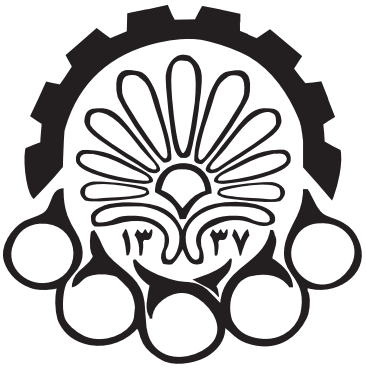


دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

عنوان

یادگیری ماشین

استاد درس

دکتر فاطمه شاکری

پروژه دوم

رگرسیون لجستیک و درخت تصمیم

مهر ۱۴۰۳

۱ مقدمه

در این تمرین، هدف در کنار پیاده‌سازی و مفاهیم توضیح داده شده در کلاس، آشنایی شما با مفاهیم شناخته شده و مرتبط با در یادیگری ماشین است که به دلیل کمبود وقت امکان مطرح شدن در کلاس را نداشته است. لطفا با توجه به کتابخانه‌هایی که استفاده می‌کنید در ابتدای فایل کد، قبل از اجرا، در صورتی که خروجی متغییر دارند، مقدار SEED را شماره‌ی گروه خود برای آن کتابخانه‌ها قرار داده و تعریف کنید تا در صورت اجرا از سمت ما نتایج با خروجی شما، یکسان باشد.

۲ به سوالات زیر به صورت پاسخ دهید:

۱. شیوه‌ی عمل k-Fold Cross Validation در کلاس توضیح داده شد، دو روش معروف دیگر Monte-Carlo Cross Validation و Nested Cross-Validation است، آن‌ها را توضیح داده و معایب و مزایای هر سه را بیان کنید. این روش‌ها چگونه بر روی بایاس و واریانس مدل اثر می‌گذرانند.

۲. در بخش زیادی از مجموعه داده‌ها مقدار نمونه‌ی موجود از هر کلاس نزدیک به هم نیست. در اصطلاح به چنین مجموعه داده‌هایی Imbalanced Dataset می‌گویند. نتیجه‌ی بایانس واریانس یک مدل با چنین مجموعه داده‌ای چگونه است؟ راهکارهایی که برای حل این مشکل وجود دارد را تحقیق کرده و بیان کنید.

۳. چگونه Vectorization ، کارایی محاسباتی را در پردازش داده‌ها بهبود بخشیده است؟ تفاوت‌های استفاده از عملیات Vectorization در مقابل استفاده از روش‌های سنتی مانند حلقه چیست؟

۴. آیا دقت درخت تصمیم روی داده‌های آموزشی همواره صد درصد است؟ اگر جواب مثبت است با استدلال و اگر جواب منفی است با مثال نقض بیان کنید.

۵. هرس کردن را برای درخت مربوط به داده‌های صفحه ۳۳ (اسلاید درخت تصمیم) که قبلا بدست آورده‌ایم، با روش Replacement با معیار Cost-Complexity یک بار با $\alpha = 0.1$ و یک بار با مقدار $\alpha = 0.9$ انجام دهید.

۶. در مورد ریشه‌ی نام‌گذاری روش Bootstrap ، مزایا، معایب آن و نیز شیوه‌ی ارزیابی Bootstrap 0.632 تحقیق کنید.

۳ کد

منظور از تحلیل کنید، بیان دقت، Recall، Precision، F1 score و رسم ماتریس درهم ریختگی و برداشت خود از آن‌ها است.

۱.۳ درخت تصمیم

در این بخش با استفاده از کلاس درخت تصمیم موجود در سایکیت لرن [۱] به اجرای مدل بر روی مجموعه داده خواهید پرداخت. درباره‌ی نوع خروجی، امکانات و گزینه‌های اختیاری آن قبل از انجام مراحل پایین تحقیق کنید (از موارد مهم که باید در قسمت تحلیل استفاده کنید plot_tree از بخش sklearn.tree می‌باشد). توضیح مناسبی درباره‌ی جزئیاتی که مورد استفاده قرار دادید، بدهید.

در گام‌های آینده، مجموعه داده‌ی گفته شده را با حفظ توزیع [۲] به ۷۰ درصد جهت آموزش و ۳۰ درصد جهت آزمایش تقسیم کنید.

- **گام ۱:** ابتدا با توجه به فراداده‌ی مجموعه داده (اینجا کلیک کنید) یک تحلیل کاوشگرانه بر روی داده‌ها [۳] انجام داده و پیش پردازش لازم بر روی مجموعه داده‌ی خود را مطابق تحلیل خود انجام دهید.
- **گام ۲:** با استفاده از درخت تصمیم موجود در سایکیت لرن، مدل خود را بر روی مجموعه داده آموزش، آموزش داده و خروجی را بر روی مجموعه آزمایش، تحلیل کنید.
- **گام ۳:** با انجام هرس پسین بر درخت گام قبل، از overfit شدن بر مجموعه آموزش جلوگیری کنید و خروجی مجموعه آزمایش این مدل را با مدل گام قبل مقایسه کنید. در کنار تحلیل، نمودار ROC را رسم کرده و مقدار AUC را برای هر یک محاسبه کنید .
- **گام ۴:** الگوریتم رگرسیون لجستیک از سایکیت لرن را بر روی مجموعه داده پیاده و نتیجه را تحلیل کنید. به مقایسه نتیجه‌ی این گام و گام قبل بپردازید.
- **گام ۵:** بردار احتمالاتی خروجی گام ۳ و گام ۴ را میانگین گرفته و نتیجه‌ی خروجی را تحلیل کنید.

