Data Science Hw 3

工科海洋四 B07505015 梁瑞翔

Problem 1

(1)

I use the LassoCV estimator, which is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.. The features with the highest absolute coef value are considered the most important. I set the value which is positive to one, and negative number to zero in label.txt. The result shows there are 19 significant features

(2)

Hsa.1131	T92451	3' UTR	1	118219	"TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN);.
Hsa.140 M87789	gene	1		IG GAMMA-1 CHAIN C REGION (HUMAN);.	
Hsa.4689	T95018	3' UTR	2a	120032	40S RIBOSOMAL PROTEIN S18 (Homo sapiens)
Hsa.3004	H55933	3' UTR	1	203417	H.sapiens mRNA for homologue to yeast ribosomal protein L41.
Hsa.3002	R22197	3' UTR	1	130829	60S RIBOSOMAL PROTEIN L32 (HUMAN);.
Hsa.13491	R39465	3' UTR	2a	23933	EUKARYOTIC INITIATION FACTOR 4A (Oryctolagus cuniculus)
Hsa.8147	M63391	gene	1		"Human desmin gene, complete cds.
Hsa.2357	T52342	3' UTR	1	72028	Human tral mRNA for human homologue of murine tumor rejection antigen gp96.
Hsa.316 M94132	gene	1		Human mucin 2 (MUC2) mRNA sequence.	
Hsa.31 T57780	3' UTR	1	80626	IG LAMBDA CHAIN C REGIONS (HUMAN).	
Hsa.832 T51023	3' UTR	1	75127	HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN).	
Hsa.2688	X60489	gene	1		Human mRNA for elongation factor-1-beta.
Hsa.45604	H88360	3' UTR	2a	252849	"GUANINE NUCLEOTIDE-BINDING PROTEIN G(OLF), ALPHA SUBUNIT (Rattus norvegicus)
Hsa.539 U14971	gene	1		"Human ribosomal protein S9 mRNA, complete cds.	
Hsa.1534	J00231	gene	1		Human Ig gamma3 heavy chain disease OMM protein mRNA.
Hsa.1737	T72175	3' UTR	1	85528	IG KAPPA CHAIN PRECURSOR V-III REGION (HUMAN);.
Hsa.5821	X57351	gene	1		INTERFERON-INDUCIBLE PROTEIN 1-8D (HUMAN);contains MSR1 repetitive element ;.
Hsa.8068	T57619	3' UTR	2a	75437	40S RIBOSOMAL PROTEIN S6 (Nicotiana tabacum)

19 features are selected, and code is at page 5

Reference: Model-based and sequential feature selection — scikit-learn 1.0.1 documentation

Problem 2

(1)

(a)

I choose PSO for feature selection

1. Initialize parameters like position and velocity, number of particles.
2. Initialize the particle set.
3. while(until reaching the max iteration):
    Update the position
    Update the fitness

(b) Objective function: fitModel(), which returns global best position and cost value

(c) tunable parameters are particle numbers, max iteration, c1 c2 r1 r2 to calculate the optimal position and velocity.

initPopulation() to initialize the particle set

costFunction() to calculate the fitness in the objective function in each interation

(2)

