Python - projekt
Lukáš Marek

1 Teorie

1.1 Používané algoritmy

1.1.1 Ridge regrese

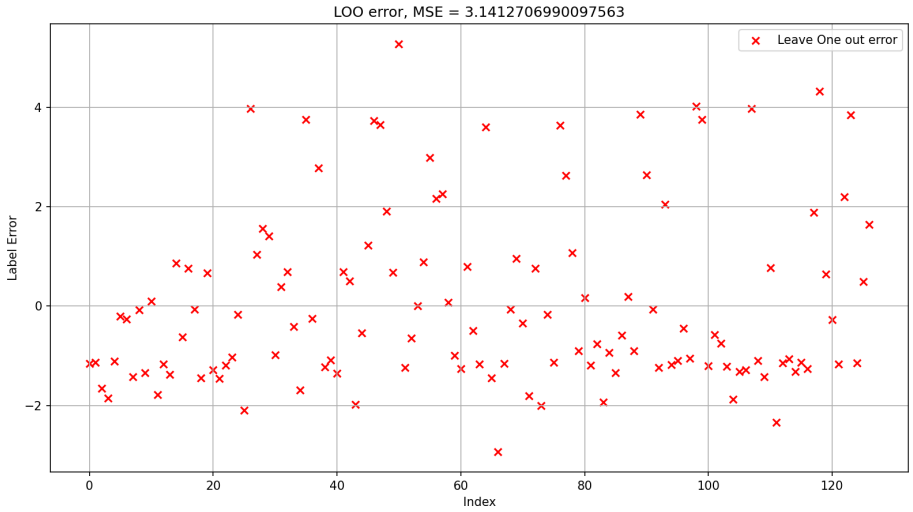
Mějme datovou matici $\mathbb{X}$ typu $m \times n$ (máme tedy n různých datových bodů v m rozměrném prostoru) a vektor $\mathbf{y}$ hodnot vysvětlované proměnné. Cílem našeho snažení je pak nalezení takové lineární kombinace (v podobě vektoru parametrů $\boldsymbol{\theta}$) vysvětlujících proměnných, která co nejlépe vysvětluje získané hodnoty proměnné y. V naprosté většině případů platí $n \geq m$ a pokud je alespoň $m + 1$ řádků v rozšířené matici $(\mathbb{X}|\mathbf{y})$ lineárně nezávislých, tak neexistuje přesné řešení soustavy $\mathbb{X}\boldsymbol{\theta} = \mathbf{y}$. Lineární regrese se pak snaží nalézt aproximativní řešení, které minimalizuje jistou *cost*-funkci $J_{\mathbb{X}}(\boldsymbol{\theta})$. Nejběžněji používanou metodou je metoda nejmenších čtverců, která minimalizuje funkci $J_{\mathbb{X}}(\boldsymbol{\theta}) = (\mathbb{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbb{X}\boldsymbol{\theta} - \mathbf{y})$. Snadnou derivací lze odvodit, že vektor $\boldsymbol{\theta}$ minimalizující tuto funkci lze získat analyticky ze vztahu $\boldsymbol{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y}$. Ve velkém počtu případů je tento vztah plně dostačující. Pokud ale máme silně korelovaná data, může být matice $\mathbb{X}^T\mathbb{X}$ singulární (a tedy neinvertovatelná) a nebo špatně podmíněná a celý výpočet je pak numericky nestabilní (chyby v měření mají velký vliv na celkové řešení). Ridge regrese se snaží řešit tento problém zavedením regularizačního parametru α do vztahu pro $\boldsymbol{\theta} = (\mathbb{X}^T\mathbb{X} + \alpha\mathbb{I})^{-1}\mathbb{X}^T\mathbf{y}$, kde $\mathbb{I}$ značí jednotkovou matici příslušných rozměrů (matice $\mathbb{X}^T\mathbb{X}$ je totiž symetrická a tedy čtvercová). Z tohoto důvodu je někdy tato korekce označována jako $L2$-regularizace.

