**تمرین 2 -بخش 2**

**سوال 1**

رگرسیون Ridge: زمانی بسیار مناسب است که تعداد زیادی متغیر پیشبین وجود دارد که ضرایب آنها غیرصفراست و از توزیع نرمال استخراج شده باشند. بهویژه، این روش زمانی بسیار مناسب است که تعداد متغیرهای موجود در مدل زیاد و یا همخطی چندگانه شدید وجود داشته باشد. در این هنگام واریانس برآوردگرها متورم و به شکل قله خود را نشان میدهد. در رگرسیون ستیغی از تابع زیان درجه ۲ استفاده می‌شود. به این ترتیب مقدار جریمه (Penalty) برای مدل رگرسیونی، به صورت مجموع مربعات ضرایب مشخص می‌شود.

رگرسیون Lasso: هدف از این رگرسیون به عنوان جایگزین خطی این است که دقت پیشبینی را بهبود داده و امکان تفسیر مدل را با ایجاد زیرگرد آیههای کوچکتری از متغیرهای کمکی با بیشترین اثر را فراهم آورد.

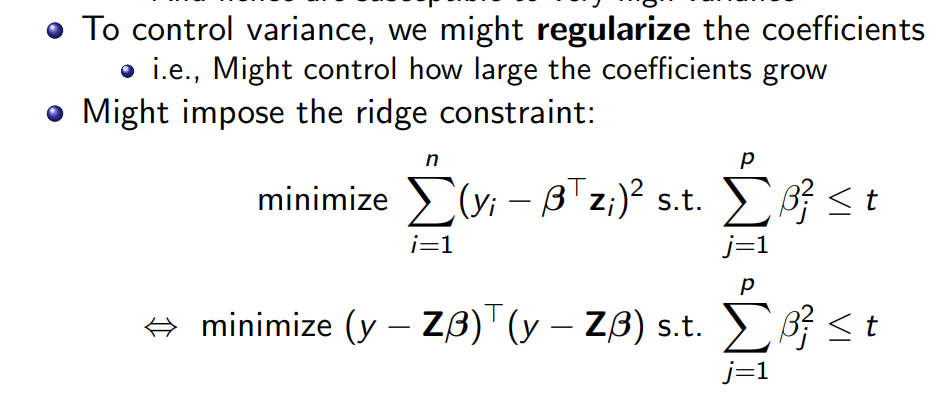
مقایسه:  رگرسیون لسو و ستیغی، به‌ترتیب بیشترین درصد واریانس و کمترین میانگین خطا را در مقایسه با رگرسیون خطی دارند.

منبع: <https://japr.ut.ac.ir/article_78578.html>

**سوال 2**

برای کنترل پراکندگی، ضریب مناسب regularization توسط لاندا بهینه میشود:

منبع: <https://statweb.stanford.edu/~owen/courses/305a/Rudyregularization.pdf>



**سوال 3**

K بزرگتر به معنای Bias کمتر نسبت به Overestimating خطای مورد انتظار است (زیرا فولدهای ترین به کل مجموعه داده نزدیکتر خواهد بود) اما واریانس بالاتر و زمان اجرای بالاتر را به همراه خواهد داشت.

منبع: <https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/>

**سوال4**

اعتبار سنجی متقابل Leave-one-out یک مورد خاص اعتبار سنجی متقابل(Cross Validation) است که در آن تعداد برابر برابر با تعداد نمونه های مجموعه داده است. بنابراین ، الگوریتم یادگیری برای هر نمونه یک بار استفاده می شود. بنابراین زمانی برای تخمین عملکرد الگوریتم های یادگیری ماشین استفاده می شود که از آنها برای پیش بینی داده هایی که برای ترین مدل استفاده نمی شوند.