187 features are selected, and code is at page 7

```
Hsa.1908        L07032  gene    1               "Human protein kinase C theta (PKC) mRNA, complete cds.
Hsa.36696       D59253  gene    1               Human mRNA for NCBP interacting protein 1.
Hsa.3160        D26069  gene    1               "Human mRNA (KIAA0041) for ORF, partial cds.
Hsa.8597        H23135  3' UTR  2a      52302   SKD1 PROTEIN (Mus musculus)
Hsa.1491        M35531  gene    1               "Human GDP-L-fucose:beta-D-galactoside 2-alpha-1-fucosyltransferase mRNA, complete cds.
Hsa.2933        T69748  3' UTR  1       108186  PHOSPHOGLYCERATE KINASE 1 (HUMAN);.
Hsa.26673       R76825  3' UTR  2a      143867  RAN-SPECIFIC GTPASE-ACTIVATING PROTEIN (Homo sapiens)
Hsa.2471        H20512  3' UTR  1       173142  H.sapiens mRNA for Sm protein G.
Hsa.8093        R74349  3' UTR  2a      157011  TUBULIN ALPHA-2 CHAIN (Drosophila melanogaster)
Hsa.41299       U37673  gene    1               "Human neuron-specific vesicle coat protein and cerebellar degeneration antigen (beta-NAP) mRNA, complete cds.
Hsa.1875        L06175  gene    1               "Homo Sapiens P5-1 mRNA, complete cds.
Hsa.3098        X16316  gene    1               VAV ONCOGENE (HUMAN);.
Hsa.59  D00761  gene    1               PROTEASOME COMPONENT C5 (HUMAN);.
Hsa.2881        X17651  gene    1               Human Myf-4 mRNA for myogenic determination factor.
Hsa.9816        T50321  3' UTR  2a      72106   APOLIPOPROTEIN A-I PRECURSOR (Homo sapiens)
Hsa.767 J04162  gene    1               "Human leukocyte IgG receptor (Fc-gamma-R) mRNA, complete cds.
Hsa.2486        D14695  gene    1               Human mRNA (KIAA0025) for ORF (complete cds) and PIGHEP3 homologous region.
Hsa.8606        T70063  3' UTR  2a      80947   EUKARYOTIC INITIATION FACTOR 4 GAMMA (Oryctolagus cuniculus)
Hsa.40206       R99916  3' UTR  2a      201708  GENOME POLYPROTEIN (Human rhinovirus 89)
Hsa.2409        U24077  gene    1               "Human p58 natural killer cell receptor precursor mRNA, clone cl-39, complete cds.
Hsa.2774        Z17240  gene    1               Homo sapiens for mRNA encoding HMG2B.
Hsa.17  D00265  gene    1               "Human cytochrome c mRNA, carboxyl-terminal region and 3' non-coding region.
Hsa.22762       H17434  3' UTR  1       50609   NUCLEOLIN (HUMAN);.
Hsa.551 L02426  gene    1               "Human 26S protease (S4) regulatory subunit mRNA, complete cds.
Hsa.3225        U27699  gene    1               "Human pephBGT-1 betaine-GABA transporter mRNA, complete cds.
Hsa.22  X57974  gene    1               H.sapiens mRNA for keratinocyte transglutaminase.
Hsa.205 L12350  gene    1               THROMBOSPONDIN 2 PRECURSOR (HUMAN);.
Hsa.983 L13977  gene    1               "Human prolylcarboxypeptidase mRNA, complete cds.
Hsa.895 X72018  gene    1               H.sapiens hTGR 1 mRNA.
Hsa.2780        X65438  gene    1               H.sapiens U21.1 mRNA.
Hsa.3301        X89430  gene    1               H.sapiens mRNA for methyl CpG binding protein 2.
Hsa.27285       R70790  3' UTR  2a      142585  GTP:AMP PHOSPHOTRANSFERASE MITOCHONDRIAL (Rattus norvegicus)
Hsa.2926        X59842  gene    1               Human PBX2 mRNA.
```

```
Hsa.2962        T92259  3' UTR  1       118111  PROTEASOME IOTA CHAIN (HUMAN);.
