je možné získať rôzne rozhodovacie hranice s rôznymi tréningovými dátami. Techniky ako neurónové siete alebo rozhodovacie stromy sú dobrými príkladmi na tento účel, pretože ich nestabilnosť je kontrolovateľná výberom ich parametrov.

5.1.2 Boosting

Boosting [13] je ďalším algoritmom na výber členov ensemble systému. Bolo dokázané, že slabý klasifikátor (iba o niečo úspešnejší ako náhodné hádanie) je možné pretvoriť na silný klasifikátor, ktorý poskytuje správne predikcie pre všetky prípady z ľubovoľne malej časti prípadov. Rovnako ako bagging, boosting vytvára súbor klasifikátorov prevzorkovaním dát s použitím hlasovania väčšiny. Prevzorkovanie je pri boostingu zlepšené, aby každému klasifikátoru boli poskytnuté najviac informatívne trénovacie dáta. Boosting vytvára tri slabé klasifikátory:

- Prvý klasifikátor C_1 je trénovaný na náhodnej podmnožine dostupných dát.
- Tréningová podsada pre klasifikátor C_2 je vybraná ako najinformatívnejšia podsada vo vzťahu k C_1. Klasifikátor C_2 je teda trénovaný na dátach, na ktorých mal C_1 polovičnú úspešnosť.
- Tretí klasifikátor C_3 je trénovaný na dátach, kde C_1 a C_2 nesúhlasili.
- Nakoniec sú klasifikátory skombinované hlasovaním väčšiny.

5.2 Spojenie klasifikátorov

Po natrénovaní jednotlivých klasifikátorov je potrebné ich spojiť najvhodnejším spôsobom. Tieto metódy môžeme rozdeliť do dvoch kategórií: netrénovateľné a trénovateľné. Netrénovateľné sú použiteľné, ak samostatný klasifikátor poskytuje porovnateľné výsledky vo väčšine častí priestoru vlastností. Trénovateľné metódy dynamicky menia rozhodovacie pravidlá podľa špecifického typu klasifikovaného prípadu, sú užitočné v prípadoch, keď klasifikátory konštantne správne alebo nesprávne klasifikujú určité prípady.

5.2.1 Trénovateľné metódy

Medzi trénovateľné metódy patria nasledovné:

- *Stacked generalization:* Snahou je vytvoriť metaklasifikátor s využitím poznatkov o presnoti klasifikátorov. Dostupné dáta sú rozdelené na tréningovú a testovaciu sadu. Všetky nástroje sú trénované na tréningovej sade, testovacia sada je použitá na určenie výkonnosti klasifikátorov a tvorbu metaklasifikátoru

- *Rozhodcovské stromy:* Jedná sa o prístup zdola nahor prí tvorbe ensemble systému. V prvom kroku sú dáta rozdelené do konečného počtu neprekrývajúcich sa podsád a každá z nich je určená na trénovanie toho istého typu klasifikátora. Pre každý pár je vytvorený tzv. rozhodca a proces sa rekurzívne opakuje pokým nezostane len jeden klasifikátor na aktuálnej úrovni.

5.2.2 Netrénovateľné metódy

Do kategórie netrénovateľných metód môžeme zaradiť nasledujúce metódy:

- *Hlasovanie väčšiny:* Najjednoduchšia metóda, ktorá priradí objektu triedu na základe počtu hlasov jednotlivých klasifikátorov. Existujú 3 typy: jednotné hlasovanie, kde sa všetky klasifikátory musia zhodnúť na predikcií; jednoduché hlasovanie v ktorom sa aspoň polovica klasifikátorov musí zhodnúť na rozhodnutí; väčšinové hlasovanie kde je výsledok daný podľa počtu hlasov.

- *Váhované hlasovanie väčšiny:* Tento prístup je vylepšením predchádzajúceho prístupu, jednotlivé klasifikátory majú rôznu váhu v konečnom rozhodovaní a tieto váhy by mali byť prispôsobené ich presnosti.

- *Bayesova kombinácia:* Váhovaná metóda, váha spojená s klasifikátorom je posterior(TODO) pravdepodobnosť klasifikátora na tréningovej sade.