1.1.2 DBSCAN

DBSCAN je jedna z metod shlukové analýzy. Pokud si každý prvek v datech představím jako bod v nějakém prostoru, tak cílem shlukové analýzy je rozdělení těchto bodů do skupin (shluků či klastrů) v závislosti na jejich mezi bodové vzdálenosti (body ve stejné skupině by měli být co nejblíže u sebe a naopak body z různých skupin by měli mít co největší vzdálenost mezi sebou).
DBSCAN se snaží rozdělit body do shluků na základě jisté lokální hustoty. Použitím nastavitelných parametrů ϵ a *minPts*. DBSCAN postupně přiřazuje každému bodu jisté označení. Nejdříve se zvolí náhodně bod p. Následně se určí všechny body N ve vzdálenosti ϵ od p (včetně p samotného). Pokud je těchto bodů méně než *minPts*, tak je bod p označen jako šum a náhodně vybereme nový bod. Naopak, pokud je těchto bodů alespoň *minPts*, tak se bod p označí jako centrový bod a přiřadí se mu číslo skupiny. Následně postupně procházíme všechny body q v množině N (kromě p). Pokud nemá bod q přiřazenou skupinu (jedná se tedy o nenavštíveny bod nebo o šum), tak se přidá bod q do stejné skupiny jako p. Potom se vyberou všechny body M ve vzdálenosti ϵ od q a znova se určí jejich počet. Je-li jich méně než *minPts*, označíme q jako okrajový bod a vybere nový bod z N. Pokud je jich alespoň *minPts*, tak se q označí za centrový a všechny body z M se přidají do množiny N a vybereme další bod z N. Takto celý proces opakujeme dokud nemá každý datový bod přiřazené označení.
Výhodou DBSCAN algoritmu oproti například populárnímu k-means, je že není třeba dopředu zvolit počet shluků. Navíc jsou v DBSCAN hranice mezi shluky určeny libovolnou křivkou, zatímco v k-means to jsou vždy přímky (rovniny). Naopak je u DBSCAN nelehké správné určení parametrů ϵ a *minPts* a popřípadě i metriky určující vzdálenost mezi body. Populárním výchozím bodem pro tyto parametry je zvolení $minPts = 2 * d$, kde d je dimenze shlukovaných bodů. Následně se pro každý bod nalezne vzdálenost $minPts - 1$-ního nejbližšího souseda využitím grafu k-nejbližších sousedů a všechny tyto vzdálenosti se sestupně uspořádají. Parametr ϵ se pak zvolí jako zlomový bod ("loket") v grafu takto seřazených vzdálenosti.

1.2 Data

V celé této práci jsou zpracovávána neveřejná obličejová data pacientů trpících různými úrovněmi částečné (jednostranné) paralýzy obličeje. Míra této paralýzy je určena doktorem na základě série běžných obličejových úkonů (úsměv, mračení, špulení rtů atd.) a řídí se House-Brackmannovou škálou. Ta pacientům přiřadí hodnocení od 1 do 6 (1 znamená zdravou funkci obličejových svalů a 6 je úplná paralýza). V průběhu těchto cvičení byly pacienti natáčeni a při dalším zpracování byly získány námi používaná data.

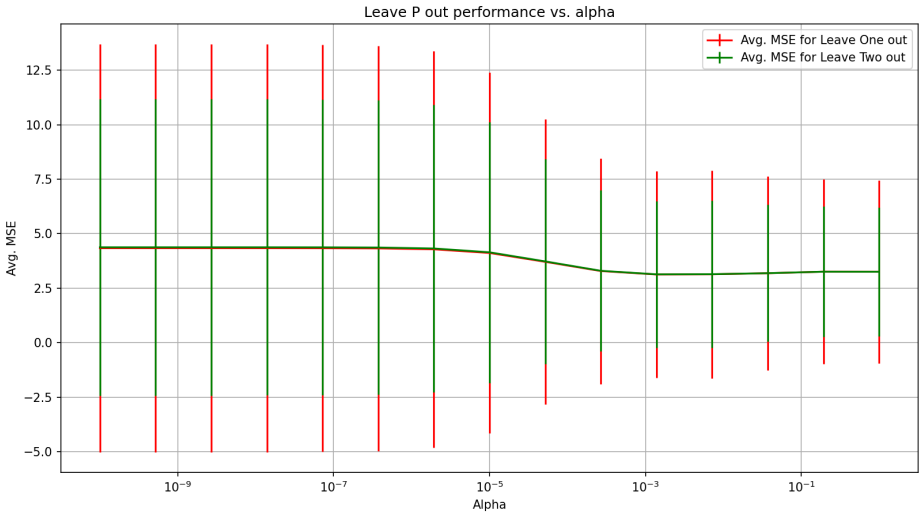


Obrázek 1: Rozdíly mezi skutečným a předpovězeným HB hodnocením pro jednotlivá vynechaná měření.

