

Téma 1: PREDIKCE VÝSKYTU HOREČKY DENGUE

Motivace

Pro různá místa na světě jsou typické i výskyty specifických nemocí. Horečka dengue je jednou z nemocí přenášenou převážně hmyzem v oblastech Karibiku, Filipín, Malajsie, Tchaj-wanu nebo Indie. Do ostatních regionů se tato nemoc dostává kvůli stále se zvětšujícímu počtu turistů. V dnešní době neexistuje účinná vakcína, léčba je pouze symptomatická. Ačkoliv většina nemocných se vyléčí bez jakýchkoliv následků, u některých skupin pacientů (děti, osoby s chronickými nemocemi) může dengue vést ke vzniku závažných zdravotních komplikací (vnitřní krvácení, šokový syndrom). Predikce výskytu dengue napomáhá např. včasnému zajištění bezpečnostních opatření při cestování do daného regionu nebo doporučení pro místní obyvatelstvo. Vzhledem ke způsobu přenosu nemoci (komáři), predikci jejího výskytu lze úspěšně provést za použití meteorologických, sociodemografických aj. dat.

Cíl projektu

Cílem je predikovat **incidenci horečky dengue** s co nejvyšší přesností na základě dostupných meteorologických dat.

Datové soubory

Data pochází z oblasti San-Juan (Portoriko). K dispozici jsou různé meteorologické parametry měřené v dané lokalitě a údaje o výskytu horečky dengue v místní populaci. Meteorologické parametry mohou sloužit jako příznaky pro trénování predikčního modelu. Údaje o výskytu dengue v místní populaci mohou sloužit jako očekávané výstupní hodnoty modelu. Jelikož se jedná o predikční úlohu, pro trénování modelu by se měla použít pouze data z období předcházejícího okamžiku predikce, tj. např. pro predikci dengue v 5. týdnu sledování lze využít pouze data z předchozího období (tj. z 1.-4. týdne). Délka sledovaného a analyzovaného časového okna přitom musí respektovat známé souvislosti mezi predikovanou hodnotou a použitými příznaky. Pokud se jedná o včasnou predikci dengue s předstihem 2 týdny dopředu, pak pro predikci např. v 5. týdnu sezóny lze pro trénování využít pouze data z období 1.-2. týden.

SanJuanData.csv – shrnuje meteorologické údaje a informaci o výskytu horečky dengue v dané lokalitě.

Unikátní meteorologická data pocházející z globálního celosvětového měření různých veličin. Výsledné parametry jsou získány v rámci komplexního zpracování a analýzy všech nasbíraných údajů. Data jsou sbírána s frekvencí 1x denně po dobu několika kalendářních let.

Údaj o výskytu horečky dengue je vyjádřen jako celkové množství pacientů s diagnostikovanou horečkou dengue. Tento údaj zahrnuje pacienty s pozitivním PCR nálezem pro jeden ze čtyř antigeně odlišných sérotypů dengue (DENV1, DENV2, DENV3, DENV4 – „přímá“ diagnostika přítomnosti samotného viru dengue) a pacienty s pozitivním sérologickým vyšetřením („nepřímá“ diagnostika dengue - detekce přítomnosti protilátek třídy IgM a IgG produkovaných organismem v případě kontaktu pacienta s virem). Sběr těchto dat se provádí v tzv. sezóně sledování dengue, která začíná týdnem s nejnižší incidencí nemoci v dané lokalitě (zjištění za celou dobu sledování dengue v dané lokalitě) a končí po uplynutí 12 měsíců. Začátek sezóny sledování nemoci nekoresponduje se začátkem kalendářního roku. Hodnoty jsou zaznamenávány 1x týdně.

Year: rok měření

Month: měsíc měření

Day: pořadí dne

airTemp: teplota vzduchu (K)

dewPoint: teplota rosného bodu (K)
humidityRelative: relativní vlhkost vzduchu (%)
humiditySpecific: měrná vlhkost vzduchu (g kg-1)
tempMax: maximální denní teplota vzduchu (K)
tempMin: minimální denní teplota vzduchu (K)
tempAvg: průměrná denní teplota vzduchu (K)
tempMinMaxDiff: denní rozpětí mezi minimální a maximální teplotou (K)
Precipitation: denní množství atmosférických srážek (kg m-2)
dengueCases: celkové množství pacientů s diagnostikovanou horečkou dengue (použijte pro odvození očekávaných hodnot incidence onemocnění)
populationTotal: odhadovaná velikost sledované populace v daném roce

Chybějící hodnoty měřených veličin jsou označeny: NaN

Výstup projektu

Výstupem projektu budou hodnoty **incidence dengue pro každý sledovaný týden s předstihem 1 měsíce**. Pro vyhodnocení úspěšnosti predikce použijte *průměrnou absolutní odchylku* (mean absolute error, MAE) mezi predikovanou hodnotou incidence y a očekávanou hodnotou incidence d (n – počet predikovaných hodnot):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - d_i|$$