- *Váhovanie entropie:* Rovnako sa jedná o váhovanú metódu a váhy klasifikátorov sú nepriamo úmerné entropií klasifikačného vektoru.

5.3 Nástroje pre predikciu stability využívajúce strojové učenie

Kapitola 6

Typografické a jazykové zásady

Při tisku odborného textu typu *technická zpráva* (anglicky *technical report*), ke kterému patří například i text kvalifikačních prací, se často volí formát A4 a často se tiskne pouze po jedné straně papíru. V takovém případě volte levý okraj všech stránek o něco větší než pravý – v tomto místě budou papíry svázány a technologie vazby si tento požadavek vynucuje. Při vazbě s pevným hřbetem by se levý okraj měl dělat o něco širší pro tlusté svazky, protože se stránky budou hůře rozevírat a levý okraj se tak bude oku méně odhalovat.

Horní a spodní okraj volte stejně veliký, případně potištěnou část posuňte mírně nahoru (horní okraj menší než dolní). Počítejte s tím, že při vazbě budou okraje mírně oříznuty.

Pro sazbu na stránku formátu A4 je vhodné používat pro základní text písmo stupně (velikosti) 11 bodů. Volte šířku sazby 15 až 16 centimetrů a výšku 22 až 23 centimetrů (včetně případných hlaviček a patiček). Proklad mezi řádky se volí 120 procent stupně použitého základního písma, což je optimální hodnota pro rychlost čtení souvislého textu. V případě použití systému LaTeX ponecháme implicitní nastavení. Při psaní kvalifikační práce se řiďte příslušnými závaznými požadavky.

Stupeň písma u nadpisů různé úrovně volíme podle standardních typografických pravidel. Pro všechny uvedené druhy nadpisů se obvykle používá polotučné nebo tučné písmo (jednotně buď všude polotučné nebo všude tučné). Proklad se volí tak, aby se následující text běžných odstavců sázel pokud možno na *pevný rejstřík*, to znamená jakoby na linky s předem definovanou a pevnou roztečí.

Uspořádání jednotlivých částí textu musí být přehledné a logické. Je třeba odlišit názvy kapitol a podkapitol – píšeme je malými písmeny kromě velkých začátečních písmen. U jednotlivých odstavců textu odsazujeme první řádek odstavce asi o jeden až dva čtverčíky (vždy o stejnou, předem zvolenou hodnotu), tedy přibližně o dvě šířky velkého písmene M základního textu. Poslední řádek předchozího odstavce a první řádek následujícího odstavce se v takovém případě neoddělují svislou mezerou. Proklad mezi těmito řádky je stejný jako proklad mezi řádky uvnitř odstavce.

Při vkládání obrázků volte jejich rozměry tak, aby nepřesáhly oblast, do které se tiskne text (tj. okraje textu ze všech stran). Pro velké obrázky vyčleňte samostatnou stránku. Obrázky nebo tabulky o rozměrech větších než A4 umístěte do písemné zprávy formou skládanky všité do přílohy nebo vložené do záložek na zadní desce.

Obrázky i tabulky musí být pořadově očíslovány. Číslování se volí buď průběžné v rámci celého textu, nebo – což bývá praktičtější – průběžné v rámci kapitoly. V druhém případě se číslo tabulky nebo obrázku skládá z čísla kapitoly a čísla obrázku/tabulky v rámci kapitoly – čísla jsou oddělena tečkou. Čísla podkapitol nemají na číslování obrázků a tabulek žádný vliv.

Tabulky a obrázky používají své vlastní, nezávislé číselné řady. Z toho vyplývá, že v odkazech uvnitř textu musíme kromě čísla udat i informaci o tom, zda se jedná o obrázek či tabulku (například „... *viz tabulka 2.7* ...“). Dodržování této zásady je ostatně velmi přirozené.

Pro odkazy na stránky, na čísla kapitol a podkapitol, na čísla obrázků a tabulek a v dalších podobných příkladech využíváme speciálních prostředků DTP programu, které zajistí vygenerování správného čísla i v případě, že se text posune díky změnám samotného textu nebo díky úpravě parametrů sazby. Příkladem takového prostředku v systému LaTeX je odkaz na číslo odpovídající umístění značky v textu, například návěští (`\ref{navesti}` – podle umístění návěští se bude jednat o číslo kapitoly, podkapitoly, obrázku, tabulky nebo podobného číslovaného prvku), na stránku, která obsahuje danou značku (`\pageref{navesti}`), nebo na literární odkaz (`\cite{identifikator}`).