منبع: <https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_469#fromHistory>

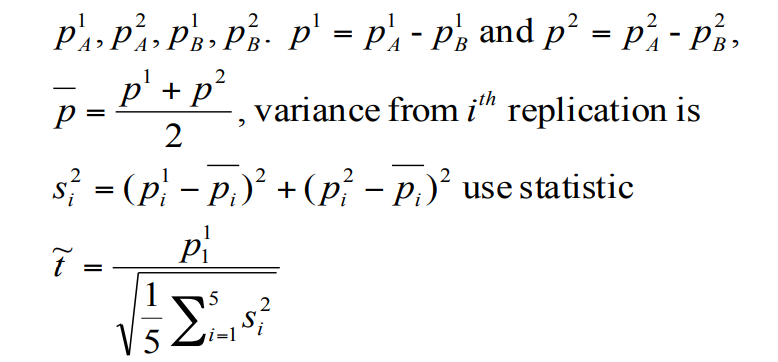
**سوال5 (امتیازی)**

Bootstrapping تکنیکی است که در بسیاری از شرایط مانند اعتبارسنجی عملکرد مدل ، ensemble methods ، تخمین Bias و واریانس مدل رفتار میکند این روش با جایگزینی نمونه‌گیری از داده‌های اصلی کار می‌کند و نقاط داده "انتخابی" را به عنوان موارد تست انتخاب می‌کند. ما می‌توانیم این کار را چندین بار انجام دهیم و میانگین نمره را به عنوان تخمینی از عملکرد مدل خود محاسبه کنیم. از طرف دیگر، اعتبار سنجی متقابل تکنیکی برای اعتبار سنجی عملکرد مدل است و با تقسیم داده‌های آموزشی به بخش‌های k انجام می‌شود. ما بخش‌های k - ۱ را به عنوان مجموعه آموزشی خود انتخاب کرده و از قسمت "نگه‌داشته شده" به عنوان مجموعه تست خود استفاده می‌کنیم. در نهایت ما میانگین k را به عنوان برآورد عملکرد خود در نظر می‌گیریم.

منبع: <https://nirpyresearch.com/kfold-montecarlo-cross-validation-bootstrap-primer/>

**سوال6**

در repeated cross-validation ، روش اعتبار سنجی n بار تکرار می شود و n پارتیشن تصادفی از نمونه اصلی حاصل می شود. نتایج n بازهم به طور متوسط با هم ترکیب می شوند تا یک تخمین واحد تولید کنند. اگر پنج بار این کار را انجام دهید و سپس از معادله زیر برای الگوریتم های A و B استفاده کنید ، 5X2-fold Cross validation را نشان داده اید.



منبع: <http://aivi.csee.usf.edu/~hall/dm/Eval2fold.pdf>

**بخش 4**

**سوال 1**

رگرسیون لجستیک ، به طور پیش فرض ، به مشکلات طبقه بندی دو کلاس محدود می شود. برخی از برنامه های افزودنی مانند one-vs-rest می توانند از رگرسیون لجستیک برای مشکلات طبقه بندی چند کلاس استفاده کنند ، اگرچه آنها نیاز دارند که مسئله طبقه بندی ابتدا به چندین مسئله طبقه بندی باینری تبدیل شود. الگوریتم رگرسیون لجستیک چند کلاس شامل تغییر تابع هزینه به توزیع احتمال چند جمله ای است.

**سوال2**

|  |  |  |  |
| --- | --- | --- | --- |
| recall\_score | precision\_score | f1\_score | سوال |
| 0.7941120607787274 | 0.7878174982318125 | 0.7878174982318125 | سوال 4 |
| 0.8268352749965882 | 0.8303643090042638 | 0.8264242782386926 | سوال 5 |
| 4.12% | 5.40% | 4.90% | میزان بهبود 5 به 4 |

همانطور که در جدول مشاهده میکنید، متریک های عملکرد در سوال 5 درصد بهبود داشته اند.

**سوال3**

|  |  |  |  |
| --- | --- | --- | --- |
| recall\_score | precision\_score | f1\_score | سوال |
| 0.677805 | 0.665207491 | 0.670662318 | سوال 1 |
| 0.716901203 | 0.697554937 | 0.704981995 | سوال 6 |
| 5.12% | 4.86% | 5.77% | میزان بهبود 6 به 1 |

همانطور که مشاهده میکنید، متریک های عملکردی در سوال 6 بهبود داشته اند.