Další informace o horečce dengue např. zde: http://www.khsstc.cz/dokumenty/horecka-dengue-4658_4658_161_1.html

Téma 2: DETEKCE SEPSE NA ZÁKLADĚ KLINICKÝCH DAT

Motivace

Sepse je život ohrožující stav, který je v českém prostředí často označován nepřesně jako “otrava krve”. Reálně je ale sepsa jednou z nejzávažnějších a nejčastějších komplikací v traumatologii, chirurgii a obecně intenzivní péči. Jedná se o celkovou reakci organismu na infekci, může přecházet v septický šok zahrnující i orgánové selhávání. Mortalita spojená se sepsí v posledních letech “porazila” i mortalitu na infarkt myokardu. Náklady na léčbu těžké sepsy se pohybují v řádech statisíců Kč. Vzhledem k oběma uvedeným okolnostem je nasnadě automaticky detekovat nástup septického stavu a tím snížit nároky na personál JIP a zároveň eliminovat lidský faktor. Septický stav je možné detekovat na základě sledování vitálních funkcí pacienta (krevní tlak, teplota, tep, ...) nebo laboratorních vyšetření. Pomoci mohou i demografické charakteristiky (věk, pohlaví,...).

Cíl projektu

Cílem je určit, zda se pacient nachází v septickém stavu či nikoli na základě dostupných klinických dat.

Datové soubory

dataSepsis.csv – klinická data pacientů JIP a údaje o stavu pacienta.

Vitální charakteristiky, výsledky laboratorního vyšetření a demografické charakteristiky mohou sloužit k trénování klasifikačního modelu za účelem hodnocení stavu pacienta (septický či neseptický). Jednotlivé hodnoty se sbírají v různých časech s rozdílnými intervaly mezi měřeními. Z tohoto důvodu může v záznamech část údajů chybět.

Vitální charakteristiky

HR	Tepová frekvence (bpm)
O2Sat	Saturace O2 (%)
Temp	Teplota těla (° C)
SBP	Systolický tlak (mmHg)
MAP	Střední arteriální tlak (mmHg)
DBP	Diastolický tlak (mmHg)
Resp	Dechová frekvence (počet dechů za minutu)
EtCO2	Obsah CO2 ve vzduchu na konci výdechu (end-tidal CO2) (mmHg)

Laboratorní vyšetření

BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bikarbonáty (mmol/L)
FiO2	koncentrace O2 ve vdechovaném vzduchu (%)
pH	N/A
PaCO2	Parciální tlak CO2 v arteriální krvi (mmHg)
SaO2	Saturace O2 v arteriální krvi (%)
AST	Aspartátaminotransferáza (IU/L)
BUN	Močovinový dusík v krvi (mg/dL)

Alkalinephos	Alkalická fosfatáza (IU/L)
Calcium	Vápník (mg/dL)
Chloride	Chloridy (mmol/L)
Creatinine	Kreatinin (mg/dL)
Bilirubin_direct	Bilirubin přímý (mg/dL)
Glucose	Glykemie (mg/dL)
Lactate	Laktát (mg/dL)
Magnesium	Hořčík (mmol/dL)
Phosphate	Fosfát (mg/dL)
Potassium	Draslík (mmol/L)
Bilirubin_total	Bilirubin celkový (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematokrit (%)
Hgb	Hemoglobin (g/dL)
PTT	Aktivovaný parciální tromboplastinový čas (s)
WBC	Leukocyty (*10 ³ /μL)
Fibrinogen	Fibrinogen (mg/dL)
Platelets	Trombocyty (count*10 ³ /μL)
Demografické charakteristiky	
Age	Věk
Gender	Pohlaví: žena (0) or muž (1)
Unit1	ID JIP (MICU)
Unit2	ID JIP (SICU)
HospAdmTime	Počet hodin mezi hospitalizací a umístěním na JIP
ICULOS	Doba strávená na JIP (hod)
Stav pacienta	
isSepsis	0 – neseptický, 1 – septický (odhadovaná veličina)

Chybějící hodnoty měřených veličin jsou označeny: NaN

Výstup projektu

Výstupem projektu budou hodnoty (0/1) indikující **stav pacienta**. Pro vyhodnocení úspěšnosti klasifikace dat použijte *matici záměn* (confusion matrix) a *Se*, *Sp*, *Acc* (v případě rovnoměrně zastoupených klasifikačních skupin) či *F-measure* (v případě nerovnoměrně zastoupených klasifikačních skupin):

$$Se = TP / (TP + FN)$$

$$Acc = (TP + TN) / (TP + TN + FP + FN)$$

$$Sp = TN / (TN + FP)$$

$$PPV = TP / (TP + FP)$$

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

Téma 3: ROZPOZNÁNÍ ČÍSLIC Z OBRAZOVÝCH DAT (OPTICAL CHARACTER RECOGNITION, OCR)