Rovnice, na které se budeme v textu odvolávat, opatříme pořadovými čísly při pravém okraji příslušného řádku. Tato pořadová čísla se píší v kulatých závorkách. Číslování rovnic může být průběžné v textu nebo v jednotlivých kapitolách.

Jste-li na pochybách při sazbě matematického textu, snažte se dodržet způsob sazby definovaný systémem LaTeX. Obsahuje-li vaše práce velké množství matematických formulí, doporučujeme dát přednost použití systému LaTeX.

Mezeru neděláme tam, kde se spojují číslice s písmeny v jedno slovo nebo v jeden znak – například *25krát*.

Členicí (interpunkční) znaménka tečka, čárka, středník, dvojtečka, otazník a vykřičník, jakož i uzavírací závorky a uvozovky se přimykají k předcházejícímu slovu bez mezery. Mezera se dělá až za nimi. To se ovšem netýká desetinné čárky (nebo desetinné tečky). Otevírací závorka a přední uvozovky se přimykají k následujícímu slovu a mezera se vynechává před nimi – (takto) a „takto“.

Pro spojovací a rozdělovací čárku a pomlčku nepoužíváme stejný znak. Pro pomlčku je vyhrazen jiný znak (delší). V systému TeX (LaTeX) se spojovací čárka zapisuje jako jeden znak „pomlčka“ (například „Brno-město“), pro sázení textu ve smyslu intervalu nebo dvojic, soupeřů a podobně se ve zdrojovém textu používá dvojice znaků „pomlčka“ (například „zápas Sparta – Slavie“; „cena 23–25 korun“), pro výrazné oddělení části věty, pro výrazné oddělení vložené věty, pro vyjádření nevyslovené myšlenky a v dalších situacích (viz Pravidla českého pravopisu) se používá nejdelší typ pomlčky, která se ve zdrojovém textu zapisuje jako trojice znaků „pomlčka“ (například „Další pojem — jakkoliv se může zdát nevýznamný — bude neformálně definován v následujícím odstavci.“). Při sazbě matematického mínus se při sazbě používá rovněž odlišný znak. V systému TeX je ve zdrojovém textu zapsán jako normální mínus (tj. znak „pomlčka“). Sazba v matematickém prostředí, kdy se vzoreček uzavírá mezi dolary, zajistí vygenerování správného výstupu.

Lomítko se píše bez mezer. Například školní rok 2008/2009.

Pravidla pro psaní zkratek jsou uvedena v Pravidlech českého pravopisu [7]. I z jiných důvodů je vhodné, abyste tuto knihu měli po ruce.

6.1 Co to je normovaná stránka?

Pojem *normovaná stránka* se vztahuje k posuzování objemu práce, nikoliv k počtu vytištěných listů. Z historického hlediska jde o počet stránek rukopisu, který se psal psacím strojem na speciální předtištěné formuláře při dodržení průměrné délky řádku 60 znaků a při 30 řádcích na stránku rukopisu. Vzhledem k zápisu korekturních značek se používalo řádkování 2 (ob jeden řádek). Tyto údaje (počet znaků na řádek, počet řádků a proklad

mezi nimi) se nijak nevztahují ke konečnému vytištěnému výsledku. Používají se pouze pro posouzení rozsahu. Jednou normovanou stránkou se tedy rozumí 60*30 = 1800 znaků. Obrázky zařazené do textu se započítávají do rozsahu písemné práce odhadem jako množství textu, které by ve výsledném dokumentu potisklo stejně velkou plochu.