[1] scikit-learn
[2] Stratify
[3] Exploratory data analysis

۲.۳ رگرسیون لجستیک

توجه کنید که انتخاب ابرپارامترهای تعداد دوره‌ی آموزش و مقدار منظم سازی بر عهده‌ی خود شما می‌باشد.

همچنین در گام‌های آینده، مجموعه داده‌ی گفته شده را با حفظ توزیع به ۷۰ درصد جهت آموزش و ۳۰ درصد جهت آزمایش تقسیم کنید.

- **گام ۱:** کدهای خواسته شده در بخش نوت بوک را کامل کنید. باید حتما از عملیات vectorized استفاده کنید. مفهوم vectorized در سوالات بخش اول خواسته شده است.
- **گام ۲:** ۱۰۰ نمونه از هرکدام از ۲ توزیع زیر که هر یک، یک کلاس را نمایندگی می‌کنند، تولید کرده و به عنوان مجموعه داده‌ی خود در نظر بگیرید. مقادیر X را به صورت یونیفرم تولید کنید.

$$y_i = -5 * \sin(0.5X_i) + 12 + \text{نویز}$$

$$y'_i = 5 * \sin(0.5X'_i) + \text{نویز}$$

- **گام ۳:** مدل رگرسیون لجستیک خود را بر روی داده‌ی آموزش، بدون تغییر دادن مقدار پیش فرض برای degree آموزش دهید. با استفاده از تابع plot_decision_boundary داده شده در کد، نتیجه‌ی مدل را نمایش داده و تحلیل کنید. نتیجه‌ی مدل بر روی مجموعه داده‌ی آزمایش چگونه است.
- **گام ۴:** در کد یک تابع با نام transform_features تعریف شده است که از PolynomialFeatures در sklearn.preprocessing استفاده می‌کند. توضیح کوتاهی از این تابع بدهید.
- **گام ۵:** مقدار پیش فرض شده‌ی degree را به ۳ تغییر داده و گام سوم را تکرار کنید. دلیل تغییر خروجی را صرفا بیان کنید.

۳.۳ کالیبراسیون مدل

هدف کالیبره کردن خروجی‌های احتمالاتی درخت تصمیم با استفاده از رگرسیون لجستیک برای بهبود دقت احتمالات است.

- **گام ۱:** مجموعه داده‌ها را به سه بخش آموزشی، آزمایش و کالیبراسیون (۷۰ - ۱۵ - ۱۵ درصد) با حفظ توزیع تقسیم کنید.
- **گام ۲:** درخت تصمیم را با استفاده از مجموعه آموزشی آموزش دهید و آن را روی مجموعه آزمایش ارزیابی کنید. نتیجه را تحلیل کنید.
- **گام ۳:** از مجموعه کالیبراسیون برای کالیبره کردن خروجی‌های احتمالی درخت تصمیم با استفاده از رگرسیون لجستیک استفاده کنید. در اینجا، احتمالات پیش‌بینی شده توسط درخت تصمیم به عنوان ورودی به مدل رگرسیون لجستیک داده می‌شوند.

- **گام ۴:** احتمالات کالیبره شده را با احتمالات اصلی مقایسه کرده و نمودارهای قابلیت اطمینان [۴] را رسم کنید.
- **گام ۵:** برای هر دو مدل درخت تصمیم اصلی و درخت تصمیم کالیبره شده، نمودار قابلیت اطمینان را رسم کنید. محور افقی باید نشان دهنده‌ی احتمالات پیش‌بینی‌شده و محور عمودی باید نشان دهنده‌ی احتمالات واقعی (نسبت مثبت‌ها) باشد. یک خط مرجع که نشان‌دهنده یک مدل کاملاً کالیبره شده است ($y = x$) اضافه کنید.
- **گام ۶:** تفاوت‌های بین مدل اصلی و کالیبره شده را بر اساس نمودار قابلیت اطمینان توضیح دهید.

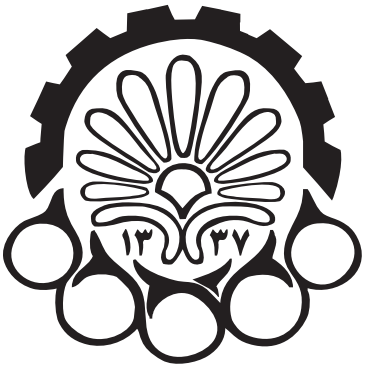
۴.۳ موارد تحویلی

۱. یک نوت‌بوک Jupyter (.ipynb) شامل موارد خواسته شده در صورت پروژه

۲. یک گزارش در قالب PDF با توجه به قالب داده شده.

۳. هر دو را در قالب یک فایل ZIP با نام HW2#groupid ارسال کنید.

[4] Reliability Diagrams



Amirkabir University of Technology

(Tehran Polytechnic)

Department Of Mathematics and Computer Science

Course

Machine Learning

Instructor

Dr. Fatemeh Shakeri

Exercise Two

Logistic Regression and Decision Tree

Oct 2024