Strojové učenie - Projekt

Rastislav Simeunovič

Úvod

V mojom projekte sa budem venovať regresii. Budem robiť lineárnu aj polynomiálnu regresiu. Tak isto som pri týchto regresiách skúšal vynechať aj niektoré featury, aby som videl ako to ovplyvní výsledok.
Regresie boli vykonávané na datasete Abalone, ktorý obsahuje informácie o ústriciach[1] a z týchto informácií sa budem snažiť predikovať vek ústrice.

Dataset

Ako je už v úvode spomenuté, tak v projekte používam dataset Abalone[9].
Tento dataset obsahuje informácie o ústriciach a pochádza pobrežného výskumného laboratória v Tasmánií avšak nikde som nenašiel informácie, či tieto ústrice pochádzaju tak isto z Tasmánie alebo aj okolia napríklad z Austrálie a Nového Zélandu prípadne z celého sveta.
Dataset bol zhotovený v roku 1995 a nachádzajú sa v ňom informácie o 4177 ústriciach. O každej ústrici je tam 9 údajov a to pohlavie, dĺžka, šírka, priemer, váha, váha lastúry, váha vnútornosti, váha mäsa a vek.
Všetky údaje až na vek sa získavajú presnými meraniami[2]. Vek ústric sa zisťuje podobne ako u stromov. Keď sa ústrica prereže, tak na nej vidno kruhy, ktore sú veľmi podobné ako letokruhy stromov. Tieto kruhy sa spočítali pod mikroskopom a pripočitaním hodnoty 1.5 sa dostane vek ústrice. Vek ústric v tomto datasete sa pohybuje v rozmedzí od 1 po 29 rokov.

Predošlé práce

Na tomto datasete bolo už vykonávaných viacero projektov, ktoré majú súvis so strojovým učením.[4]
Patria medzi ne *Age of Abalones using Physical Characteristics* , *Extending and benchmarking Cascade-Correlation*, *A Quantitative Comparison of Dystal and Backpropagation*.

Age of Abalones using Physical Characteristics

Prvý z nich sa venoval klasifikácií, keď si vek rozdelili na 9 class po 4 ročných skokoch - 1 až 4, 5 až 9 atď.
Dáta si potom rozdelili na trénovaciu a testovaciu množinu. Trénovaciu množinu tvorilo 75% dát a brali dáta v takom poradí ako boli v súbore.
Väčšina dát z datasetu sú z classy 2 a 3 tj. 5 až 12 ročné ústrice. Následne použili 2 clustrovacie algoritmy - k-means a Hierarchical clustering a tieto potom porovnali.
K-means algoritmus následne spustili cca 50-krát, kedže veľmi záleži ako sa nainicializujú vstupné stredy clustrov a hodnoty potom spriemerovali.
S algoritmom k-means dosiahli presnosť okolo 62% s druhým dosiahli presnosť iba 6.23% - prečo tak málo vysvetľujú v článku.
Článok vznikol na University of Wisconsin v roku 2010. Tento článok som spomenul najmä preto, lebo sa budem v tomto texte na neho ešte párkrát odkazovať aj napriek tomu, že nerobím klasifikačnú úlohu, tak niektoré veci som využil aj u seba.

Použité technológie

Kód bol napísaný v jazyku Python[7], pričom kľúčove bolo najmä využitie knižníc. Všetko to boli knižnice, ktoré sme využívali na cvičeniach.
Na prácu s maticami som použil knižnicu NumPy[5]. Na spracovanie vstupných dát som použil Pandas[6].
Ďalej som používal sklearn[8] na regresiu - aj na lineárnu aj na polynomiálnu.
Na plotovanie 2D a 3D obrázkov som použil matplotlib[3].