Orientační rozsah práce v normostranách lze v programu Microsoft Word zjistit pomocí funkce *Počet slov* v menu *Nástroje*, když hodnotu *Znaky (včetně mezer)* vydělíte konstantou 1800. Do rozsahu práce se započítává pouze text uvedený v jádru práce. Části jako abstrakt, klíčová slova, prohlášení, obsah, literatura nebo přílohy se do rozsahu práce nepočítají. Je proto nutné nejdříve označit jádro práce a teprve pak si nechat spočítat počet znaků. Přibližný rozsah obrázků odhadnete ručně. Podobně lze postupovat i při použití OpenOffice. Při použití systému LaTeX pro sazbu je situace trochu složitější. Pro hrubý odhad počtu normostran lze využít součet velikostí zdrojových souborů práce podělený konstantou cca 2000 (normálně bychom dělili konstantou 1800, jenže ve zdrojových souborech jsou i vyznačovací příkazy, které se do rozsahu nepočítají). Pro přesnější odhad lze pak vyextrahovat holý text z PDF (např. metodou cut-and-paste nebo *Save as Text...*) a jeho velikost podělit konstantou 1800.

Kapitola 7

Závěr

Závěrečná kapitola obsahuje zhodnocení dosažených výsledků se zvlášť vyznačeným vlastním přínosem studenta. Povinně se zde objeví i zhodnocení z pohledu dalšího vývoje projektu, student uvede náměty vycházející ze zkušeností s řešeným projektem a uvede rovněž návaznosti na právě dokončené projekty.

Literatura

[1] Alberts, B.: *Základy buněčné biologie: úvod do molekulární biologie buňky*. 1998, ISBN 80-902-9062-0.

[2] Alpaydin, E.: *Introduction to Machine Learning*. MIT Press, 2010, ISBN 978-0-262-01243-0.

[3] Breiman, L.: Bagging Predictors. *Machine Learning*, 1996.

[4] Breiman, L.: Random Forests. *Machine Learning*, ročník ročník 45, 2001.

[5] Flegr, J.: *Úvod do evoluční biologie*. Academia, 2007, ISBN 978-80-200-1539-6.

[6] Gromiha, M. M.: *Protein Bioinformatics: From sequence to function*. Elsevier, 2010, ISBN 978-81-312-2297-3.

[7] Hlavsa, Z.; aj.: *Pravidla českého pravopisu*. Academia, 2005, ISBN 80-200-1327-X.

[8] Mařík, O. L. J., V.; Štěpánková: *Umělá inteligence. 1*. Academia, 1993, ISBN 80-200-0496-3.

[9] Polikar, R.: Ensemble based systems in decision making. *IEEE Circuits Syst Mag.*, 2006.

[10] Ponnunswamy, P.; Gromiha, M.: On the conformational stability of folded proteins. *Journal of theoretical biology*, ročník 166, č. 1, 1994: s. 63–74.

[11] Rybička, J.: *LaTeX pro začátečníky*. Konvoj, 1999, ISBN ISBN 80-85615-77-0.

[12] Rábová, Z.; Hanáček, P.; Peringer, P.; aj.: *Užitečné rady pro psaní odborného textu*. FIT VUT v Brně, Listopad 2008, [Online; navštíveno 12.05.2015].
URL http://www.fit.vutbr.cz/info/statnice/psani_textu.html

[13] Schapire, R. E.: The Strength of Weak Learnability. *Machine Learning*, 1990.

[14] Voet, J. G. P. C. W., D.; Voet: *Fundamentals of Biochemistry: Life at the Molecular Level, 3rd Edition*. John Wiley and Sons, Inc., 2008, ISBN 0470129301.

[15] Whitford, D.: *Proteins: Structure and function*. Wiley, 2005.

[16] Řehout, J. B. B. a., V.; Čítek: *Základy genetiky a poradenství*. Únor 2003, [Online; cit. 17.12.2017].
URL http://www.zsf.jcu.cz/cs/katedra/katedra-klinickych-a-preklinickych-oboru/import/ucebni_texty/zaklady-genetiky-a-poradenstvi

Příloha A

Jak pracovat s touto šablonou

V této kapitole je uveden popis jednotlivých částí šablony, po kterém následuje stručný návod, jak s touto šablonou pracovat.