Hsa.3117        R49231  3' UTR  1       38397   MITOCHONDRIAL PHOSPHATE CARRIER PROTEIN PRECURSOR (HUMAN);.
Hsa.975 D78152  gene    1               Human mRNA for annexin IV (carbohydrate-binding protein p33/41).
Hsa.2298        U03865  gene    1               "Human adrenergic alpha-1b receptor protein mRNA, complete cds.
Hsa.9711        T50500  3' UTR  2a      77136   PLACENTAL THROMBIN INHIBITOR (Homo sapiens)
Hsa.2419        D45887  gene    1               "Human mRNA for calmodulin, complete cds.
Hsa.557 H65019  3' UTR  1       238799  60 KD RO PROTEIN (HUMAN);.
Hsa.42826       H66834  3' UTR  2a      210698  NON-ERYTHROID PROTEIN 4.1 (Homo sapiens)
Hsa.879 H41129  3' UTR  1       175539  GALECTIN-1 (HUMAN);contains Alu repetitive element;.
Hsa.8837        T40568  3' UTR  2a      60605   MSS4 PROTEIN (Saccharomyces cerevisiae)
Hsa.17514       X89985  gene    1               H.sapiens mRNA for BCL7B protein.
Hsa.1045        H05899  3' UTR  1       43338   HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEINS C1/C2 (HUMAN);.
Hsa.9631        T49397  3' UTR  2a      67478   SHC TRANSFORMING PROTEINS 46.8 KD AND 51.7 KD PRECURSOR (Homo sapiens)
Hsa.1712        X17273  gene    1               Human HLA G (HLA 6.0) mRNA for non classical class I transplantation antigen.
Hsa.3225        U27699  gene    1               "Human pephBGT-1 betaine-GABA transporter mRNA, complete cds.
i
Hsa.638 L06111  gene    1               "Human L-type voltage-gated calcium channel B subunit mRNA for isoform b, complete cds.
Hsa.14914       T71260  3' UTR  2a      110197  SIGNAL RECOGNITION PARTICLE 14 KD PROTEIN (Canis familiaris)
Hsa.2191        T40645  3' UTR  1       60737   "Human Wiskott-Aldrich syndrome (WAS) mRNA, complete cds.
Hsa.541 U14973  gene    1               "Human ribosomal protein S29 mRNA, complete cds.
Hsa.2499        D26068  gene    1               "Human mRNA (KIAA0038) for ORF, partial cds.
Hsa.23608       H24250  3' UTR  2a      52050   SEVENLESS PROTEIN (Drosophila melanogaster)
Hsa.41347       U37408  gene    1               "Human CtBP mRNA, complete cds.
Hsa.41260       L11706  gene    1               "Human hormone-sensitive lipase (LIPE) gene, complete cds.
Hsa.120 D14662  gene    1               "Human mRNA for ORF, complete cds.
Hsa.32734       R77780  3' UTR  2a      145300  TRANSPOSABLE ELEMENT ACTIVATOR (Zea mays)
Hsa.22968       R20666  3' UTR  2a      26418   PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1 (Homo sapiens)
Hsa.919 H66976  3' UTR  1       212229  "HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DP(1) ALPHA CHAIN (HUMAN);.
Hsa.6458        H89087  3' UTR  2a      253224  SPLICING FACTOR SC35 (Homo sapiens)
Hsa.45754       H89688  3' UTR  2a      250489  GENERAL NEGATIVE REGULATOR OF TRANSCRIPTION SUBUNIT 4 (Saccharomyces cerevisiae)
Hsa.1578        D42046  gene    1               "Human mRNA (KIAA0083) for ORF (related to yeast gene in chromosome VIII), partial cds.
Hsa.359 U09413  gene    1               "Human zinc finger protein ZNF135 mRNA, complete cds.
Hsa.22614       R37276  3' UTR  2a      25988   EUKARYOTIC INITIATION FACTOR 4 GAMMA (Homo sapiens)
Hsa.2505        D13634  gene    1               "Human mRNA for ORF, complete cds.
```