Motivace

Rozpoznání textu je jednou ze základních úloh automatického zpracování obrazů s výskytem textových dat. Typické využití lze nalézt např. při rozpoznávání poznávacích značek automobilů ze silničních kamer, strojovém čtení archivních lékařských zpráv, či u pomůcek pro nevidomé. Algoritmus by měl být schopen rozpoznat v obrazu znak za různě ztížených podmínek (snížená viditelnost, geometrické zkreslení, šum, obrazové artefakty) a bez ohledu na typ písma, kterým byl znak napsán.

Cíl projektu

Cílem projektu je vytvořit klasifikační model, který ve vstupním obrazu rozpozná číslice z intervalu 0–9 a přiřadí jim odpovídající hodnotu dle zadání.

Datové soubory











Data pro trénování a testování modelu jsou dostupná ve složce trainData. Finální testování bude provedeno na skryté sadě dat. Udržujte proto strukturu dat, která vám byla poskytnuta (v opačném případě nebude možné projekt automaticky vyhodnotit a bude hodnocen 0 body).

characterName_id.png – Obrazová data znaku ve formátu *.png (barevná hloubka 8 bitů, 3 kanály).

references.csv – Tabulka obsahující referenční hodnoty (požadované třídy) pro jednotlivé obrázky.

Název třídy odpovídá danému znaku (např. znak „3“ je zařazen do třídy 3). Výjimku tvoří znak „0“, který je klasifikován do třídy označené 10.

Ukázka dat:

Vstupní data:										
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
Výstup:	10	3	7	5	2	8	6	9	4	1

Výstup projektu

Pro vyhodnocení úspěšnosti klasifikace dat použijte matici záměn (confusion matrix) a F1 skóre.

$$F = 2 \frac{PPV * Se}{PPV + Se}$$

TESTOVÁNÍ ODEVZDANÝCH ALGORITMŮ A OBHAJOBA PROJEKTŮ

Nezávislé testování odevzdaných algoritmů

Odevzdané algoritmy **budou testovány na nezávislém testovacím souboru dat** (testovací soubory mají formát shodný s trénovacími soubory). Finální nezávislé testování bude zajištěno vyučujícími. Z tohoto důvodu vás žádáme o odevzdání výstupů ve standardizovaném formátu, viz poskytnuté kódy ke každému zadání. Výstupy odevzdá vybraný člen týmu (kontaktní osoba) za celý tým. O výsledcích testování se dozvíte v rámci obhajoby projektů.

Obhajoba projektu

Řešiteli projektu budou týmy složené z maximálně 3 studentů/studentek.

Součástí úspěšného splnění zadání bude obhajoba projektu. Prezentace se zúčastní všichni členové týmu. V rámci obhajoby se každý tým vyjádří ke třem částem projektu:

1. *Příprava a předzpracování dat*

Tato oblast může zahrnovat: parsování dat, škálování či standardizaci příznaků, ověření rovnoměrnosti zastoupení jednotlivých klasifikačních skupin a případnou augmentaci méně zastoupené skupiny dat, detekce a odstranění odlehlých hodnot, nahrazování chybějících hodnot, interpolaci, převzorkování, překódování, způsob rozdělení na trénovací a testovací množinu, selekce vhodných příznaků,...)

2. *Použitá predikční/klasifikační metoda*

Např.: Výběr vhodného modelu, nastavení modelu, ověření úspěšnosti predikce/klasifikace,...

3. *Interpretace a diskuze výsledků*

Např.: Rozebrání použitých příznaků a parametrů modelu a jejich vlivu na výsledek, hodnocení generalizační schopnosti modelu, interpretace výsledků vzhledem k povaze zadaného problému, porovnání s výsledky z literatury, limitace a výhody použitého postupu...)

Přiřazení studenta k části prezentace bude provedeno náhodně, proto každý student musí být připraven prezentovat jakoukoli část projektu. Hodnocená bude odborná úroveň projektu i schopnost reagovat na případné dotazy publika a komise.

Délka obhajoby projektu je max 12 min, včetně prezentace výsledků a diskuze. Tomu odpovídá cca 5-8 slajdů prezentace (*.ppt) a 5 min diskuze. Účast týmů na prezentacích výsledků ostatních skupin není povinná.

Důležité termíny

Nejzazší termín odevzdání: 11. prosince

Termín obhajoby: 17. prosince, začátek 8 h (zápočtový týden). Pořadí prezentací bude dohodnuto s týmy.