Jedná se o přechodnou verzi šablony. Nová verze bude zveřejněna do konce roku 2017 a bude navíc obsahovat nové pokyny ke správnému využití šablony, závazné pokyny k vypracování bakalářských a diplomových prací (rekapitulace pokynů, které jsou dostupné na webu) a nezávazná doporučení od vybraných vedoucích, která již teď najdete na webu (viz odkazy v souboru s literaturou). Jediné soubory, které se v nové verzi změní, budou `projekt-01-kapitoly-chapters.tex` a `projekt-30-prilohy-appendices.tex`, jejichž obsah každý student vymaže a nahradí vlastním. Šablonu lze tedy bez problémů využít i v současné verzi.

Popis částí šablony

Po rozbalení šablony naleznete následující soubory a adresáře:

bib-styles Styly literatury (viz níže).

obrazky-figures Adresář pro Vaše obrázky. Nyní obsahuje placeholder.pdf (tzv. TODO obrázek, který lze použít jako pomůcku při tvorbě technické zprávy), který se s prací neodevzdává. Název adresáře je vhodné zkrátit, aby byl jen ve zvoleném jazyce.

template-fig Obrázky šablony (znak VUT).

fitthesis.cls Šablona (definice vzhledu).

Makefile Makefile pro překlad, počítání normostran, sbalení apod. (viz níže).

projekt-01-kapitoly-chapters.tex Soubor pro Váš text (obsah nahraďte).

projekt-20-literatura-bibliography.bib Seznam literatury (viz níže).

projekt-30-prilohy-appendices.tex Soubor pro přílohy (obsah nahraďte).

projekt.tex Hlavní soubor práce – definice formálních částí.

Výchozí styl literatury (czechiso) je od Ing. Martínka, přičemž anglická verze (englishiso) je jeho překladem s drobnými modifikacemi. Oproti normě jsou v něm určité odlišnosti, ale

na FIT je dlouhodobě akceptován. Alternativně můžete využít styl od Ing. Radima Loskota nebo od Ing. Radka Pyšného[1]. Alternativní styly obsahují určitá vylepšení, ale zatím nebyly řádně otestovány větším množstvím uživatelů. Lze je považovat za beta verze pro zájemce, kteří svoji práci chtějí mít dokonalou do detailů a neváhají si nastudovat detaily správného formátování citací, aby si mohli ověřit, že je vysázený výsledek v pořádku.

Makefile kromě překladu do PDF nabízí i další funkce:

- přejmenování souborů (viz níže),
- počítání normostran,
- spuštění vlny pro doplnění nezlomitelných mezer,
- sbalení výsledku pro odeslání vedoucímu ke kontrole (zkontrolujte, zda sbalí všechny Vámi přidané soubory, a případně doplňte).

Nezapomeňte, že vlna neřeší všechny nezlomitelné mezery. Vždy je třeba manuální kontrola, zda na konci řádku nezůstalo něco nevhodného – viz Internetová jazyková příručka[2].

Pozor na číslování stránek! Pokud má obsah 2 strany a na 2. jsou jen „Přílohy“ a „Seznam příloh“ (ale žádná příloha tam není), z nějakého důvodu se posune číslování stránek o 1 (obsah „nesedí“). Stejný efekt má, když je na 2. či 3. stránce obsahu jen „Literatura“ a je možné, že tohoto problému lze dosáhnout i jinak. Řešení je několik (od úpravy obsahu, přes nastavení počítadla až po sofistikovanější metody). **Před odevzdáním proto vždy překontrolujte číslování stran!**

Doporučený postup práce se šablonou

1. **Zkontrolujte, zda máte aktuální verzi šablony.** Máte-li šablonu z předchozího roku, na stránkách fakulty již může být novější verze šablony s aktualizovanými informacemi, opravenými chybami apod.

2. **Zvolte si jazyk**, ve kterém budete psát svoji technickou zprávu (česky, slovensky nebo anglicky) a svoji volbu konzultujte s vedoucím práce (nebyla-li dohodnuta předem). Pokud Vámi zvoleným jazykem technické zprávy není čeština, nastavte příslušný parametr šablony v souboru projekt.tex (např.: `documentclass[english]{fitthesis}` a přeložte prohlášení a poděkování do angličtiny či slovenštiny.