Spracovanie dát

Dáta boli v peknom tvare a nebolo ich treba veľmi spracovávať dopredu. Trebalo na nich iba spraviť comma separate a upraviť featuru sex.
Táto featura na vstupe obsahovala hodnoty M (male), F (female) alebo I (infant). Takže miesto tohto som potreboval urobiť one hot encoding. Pridaním dummies variables nám zmizli síce featury, ktoré mali ako hodnotu char, ale dostali sme tým pádom 2 nové stĺpce s featurami.
Po týchto úpravách som dáta náhodne poprehadzoval, aby každý test bol na "iných" trénovacich dátach, a vybral som prvých 85% dát, ktoré som používal ako trénovaciu množinu. Zvyšných 15% boli testovacie dáta

Regresia vs Klasifikácia

Pôvodne som uvažoval urobiť na tomto datasete klasifikačný problém postavený na rovnakom princípe ako to bolo v článku *Age of Abalones*[4] a porovnávať výsledky s ich výsledkami prípadne urobiť podobne ako oni k-means a upraviť

ho, nakoniec som však od tejto myšlienky upustil a urobil regresiu.
Na tomto datasete by sa dali robiť rôzne regresné a klasifikačné problémy. Napríklad určiť či sa jedná o ústricu samčiu, samičiu alebo nevyvinutú. Alebo napríklad skúsiť odhadnúť váhu na základe ostatných parametrov ústrice. Nakoniec som sa rozhodol, že budem určovať vek ústrice.

Priebeh programu

Ako som spomínal po úprave dát som ich poprehadzoval, čo spôsobilo, že pravdepodobne žiadne 2 merania neboli robené na identickej trénovacej množine. Toto je na jednej strane dobré, lebo by sa mohlo stať, že by v testovacej množine bolo veľa hodnôt, ktoré sú blízko extrémov a bola by u nich veľká chyba, na druhej strane však môžeme argumentovať, že regresie vyšli nepresne resp. horšie ako iné prípady, lebo tie mali lepšiu trénovaciu množinu.
Aby som zabránil takýmto dohadom, tak som každú regresiu - lineárnu, polynomiálnu pre polynóm druhého stupňa a polynóm tretieho stupňa spustil aspoň 50-krát. Tieto testy vykonáva regresie z balíku scikit-learn[8]. Výsledky som následne spriemeroval pre jednotlivé stupňe polynómu.
Tak isto som sa snažil dáta nejako rozumne vyplotovať avšak pri toľkých featurach nám ani 3D plot nedáva moc výpovedné hodnoty, kedže sa to ťažko predstavuje len na záklde 2 prinajlepšom 3 featurách.
.

Výsledky

Program som spúšťal pre rôzne trénovacie množiny, tak isto som skúšal aj meniť - vynechávať - featury v týchto množinách a meniť aj veľkosť trénovacej množiny. Pre žiadnu testovaciu množinu som nedostal nejaké veľmi odlišné výsledky. Čo však veľmi zavážilo bolo, koľkého stupňa bol polynóm, ktorý predikoval vek ústric.

Stupeň polynómu	Priemerná chyba	Priemerná percentuálna chyba
1	1.591	16.1202
2	1.535	15.2961
3	1.796	18.3765
4	22.4753	274.9557

Všetky boli spustené 75-krát na testovacej množine s veľkosťou 3550 dát.

Ako vidieť tak pri polynóme tretieho stupňa sa chyba už zväčšuje a pri polynóme štvrtého stupňa je už chyba obrovská. Tu už dochádza k over-fittingu a výsledky už nie su moc relevantné pre stupne väčšie ako 3. Pre všetky testy mi to vyšlo veľmi podobne ako v tabuľke.

Aj pre menej featur alebo menší dataset bola polynomiálna regresia pre polynóm druhého stupňa najpresnejšia. Lineárna regresia bola vždy presnejšia ako kubická funkcia avšak iba o trochu. Vidíme že percentuálne rozdiely medzi lineárnou, kvadratickou a kubickou sa pohybujú v rozmedzí +-3%, čo je stále veľmi slušné podľa mňa.

Následne som skúšal program pre featury, keď som zahrnul iba 2 z 3 kategorií - pohlavie, rozmery a váha. Chyba pre kvadratický polynóm v žiadnom z prípadov nepresiahla 18% a veľmi málo sa líšila chyba aj pre zvyšné 2 stupne - vždy v rozmedzí +-3%.

Keď som si zobral iba jednu z kategorií, tak najpresnejšie určovala regresia vek na základe váhovej kategórie, najnepresnejšie na základe veku, to je však očakávané, keďže vek má iba 3 featury a vždy sú 2 nuly a zvyšné je jednotka.