دلایل Feature Selection:

* بهبود کارایی الگوریتم‌های یادگیری ماشین
* درک داده، کسب دانش درباره فرآیند و کمک به بصری‌سازی آن
* کاهش داده کلی، محدود کردن نیازمندی‌ها ذخیره‌سازی و احتمالا کمک به کاهش هزینه‌ها
* کاهش مجموعه ویژگی‌ها، ذخیره‌سازی منابع در دور بعدی گردآوری داده یا در طول بهره‌برداری
* سادگی و قابلیت استفاده از مدل‌های ساده‌تر و کسب سرعت

**سوال4**

Filter Methods

این روش ها از یک معیار آماری برای تخصیص یک score به هر ویژگی استفاده می نمایند. ویژگی ها با این score امتیازدهی می شوند و برای نگه داشتن و یا حذف از دیتاست انتخاب می شوند. این روش ها غالبا تک متغیره (univariate) هستند و هرویژگی را به صورت مستقل و یا با توجه به متغیرهای وابسته دیگر در نظر می گیرند. تعدادی از روش های این دسته عبارتند از Chi squared test، information gain و correlation coefficient scores.

Wrapper Methods

این روش ها مساله انتخاب ویژگی را به عنوان یک مساله جستجو در نظر می گیرند. در این روش ترکیب های مختلفی آماده می شوند، مورد ارزیابی قرار می گیرند و با ترکیب های دیگر مقایسه می شوند. یک مدل پیش بینی برای ارزیابی ترکیبی از ویژگی ها و تخصیص یک score بر پایه دقت به کار می رود. فرایند جستجو ممکنه است بر مبنای یک متدولوژی باشد مانند جستجوی best-first باشد، ممکن است تصادفی باشد مانند الگوریتم random hill-climbing و یا ممکن است heuristic باشد مانند گذرهای forward و backward برای اضافه و یا حذف کردن ویژگی ها.

**سوال 5**

انتخاب ویژگی Forward و Backward روش تضمین شده ای نیست که بهترین مدل پیش بینی کننده را به ما بدهد و این هزینه و تریدآفی است که برای دور ماندن از بیش برازش باید در نظر بگیریم. از طرفی عیب رایج روشهای فیلتر این است که آنها تعامل با طبقه بندی را نادیده می گیرند و هر ویژگی به طور مستقل در نظر گرفته می شود بنابراین از وابستگی به ویژگی چشم پوشی می شود. بعلاوه ، مشخص نیست که چگونه می توان برای رتبه بندی ترشولد را تعیین کرد تا فقط ویژگی های مورد نیاز انتخاب شود و نویز حذف شود.

**سوال6**

تحلیل تفکیک کننده خطی (به انگلیسی: Linear Discriminant Analysis، به طور مخفف LDA) و تفکیک کننده خطی فیشر روش‌های آماری هستند که از جمله در یادگیری ماشین برای پیدا کردن ترکیب خطی خصوصیاتی که به بهترین صورت دو یا چند کلاس از اشیا را از هم جدا می‌کند، استفاده می‌شوند. در این روش متغیر وابسته به صورت یک ترکیب خطی از متغیرهای دیگر مدل‌سازی می‌شود. با این حال دو روش آخر متغیر وابسته را از نوع فاصله‌ای در نظر می‌گیرند در حالی که تحلیل تشخیص خطی برای متغیرهای وابسته اسمی یا رتبه‌ای به کار می‌رود. از این رو تحلیل تشخیص خطی به رگرسیون لجستیک شباهت بیشتری دارد. تحلیل تشخیص خطی همچنین با تحلیل مؤلفه‌های اصلی و تحلیل عاملی هم شباهت دارد؛ هر دوی این روش‌های آماری برای ترکیب خطی متغیرها به شکلی که داده را به بهترین نحو توضیح بدهد به کار می‌روندیک کاربرد عمده هر دوی این روش‌ها، کاستن تعداد بعدهای داده است. با این حال این روش‌ها تفاوت عمده‌ای با هم دارند: در تحلیل تشخیص خطی، تفاوت کلاس‌ها مدل‌سازی می‌شود در حالی که در تحلیل مؤلفه‌های اصلی تفاوت کلاس‌ها نادیده گرفته می‌شود.