3. **Přejmenujte soubory.** Po rozbalení je v šabloně soubor `projekt.tex`. Pokud jej přeložíte, vznikne PDF s technickou zprávou pojmenované `projekt.pdf`. Když vedoucímu více studentů pošle `projekt.pdf` ke kontrole, musí je pracně přejmenovávat. Proto je vždy vhodné tento soubor přejmenovat tak, aby obsahoval Váš login a (případně zkrácené) téma práce. Vyhněte se však použití mezer, diakritiky a speciálních znaků. Vhodný název může být např.: „`xlogin00-Cisteni-a-extrakce-textu.tex`“. K přejmenování můžete využít i přiložený Makefile:

   ```
   make rename NAME=xlogin00-Cisteni-a-extrakce-textu
   ```

[1] BP Ing. Radka Pyšného http://www.fit.vutbr.cz/study/DP/BP.php?id=7848
[2] Internetová jazyková příručka http://prirucka.ujc.cas.cz/?id=880

4. Vyplňte požadované položky v souboru, který byl původně pojmenován `projekt.tex`, tedy typ, rok (odevzdání), název práce, svoje jméno, ústav (dle zadání), tituly a jméno vedoucího, abstrakt, klíčová slova a další formální náležitosti.

5. Nahraďte obsah souborů s kapitolami práce, literaturou a přílohami obsahem svojí technické zprávy. Jednotlivé přílohy či kapitoly práce může být výhodné uložit do samostatných souborů – rozhodnete-li se pro toto řešení, je doporučeno zachovat konvenci pro názvy souborů, přičemž za číslem bude následovat název kapitoly.

6. Nepotřebujete-li přílohy, zakomentujte příslušnou část v `projekt.tex` a příslušný soubor vyprázdněte či smažte. Nesnažte se prosím vymyslet nějakou neúčelnou přílohu jen proto, aby daný soubor bylo čím naplnit. Vhodnou přílohou může být obsah přiloženého paměťového média.

7. Nascanované zadání uložte do souboru `zadani.pdf` a povolte jeho vložení do práce parametrem šablony v projekt.tex (`documentclass[zadani]{fitthesis}`).

8. Nechcete-li odkazy tisknout barevně (tedy červený obsah – bez konzultace s vedoucím nedoporučuji), budete pro tisk vytvářet druhé PDF s tím, že nastavíte parametr šablony pro tisk: (`documentclass[zadani,print]{fitthesis}`). Barevné logo se nesmí tisknout černobíle!

9. Vzor desek, do kterých bude práce vyvázána, si vygenerujte v informačním systému fakulty u zadání. Pro disertační práci lze zapnout parametrem v šabloně (více naleznete v souboru fitthesis.cls).

10. Nezapomeňte, že zdrojové soubory i (obě verze) PDF musíte odevzdat na CD či jiném médiu přiloženém k technické zprávě.

Obsah práce se generuje standardním příkazem `\tableofcontents` (zahrnut v šabloně). Přílohy jsou v něm uvedeny úmyslně.

Pokyny pro oboustranný tisk

- **Oboustranný tisk je doporučeno konzultovat s vedoucím práce.**
- Je-li práce tištěna oboustranně a její tloušťka je menší než tloušťka desek, nevypadá to dobře.
- Zapíná se parametrem šablony: `\documentclass[twoside]{fitthesis}`
- Po vytištění oboustranného listu zkontrolujte, zda je při prosvícení sazební obrazec na obou stranách na stejné pozici. Méně kvalitní tiskárny s duplexní jednotkou mají často posun o 1–3 mm. Toto může být u některých tiskáren řešitelné tak, že vytisknete nejprve liché stránky, pak je dáte do stejného zásobníku a vytisknete sudé.
- Za titulním listem, obsahem, literaturou, úvodním listem příloh, seznamem příloh a případnými dalšími seznamy je třeba nechat volnou stránku, aby následující část začínala na liché stránce (\cleardoublepage).
- Konečný výsledek je nutné pečlivě překontrolovat.