Hsa.1273	T48041	3' UTR	1	72117	Human messenger RNA fragment for the beta-2 microglobulin.
Hsa.20 D12686	gene	1			Human mRNA for eukaryotic initiation factor 4 gamma (eIF-4 gamma).
Hsa.14763	H77536	3' UTR	2a	233349	SUCCINATE DEHYDROGENASE (Bos taurus)
Hsa.2291	H06524	3' UTR	1	44386	"GELSOLIN PRECURSOR, PLASMA (HUMAN);.
Hsa.10169	T52624	3' UTR	1	67178	H.sapiens mRNA for processing a-glucosidase I.
Hsa.2354	X52228	gene	1		Human mRNA for secreted epithelial tumour mucin antigen.
Hsa.12879	T95612	3' UTR	2a	120205	TUBULIN--TYROSINE LIGASE (Sus scrofa)
Hsa.34428	H09665	3' UTR	2a	46546	LAMIN B RECEPTOR (Gallus gallus)
Hsa.1651	M81933	gene	1		M-PHASE INDUCER PHOSPHATASE 1 (HUMAN);.
Hsa.12260	R81170	3' UTR	2a	147439	TRANSLATIONALLY CONTROLLED TUMOR PROTEIN (Homo sapiens)
Hsa.3121	H40705	3' UTR	1	191092	"MYOSIN REGULATORY LIGHT CHAIN 2, NONSARCOMERIC (HUMAN);.
Hsa.19232	R53936	3' UTR	2a	39921	PROTEIN PHOSPHATASE 2C HOMOLOG 2 (Schizosaccharomyces pombe)
Hsa.24490	R49719	3' UTR	2a	38755	GAMMA-AMINOBUTYRIC-ACID RECEPTOR BETA-4 SUBUNIT PRECURSOR (Gallus gallus)
Hsa.2391	L21993	gene	1		"Human adenylyl cyclase mRNA, 3' end of cds.
Hsa.21160	R99591	3' UTR	2a	201340	ANTIGEN WC1.1 (Bos taurus)
Hsa.2285	X12548	gene	1		Human mRNA for lysosomal acid phosphatase (EC 3.1.3.2).
Hsa.9251	T73092	3' UTR	2a	85983	EUKARYOTIC INITIATION FACTOR 4A-I (Homo sapiens)
Hsa.2654	Z14244	gene	1		H.sapiens coxVIIb mRNA for cytochrome c oxidase subunit VIIb.
Hsa.1244	J04794	gene	1		"Human aldehyde reductase mRNA, complete cds.
Hsa.42317	H62531	3' UTR	2a	208178	SPORE GERMINATION PROTEIN B2 (Bacillus subtilis)
Hsa.43540	H71122	3' UTR	2a	214653	PROBABLE G PROTEIN-COUPLED RECEPTOR R334 FROM PITUITARY GLAND (Rattus norvegicus)
Hsa.307 L22214	gene	1			"Human adenosine A1 receptor (ADORA1) mRNA exons 1-6, complete cds.
Hsa.41108	V00520	gene	1		Human germ line gene for growth hormone (presomatotropin).
Hsa.229 U04953	gene	1			"Human isoleucyl-tRNA synthetase mRNA, complete cds.
Hsa.2918	X67325	gene	1		H.sapiens p27 mRNA.
Hsa.601 J05032	gene	1			"Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
Hsa.2943	Z15114	gene	1		H.sapiens mRNA for protein kinase C gamma (partial).
Hsa.11352	T59406	3' UTR	2a	75852	LYMPHOID ENHANCER BINDING FACTOR 1 (Mus musculus)
Hsa.2089	M14016	gene	1		UROPORPHYRINOGEN DECARBOXYLASE (HUMAN);contains element PTR5 repetitive element ;.
Hsa.5346	T63370	3' UTR	2a	81523	GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT-LIKE PROTEIN 12.3 (Homo sapiens)
Hsa.2840	X53683	gene	1		Human LAG-1 mRNA.
Hsa.44676	H81802	3' UTR	2a	219929	VAV ONCOGENE (Homo sapiens)
Hsa.27560	R72164	3' UTR	2a	155799	HYPOTHETICAL 76.3 KD PROTEIN K04H4.2 IN CHROMOSOME III (Caenorhabditis elegans)
Hsa.1192	D38549	gene	1		"Human mRNA (KIAA0068) for ORF, partial cds.