2 Experimentální část

2.1 Ridge regrese

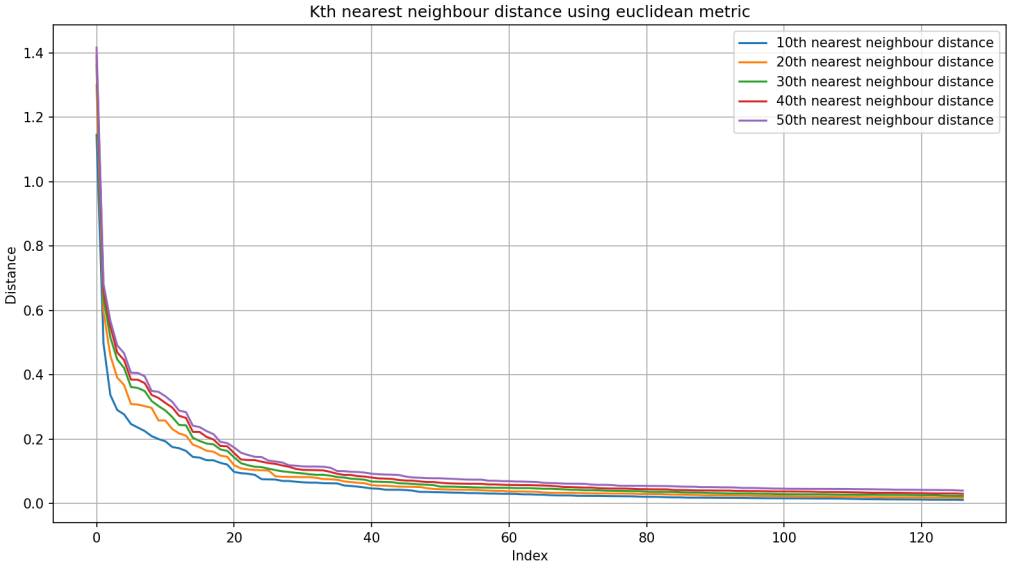
Jako první jsem použili Ridge regresi na variacích v datech jednotlivých pacientů. Přesněji jsme pro každé měření spojili jednotlivé cviky dohromady a vytvořili tak průběh pohybů (pohyb byl rozdělený podle os) jednotlivých obličejových bodů během měření. Pro každý takový bod jsme spočetly celkovou vzdálenost od kamery a nakonec jsme spočetly variaci v této vzdálenost. Z každého měření jsme tak získali 21 hodnot (variace v pohybech jednotlivých bodů). Dohromady jsme měli 127 měření a tato data společně s HB hodnocením jsme použili jako vstup pro Ridge regresi. Data jsme nejdříve rozdělili na trénovací a testovací (90 : 10). Prvotní výsledky naznačovali výraznou variabilitu v kvalitě modelu v závislosti na konkrétním rozdělení dat do trénovacích a testovacích skupin. Pro další vyhodnocení kvality modelu jsme proto použili metodu *Leave P Out*, která se používá při cross-validaci statistických modelů. Metoda je založená na postupném vynechání p hodnot z datasetu, Následně je model natrénován na zbylých datech a otestován na vynechaných p hodnotách. Cely proces se opakuje pro všechny možné kombinace vynechaných hodnot. Z tohoto důvodu se používá metoda jen pro malé hodnoty p a malé datasety. Na obrázku výše (**Obr.** 1) jsou vyobrazeny chyby pro jednotlivé vynechané měření z datasetu. Z **Obr.** 2 je pak patrné, že ani úprava parametru α v Ridge regresi neumožňuje dosažení kvalitních výsledků. Celkově lze tedy spolehlivě říci, že Ridge regrese není pravděpodobně dostatečně komplexní model, pro zpracování takto složitých dat.



Obrázek 2: MSE Leave P out pro $p = 1, 2$ v závislosti na parametru α Ridge regrese