**لزوما باعث نتایج بهتر نمیشود**، نتایج طبقه بندی توسط مدل رگرسیون لجستیک پس از PCA و LDA تقریباً مشابه است. دلیل اصلی این شباهت در نتیجه این است که ما از مجموعه های داده مشابهی در این دو پیاده سازی استفاده کرده ایم. زیرا بین متغیرهای ورودی و خروجی رابطه خطی وجود دارد. وظیفه کاهش تعداد ویژگی های ورودی بود. هر دو تکنیک کاهش بعد مشابه هستند اما هر دو دارای یک استراتژی متفاوت و الگوریتم های مختلف هستند.

منبع: <https://analyticsindiamag.com/practical-approach-to-dimensionality-reduction-using-pca-lda-and-kernel-pca/>

**سوال 7 (امتیازی)**

مقایسه روشهای یادگیری ماشین و انتخاب مدل نهایی عملی معمول در یادگیری ماشین کاربردی است. مدلها معمولاً با استفاده از روشهای نمونه گیری مجدد مانند crossvalidation ارزیابی می شوند که میانگین skill scores به طور مستقیم محاسبه و مقایسه می شوند. اگرچه ساده است ، اما این روش می تواند گمراه کننده باشد ، زیرا دشوار است فهمید که آیا تفاوت بین میانگین نمرات واقعی است یا نتیجه یک شباهت آماری.

Statistical significance tests با توجه به این فرض که از یک توزیع یکسان گرفته شده اند ، برای رفع این مشکل و کمی سازی likelihood مشاهده شده است. اگر این فرض یا فرضیه صفر رد شود ، نشان می دهد که اختلاف skill scores از نظر آماری معنی دار است. تست فرضیه آماری اگرچه بی عیب و نقص نیست ، اما می تواند اعتماد به نفس شما را در تفسیر و همچنین ارائه نتایج در هنگام انتخاب مدل بهبود بخشد. دو توضیح را ابتدا باید بدانیم: **آزمون آماری صفر**: از آزمون آماری فرضیه صفر برای مقایسه دو نمونه داده و محاسبه اطمینان آماری استفاده می شود. فرضیه صفر عبارت است از اینکه تفاوتی بین توزیع دو نمونه داده وجود ندارد. که هرگونه واریانس به دلیل سر و صدا یا شانس دیده می شود. **P-Value:** این آزمون نشان می دهد که آیا ما باید این فرضیه صفر را که هیچ تفاوتی وجود ندارد را رد کنیم یا رد نکنیم. همانطور که crossvalidation داده ها را دوباره نمونه برداری می کند ، بین مجموعه های آموزشی همپوشانی وجود دارد. این بدان معنی است که مقادیر مستقل نیستند که اغلب فرضیه آزمونهای فرضیه آماری است.

**سوال 8**

پارامتر دیگری است که برای ارزیابی کارایی الگوریتم‌های یادگیری ماشین از آن استفاده می‌شود. این پارامتر بیان‌گر کیفیت کلاس‌بندی برای یک مجموعه باینری می‌باشد. (MCC (Matthews correlation coefficient، سنجه‌ای است که بیان‌گر بستگی مابین مقادیر مشاهده شده از کلاس باینری و مقادیر پیش‌بینی شده از آن می‌باشد. مقادیر مورد انتظار برای این کمیت در بازه 1- و 1 متغیر می‌باشد. مقدار 1+، نشان دهنده پیش‌بینی دقیق و بدون خطای الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار 0، نشان دهنده پیش‌بینی تصادفی الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار 1-، نشان دهنده عدم تطابق کامل مابین موارد پیش‌بینی شده از کلاس باینری و موارد مشاهده شده از آن می‌باشد.