Hsa.2092	L06328	gene	1		"Human voltage-dependent anion channel isoform 2 (VDAC) mRNA, complete cds.
Hsa.33144	H92195	3' UTR	2a	221798	EBV-INDUCED G PROTEIN-COUPLED RECEPTOR 1 PRECURSOR (Homo sapiens)
Hsa.8007	R32804	3' UTR	1	135146	"GLUCOSE TRANSPORTER TYPE 3, BRAIN (HUMAN);contains Alu repetitive element;.
Hsa.8576	R39540	3' UTR	2a	23916	DELTA ANTIGEN (Hepatitis delta virus)
Hsa.20883	R05291	3' UTR	2a	125114	SEROTRANSFERRIN PRECURSOR (Homo sapiens)
Hsa.471 M29277	gene	1			"Human isolate JuSo MUC18 glycoprotein mRNA (3' variant), complete cds.
Hsa.2311	T97199	3' UTR	1	120285	INTEGRIN BETA-4 SUBUNIT PRECURSOR (HUMAN);.
Hsa.1277	H69869	3' UTR	1	212381	CLATHRIN LIGHT CHAIN A (HUMAN);.
Hsa.2489	D21260	gene	1		"Human mRNA (KIAA0034) for ORF (rat cathrin heavy chain homologue), complete cds.
Hsa.9691	T49703	3' UTR	2a	67944	60S ACIDIC RIBOSOMAL PROTEIN P1 (Polyorchis penicillatus)
Hsa.1095	U01038	gene	1		"Human pLK mRNA, complete cds.
Hsa.216 U03100	gene	1			"Human alpha2(E)-catenin mRNA, complete cds.
Hsa.5141	D63876	gene	1		Human mRNA for ORF.
Hsa.952 U20141	gene	1			"Human inducible nitric oxide synthase mRNA, complete cds.
Hsa.24884	R22816	3' UTR	2a	130330	INSULIN-LIKE GROWTH FACTOR IA PRECURSOR (Homo sapiens)
Hsa.1044	M65028	gene	1		"Human hnRNP type A/B protein mRNA, complete cds.
Hsa.18897	T95318	3' UTR	2a	120517	"TRANSCRIPTION INITIATION FACTOR IIF, ALPHA SUBUNIT (Homo sapiens)
Hsa.192 U03851	gene	1			"Human capping protein alpha mRNA, partial cds.
Hsa.35968	H27277	3' UTR	2a	158402	HOMEOBOX PROTEIN HOX-A4 (Homo sapiens)
Hsa.26628	R34098	3' UTR	2a	135983	RNA-DIRECTED RNA POLYMERASE (Murine coronavirus mhv)
Hsa.5464	H17646	3' UTR	1	50603	ELONGATION FACTOR 1-DELTA (HUMAN);.
Hsa.3271	X89416	gene	1		H.sapiens mRNA for protein phosphatase 5.
Hsa.1213	M55268	gene	1		"Human casein kinase II alpha' subunit mRNA, complete cds.
Hsa.2827	Z11502	gene	1		H.sapiens mRNA for intestine-specific annexin.
Hsa.32907	R77220	3' UTR	2a	144765	TUBULIN ALPHA-2 CHAIN (Mus musculus)
Hsa.43684	H72110	3' UTR	2a	213627	T-CELL RECEPTOR BETA CHAIN PRECURSOR (Oryctolagus cuniculus)
Hsa.41299	U37673	gene	1		"Human neuron-specific vesicle coat protein and cerebellar degeneration antigen (beta-NAP) mRNA, complete cds.
Hsa.3001	T79152	3' UTR	1	113545	60S RIBOSOMAL PROTEIN L19 (HUMAN);.
Hsa.10308	R39144	3' UTR	2a	26598	HEAT SHOCK FACTOR PROTEIN 2 (Homo sapiens)
Hsa.9574	X80507	gene	1		H.sapiens YAP65 mRNA.
Hsa.2115	U15782	gene	1		"Human cleavage stimulation factor 77kDa subunit mRNA, complete cds.
Hsa.539 U14971	gene	1			"Human ribosomal protein S9 mRNA, complete cds.
Hsa.2389	X80692	gene	1		H.sapiens ERK3 mRNA.
Hsa.8085	H45251	3' UTR	2a	182771	HOMEOTIC GENE REGULATOR (Drosophila melanogaster)