Analýza chýb

Na počiatku som očakával, že priemerná chyba bude daľeko menšia - maximálne na úrovni okolo 10% a v priemere tak 5-8%. Predpokladal som to, na základe toho, že trénovacia množina je dostatočne veľká a máme pomerne dosť featur. Po zvážení som však usúdil, že táto chyba je pomerne malá a vek ústric to určovalo celkom presne. Poďme sa teda pozrieť, kde mohli nastať chyby.

Živé organizmy

Jeden z dôvodov je podľa mňa ten, že ústrice sú podobne ako človek živý organizmu a jeho rast a vývoj ovplyvňuje množstvo faktorov napr. slnečné žiarenia, teplota vody atď., ktoré neboli v dátach zahrnuté. A aj keby boli zahrnuté, tak by sa to nedalo predikovať presne. Osobne si myslím, že keby máme rovnaké dáta ako boli pre ustrice aj pre človeka, tak chyba nám výjde omnoho väčšia, pretože každý živý organizmus sa vyvýja úplne inak.

Veľmi malý rozptyl dát

Dát bolo síce veľmi málo ale rozptyl bol pomerne malý. Z vyše 4000 ustric bolo cca 3400 vo veku od 5 do 12 rokov. To znamená, že cca 83% dát spadá do vekového rozmedzia, ktoré pokrýva iba štvrtinu z možného veku ústrice.
Tak isto, keď som chcel dáta predikovať iba na základe jednej featury - pohlavie, tak bola chyba okolo 25%, čo je veľmi málo. Z tohto dôvodu som usúdil, že rozptyl dát bol malý, lebo keď na základe jednej featury to určilo tak presne, tak nech by bola regresná krivka akákoľvek, tak by celkom presne fitovala dáta. Celkom presne v tomto prípade myslím okolo tých 25%, lebo na základe jednej hodnoty je to veľmi presné určenie.

Záver

Na počiatku som podľa dát očakával, že model bude presnejší, avšak keď som si lepšie pozrel dáta a chcel som zabrániť overfitovaniu tak by som povedal, že chyba okolo 16% nie je vôbec veľká a mohlo to byť ďaleko lepšie. Možno by bolo ešte zaujímave porovnať to s nejakou neurónovou sieťou a pozrieť si ake výsledky by hádzala ona. To už však nechám pre čitateľa :-)

References

[1] Abalone. http://en.wikipedia.org/wiki/abalone.

[2] Abalone Data Set README file. http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.names.

[3] matplotlib. https://matplotlib.org/.

[4] Hiran Mayukh. Age of abalones using physical characteristics: A classification problem. 2010.

[5] NumPy. http://www.numpy.org/.

[6] pandas. https://pandas.pydata.org/.

[7] Python. https://www.python.org/.

[8] scikit learn. http://scikit-learn.org/stable/.

[9] UCI Machine Learning Repository: Abalone Data Set. http://archive.ics.uci.edu/ml/datasets/abalone.