Styl odstavců

Odstavce se zarovnávají do bloku a pro jejich formátování existuje více metod. U papírové literatury je častá metoda s použitím odstavcové zarážky, kdy se u jednotlivých odstavců textu odsazuje první řádek odstavce asi o jeden až dva čtverčíky (vždy o stejnou, předem zvolenou hodnotu), tedy přibližně o dvě šířky velkého písmene M základního textu. Poslední řádek předchozího odstavce a první řádek následujícího odstavce se v takovém případě neoddělují svislou mezerou. Proklad mezi těmito řádky je stejný jako proklad mezi řádky uvnitř odstavce. [12] Další metodou je odsazení odstavců, které je časté u elektronické sazby textů. První řádek odstavce se při této metodě neodsazuje a mezi odstavce se vkládá vertikální mezera o velikosti 1/2 řádku. Obě metody lze v kvalifikační práci použít, nicméně často je vhodnější druhá z uvedených metod. Metody není vhodné kombinovat.

Jeden z výše uvedených způsobů je v šabloně nastaven jako výchozí, druhý můžete zvolit parametrem šablony „`odsaz`“.

Užitečné nástroje

Následující seznam není výčtem všech využitelných nástrojů. Máte-li vyzkoušený osvědčený nástroj, neváhejte jej využít. Pokud však nevíte, který nástroj si zvolit, můžete zvážit některý z následujících:

MikTeX LaTeX pro Windows – distribuce s jednoduchou instalací a vynikající automatizací stahování balíčků.

TeXstudio Přenositelné opensource GUI pro LaTeX. Ctrl+klik umožňuje přepínat mezi zdrojovým textem a PDF. Má integrovanou kontrolu pravopisu, zvýraznění syntaxe apod. Pro jeho využití je nejprve potřeba nainstalovat MikTeX.

WinEdt Ve Windows je dobrá kombinace WinEdt + MiKTeX. WinEdt je GUI pro Windows, pro jehož využití je nejprve potřeba nainstalovat MikTeX či TeX Live.

Kile Editor pro desktopové prostředí KDE (Linux). Umožňuje živé zobrazení náhledu. Pro jeho využití je potřeba mít nainstalovaný TeX Live a Okular.

JabRef Pěkný a jednoduchý program v Javě pro správu souborů s bibliografií (literaturou). Není potřeba se nic učit – poskytuje jednoduché okno a formulář pro editaci položek.

InkScape Přenositelný opensource editor vektorové grafiky (SVG i PDF). Vynikající nástroj pro tvorbu obrázků do odborného textu. Jeho ovládnutí je obtížnější, ale výsledky stojí za to.

GIT Vynikající pro týmovou spolupráci na projektech, ale může výrazně pomoci i jednomu autorovi. Umožňuje jednoduché verzování, zálohování a přenášení mezi více počítači.

Overleaf Online nástroj pro LaTeX. Přímo zobrazuje náhled a umožňuje jednoduchou spolupráci (vedoucí může průběžně sledovat psaní práce), vyhledávání ve zdrojovém textu kliknutím do PDF, kontrolu pravopisu apod. Zdarma jej však lze využít pouze s určitými omezeními (někomu stačí na disertaci, jiný na ně může narazit i při psaní bakalářské práce) a pro dlouhé texty je pomalejší.

Pozn.: Overleaf nepoužívá Makefile v šabloně – aby překlad fungoval, je nutné kliknout pravým tlačítkem na `projekt.tex` a zvolit „Set as Main File“.

Užitečné balíčky pro LaTeX

Studenti při sazbě textu často řeší stejné problémy. Některé z nich lze vyřešit následujícími balíčky pro LaTeX:

- `amsmath` – rozšířené možnosti sazby rovnic,
- `float, afterpage, placeins` – úprava umístění obrázků,
- `fancyvrb, alltt` – úpravy vlastností prostředí Verbatim,
- `makecell` – rozšíření možností tabulek,
- `pdflscape, rotating` – natočení stránky o 90 stupňů (pro obrázek či tabulku),
- `hyphenat` – úpravy dělení slov,
- `picture, epic, eepic` – přímé kreslení obrázků.

Některé balíčky jsou využity přímo v šabloně (v dolní části souboru fitthesis.cls). Nahlédnutí do jejich dokumentace může být rovněž užitečné.

Sloupec tabulky zarovnaný vlevo s pevnou šířkou je v šabloně definovaný „L“ (používá se jako „p“).