Hsa.3141	D29641	gene	1		"Human mRNA (KIAA0052) for ORF, partial cds.
Hsa.878 T61609	3' UTR	1	78081	LAMININ RECEPTOR (HUMAN);.	
Hsa.2870	T78489	3' UTR	1	113437	"MYOSIN REGULATORY LIGHT CHAIN 2, NONSARCOMERIC (HUMAN);.
Hsa.6546	H20543	3' UTR	2a	51631	DIHYDROXYACETONE KINASE (Citrobacter freundii)
Hsa.9218	T51858	3' UTR	2a	75035	EUKARYOTIC INITIATION FACTOR 4B (Homo sapiens)
Hsa.100 H48027	3' UTR	1	193650	ATP SYNTHASE LIPID-BINDING PROTEIN P1 PRECURSOR (HUMAN);.	
Hsa.539 U14971	gene	1		"Human ribosomal protein S9 mRNA, complete cds.	
Hsa.688 H09137	3' UTR	1	46399	UBIQUINOL-CYTOCHROME C REDUCTASE CORE PROTEIN 2 PRECURSOR (HUMAN);.	
Hsa.1804	M93010	gene	1		"Human epithelial cell marker protein 1 (HMe1) mRNA, complete cds.
Hsa.2255	L40904	gene	1		"H. sapiens peroxisome proliferator activated receptor gamma, complete cds.
Hsa.1445	J03075	gene	1		"PROTEIN KINASE C SUBSTRATE, 80 KD PROTEIN, HEAVY CHAIN (HUMAN);contains TAR1 rep
Hsa.1464	M35878	gene	1		"Human insulin-like growth factor-binding protein-3 gene, complete cds, clone HL1
Hsa.41207	X66503	gene	1		Human adenylosuccinate synthetase mRNA.
Hsa.2478	R54818	3' UTR	1	40360	"Human eukaryotic initiation factor 2B-epsilon mRNA, partial cds.
Hsa.758 M98045	gene	1		"Homo sapiens folylpolyglutamate synthetase mRNA, complete cds.	
Hsa.2156	T67905	3' UTR	1	81962	"ATP SYNTHASE GAMMA CHAIN, MITOCHONDRIAL PRECURSOR (HUMAN);.
Hsa.1385	T53868	3' UTR	1	77996	"Human peroxisomal enoyl-CoA hydratase-like protein (HPXEL) mRNA, complete cds.
Hsa.2778	X66975	gene	1		H.sapiens mRNA for heterogeneous nuclear ribonucleoprotein.
Hsa.341 M26683	gene	1		Human interferon gamma treatment inducible mRNA.	
Hsa.18664	T94579	3' UTR	1	119384	"Human chitotriosidase precursor mRNA, complete cds.
Hsa.1212	T55117	3' UTR	1	73917	ALPHA-1-ANTITRYPSIN PRECURSOR (HUMAN).
Hsa.2612	R53769	3' UTR	1	138069	TRANSCRIPTION FACTOR BTF3 (HUMAN);.
Hsa.1698	T54360	3' UTR	1	69068	GRANULINS PRECURSOR (HUMAN).
Hsa.43252	H70425	3' UTR	2a	213691	INTERFERON-ALPHA RECEPTOR PRECURSOR (Homo sapiens)
Hsa.2426	L34840	gene	1		"Human transglutaminase mRNA, complete cds.
Hsa.612 M34175	gene	1		"Human beta adaptin mRNA, complete cds.	
Hsa.2875	X69295	gene	1		H.sapiens MSX2 mRNA for transcription factor.
Hsa.1514	M64673	gene	1		"Human heat shock factor 1 (TCF5) mRNA, complete cds.
Hsa.7499	R56207	3' UTR	2a	40639	SODIUM CHANNEL PROTEIN PARA (Drosophila melanogaster)
Hsa.1504	M21339	gene	1		"Human non-histone chromosomal protein HMG-14 gene, complete cds.
Hsa.3040	X51416	gene	1		Human mRNA for steroid hormone receptor hERR1.
Hsa.2407	X79888	gene	1		H.sapiens AUH mRNA.
Hsa.25830	R36644	3' UTR	2a	136908	ACTIVIN RECEPTOR TYPE IIB PRECURSOR (Xenopus laevis)
Hsa.19003	Z46389	gene	1		Homo sapiens encoding vasodilator-stimulated phosphoprotein (VASP).
Hsa.1006	T63591	3' UTR	1	79994	60S ACIDIC RIBOSOMAL PROTEIN P0 (HUMAN);.
Hsa.451 D21261	gene	1		SM22-ALPHA HOMOLOG (HUMAN);.	
Hsa.22251	R53036	3' UTR	2a	154339	PUTATIVE GTP-BINDING PROTEIN MOV10 (Mus musculus)
Hsa.1288	T53889	3' UTR	1	78017	COMPLEMENT C1R COMPONENT PRECURSOR (HUMAN).
Hsa.485 L37792	gene	1		"Human syntaxin 1A mRNA, complete cds.	
Hsa.3067	X05276	gene	1		Human mRNA for fibroblast tropomyosin TM30 (p1).
Hsa.12241	T64012	3' UTR	2a	79817	"ACETYLCHOLINE RECEPTOR PROTEIN, DELTA CHAIN PRECURSOR (Xenopus laevis)
Hsa.19731	H25136	3' UTR	2a	161038	"INOSITOL 1,4,5-TRISPHOSPHATE-BINDING PROTEIN TYPE 2 RECEPTOR (Rattus norvegicus)
Hsa.3069	M37984	gene	1		"Human slow twitch skeletal muscle/cardiac muscle troponin C gene, complete cds.
Hsa.18589	T93885	3' UTR	2a	116939	GA BINDING PROTEIN ALPHA CHAIN (Homo sapiens)
Hsa.2219	T56674	3' UTR	1	69387	CD63 ANTIGEN (HUMAN).
Hsa.10068	T52003	3' UTR	2a	72611	CCAAT/ENHANCER BINDING PROTEIN ALPHA (Rattus norvegicus)
Hsa.24539	H06970	3' UTR	2a	44769	SERINE/THREONINE-PROTEIN KINASE PAK (Rattus norvegicus)
Hsa.1132	D16431	gene	1		"Human mRNA for hepatoma-derived growth factor, complete cds.
Hsa.3239	T78104	3' UTR	1	114499	"Human proline- arginine-rich end leucine-rich repeat protein PRELP mRNA, complete cds.
Hsa.3254	X74795	gene	1		H.sapiens P1-Cdc46 mRNA.
Hsa.32404	R71092	3' UTR	2a	142784	EBNA-2 NUCLEAR PROTEIN (Epstein-barr virus)