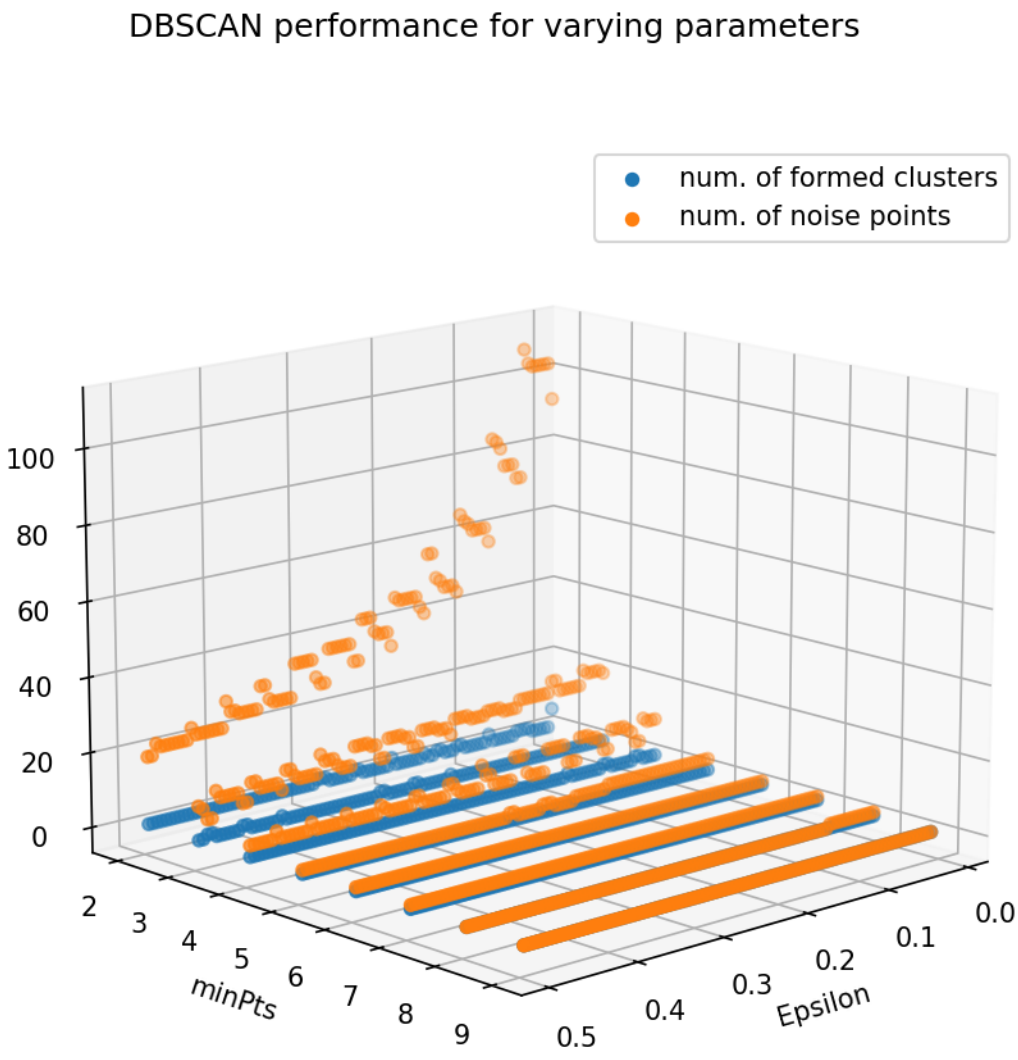
2.2 DBSCAN

Po Ridge regresi jsme se přesunuli k shlukové analýze s běžnou Euklidovskou metrikou. Jako první z parametrů DBSCANu jsme určili ϵ. Pro různé hodnoty k jsme nalezli pro každý bod vzdálenost ke k-tému nejbližšímu sousedu a tyto vzdálenosti jsme uspořádala do grafu 3 (přesnější postup je popsaný výše). Tímto jsme dostali přibližný rozsah $[0.01, 0.4]$ pro hodnoty parametru ϵ.



Obrázek 3: Seřazené vzdálenosti ke k-tým sousedům pro různé hodnoty k

Z důvodu nízkého počtu datových bodů, je nevhodné odvozovat hodnotu parametru *minPts* z dimenze dat (v našem případě je $d = 21$). Proto jsme testovali hodnoty parametru *minPts* v rozshahu od 2 do 10. Různé výsledky shlukování pro parametry z těchto rozsahů jsou vyobrazeny v grafu 4. Okamžitě je patrné, že pro vyšší hodnoty parametrů vzniká pouze jediný shluk obsahující naprostou většinu bodů. Pro klesající ϵ pak jen přibývá šumu, ale počty shluků jsou stále okolo 1 či 2. Vidíme tedy, že všechny datové body jsou velmi blízko



Obrázek 4: Počty vzniklých shluků a počty bodů klasifikovaných jako šum v závislosti na vstupních parametrech DBSCANu.

u sebe a veškeré rozdíly v datech, které by umožnili jejich rozdělení do adekvátních shluků jsou pro DBSCAN neviditelné.

3 Závěr

Negativní výsledky používaných algoritmů jsou očekávané a pouze odrazují komplexnost používaných dat. Navíc je patrné, že naprostá většina této komplexnosti je ztracena, pokud pracujeme pouze s variacemi v datech a ne s celými časovými řadami. Tento fakt nás motivuje k použití pokročilejší statistických postupů a metod strojového učení.