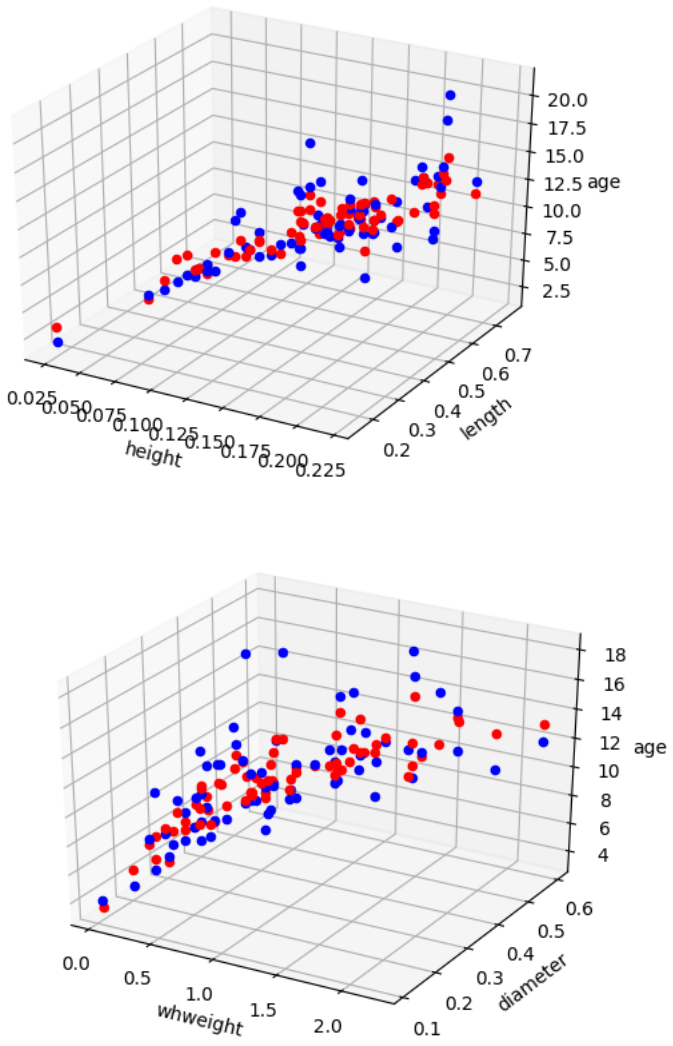

Figure 1: Skutočný vek je modrou a predikovaný červenou

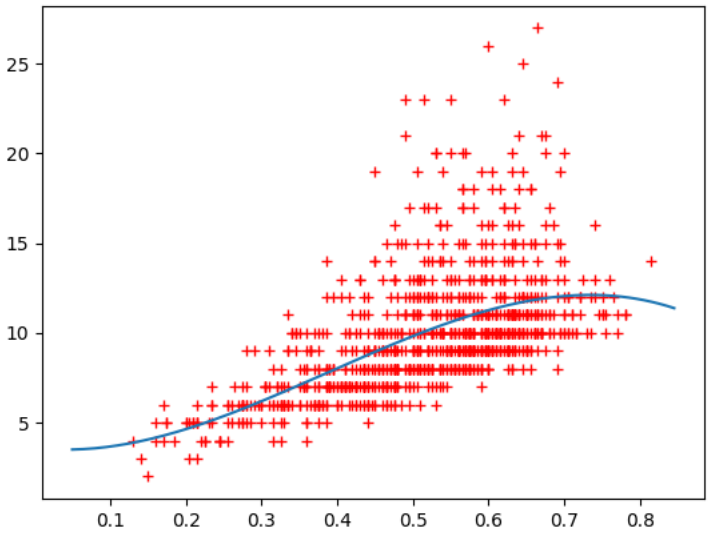

Figure 2: Fitovanie dát pomocou kubickej funkcie

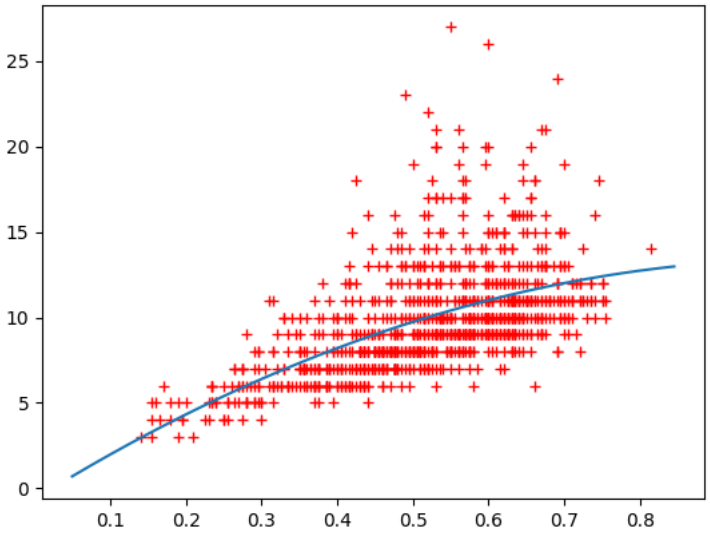

Figure 3: Fitovanie dát pomocou kvadratickej funkcie