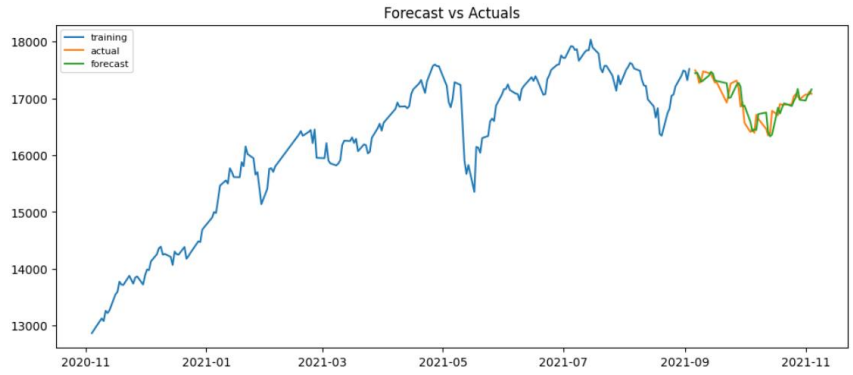
Reference: c++多元线性回归_五种优化算法实现多元线性回归_大雄行为锻炼的博客-CSDN 博客

Problem 3

(p, d, q, P, D, Q, s) = (3, 0, 1, 0, 1, 1, 12)

MSE = 23064.06888955432

(1)



```
# -*- coding: utf-8 -*-
"""hw3_p1.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1drhY2URxt9cfj4Z82qXdo6tTI5bdibqS
"""

import pandas as pd
import numpy as np
from statsmodels.tsa.seasonal import seasonal_decompose

from google.colab import drive
drive.mount('/content/drive/')

f_index = open('drive/MyDrive/hw3_prob1&2/index.txt')
gene = np.loadtxt('drive/MyDrive/hw3_prob1&2/gene.txt', dtype=float)
label = np.loadtxt('drive/MyDrive/hw3_prob1&2/label.txt', dtype=float)
feature_names = f_index.read().split('\n')

f_index.close()
label[label > 0] = 0
label[label < 0] = 1
gene = gene.transpose()

X, y = gene, label
#print(label)

import numpy as np
from sklearn.linear_model import LassoCV

lasso = LassoCV().fit(X, y)
importance = np.abs(lasso.coef_)

#print(importance)
#importance.sort()
#m = 20
```

```
#print(importance[-m:])

s = sorted(range(len(importance)), key=lambda k: importance[k])
m = 19
#print(s[-m:])

for i in range(1981,1999):
    print(feature_names[s[i]])
```

```
# -*- coding: utf-8 -*-
"""hw3_p2.ipynb

Automatically generated by Colaboratory.

Original file is located at

https://colab.research.google.com/drive/1qmxQ1Ov6dcS7LBNok4b95kgkDYRAduGo
"""

import pandas as pd
import numpy as np
from statsmodels.tsa.seasonal import seasonal_decompose

from google.colab import drive
drive.mount('/content/drive/')

f_index = open('drive/MyDrive/hw3_prob1&2/index.txt')
gene = np.loadtxt('drive/MyDrive/hw3_prob1&2/gene.txt', dtype=float)
label = np.loadtxt('drive/MyDrive/hw3_prob1&2/label.txt', dtype=float)
feature_names = f_index.read().split('\n')

f_index.close()
label[label > 0] = 0
label[label < 0] = 1
gene = gene.transpose()

X, y = gene, label
#print(label)

from sklearn import datasets

data=datasets.load_diabetes()

class LiearRegressionPSO:
    def __init__(self):
        self.x , self.y = gene, label
        self.w = 0.8
```

```
    self.c1 = 2
    self.c2 = 2
    self.r1 = 0.5
    self.r2 = 0.5
    self.pN = 30
    self.dim = 2001
    self.max_iter = 2000
    self.X = np.zeros((self.pN, self.dim))
    self.V = np.zeros((self.pN, self.dim))
    self.pbest = np.zeros((self.pN, self.dim))
    self.gbest = np.zeros((1, self.dim))
    self.p_fit = np.zeros(self.pN)
    self.fit = 100000000
def costFunction(self,x):
    h=np.sum(x.T*np.insert(self.x,0,1,axis=1),axis=1)-self.y
    diff=h**2
    return diff.sum()/self.x.shape[0]
def initPopulation(self):
    for i in range(self.pN):
        for j in range(self.dim):
            self.X[i][j] = np.random.uniform(0, 1)
            self.V[i][j] = np.random.uniform(0, 1)
            self.pbest[i] = self.X[i]
            cost = self.costFunction(self.X[i])
            self.p_fit[i] = cost
            if (cost < self.fit):
                self.fit = cost
                self.gbest = self.X[i]
def fitModel(self):
    self.initPopulation()
    costVale=[]
    for i in range(self.max_iter):
        for j in range(self.pN):
            cost= self.costFunction(self.X[j])
            if (cost<self.p_fit[j]):
                self.pbest[j]=self.X[j]
                self.p_fit[j]=cost
            if (self.p_fit[j]<self.fit):
```

```
                self.gbest= self.X[j]
                self.fit= self.p_fit[j]
        for k in range(self.pN):
            self.V[k]= self.w*self.V[k]+self.c1*self.r1*(self.pbest[k]-
self.X[k])+self.c2*self.r2*(self.gbest-self.X[k])
            self.X[k]= self.X[k]+self.V[k]
            costVale.append(self.fit)
            return self.gbest ,costVale
    def predict(self,x):
        w ,cost= self.fitModel()
        return w[0]+np.sum(w[1:].reshape(-1,1).T*x ,axis=1)

temp = LiearRegressionPSO()
print(temp.predict(gene[0]))
print(temp.gbest)
import matplotlib.pyplot as plt

gbest = temp.gbest[temp.gbest > 0.9]
m = len(gbest)

s = sorted(range(len(temp.gbest)), key=lambda k: temp.gbest[k])
#print(s[-m:])

for i in range(1812,1999):
    print(feature_names[s[i]])
```