|  |
| --- |
|  |

**بررسی مزایای الگوریتم XGBoost نسبت به الگوریتم درخت تصمیم و جنگل تصادفی**

**یک الگوریتم قدرتمند یادگیری ماشین است که به خانواده روش‌های توده‌ای تعلق دارد. این الگوریتم بر اساس ایده boosting گرادیان استوار است. در این روش چندین درخت تصمیم به صورت تکراری ساخته می‌شود و هر درخت به بازه‌های باقی‌مانده از مدل قبلی برازش می‌شوددر این الگوریتم به ازای هر بار برازش جنگل تصادفی میزان داده های به اشتباه تفسیر شده (خطا) محاسبه می گردد و در گام بعدی این داده های خطا به عنوان ورودی به جنگل های تصادفی بعدی انتقال می یابد.**

**1)پردازش اولیه**

**الف) نخست دیتاست مربوط به پیاده سازی الگوریتم درخت تصمیم را فراخوانی کرده و از دیتاست ازمون و اصلی ان به عنوان برازش استفاده نمایید همچنین کد معیار به ادرس [xgboost](https://www.kaggle.com/code/dansbecker/xgboost) را فراخوانی کرده و سلول های متناظر با مطلوب مسئله را استخراج نمایید**

**ب)اطلاعات کلی دیتاست را نمایش دهید و داده های از دست رفته را به روش Miss forest تکمیل نمایید**

**پ) داده ها را به شکل نرمال تبدیل نمایید**

**ج)ماتریس همبستگی را محاسبه نمایید و متغیر هایی که بیشترین همبستگی را با متغیر هدف دارند را به عنوان ورودی (feature) در نظر بگیرید**

**د)دیتاست را به سه بخش train test validation تفشیم کنید( train= 80% , test= 10% , validation =10%)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2. پیاده سازی الگوریتم XGBoost**

**الف)پارامتر های اساسی (hyperparameter) در الگوریتم XGBoost را نام ببرید و تاثیر ان را بر overfitting شرح دهید.**

**ب)مدل مذکور را با شرایط (n\_stimator=[20 40 60 80 100 110] ) و (learning rate=[0.1 0.2 0.3 0.01 0.02 0.03]) ارزیابی کنید**

**ج) خطا را در تمام شرایط مذکور به روش mean squared error بدست اورید**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**3. ارزش یابی و بهبود مدل (Fine tuning)**

**الف) نموار تابع خطا برای داده های train و validation را برای 6 شرط مذکور ترسیم نمایید به نحوی که محور افقی مقدار پارامتر می باشد و محور عمودی میزان تابع خطا می باشد.این روند را به ازای دو پارامتر n\_stimator و learnig rate انجام دهید**

**ب) مشخص کنید کدام میزان برای دو پارامتر اصلی دارای بهترین عملکرد نسبت به داده های validation می باشد.**

**ج) حال با داشتن پارامتر های بهینه، این مدل را با مدل درخت تصمیم و جنگل تصادفی بر حسب ماتریس در هم ریختگی (confusion matrix) مقایسه کنید و نمودار های مربوط به ان را رسم کنید.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**هنگام تحویل تمرین به نکات زیر توجه فرمایید:**

* **تمرین های دارای پیاده سازی، باید هم شامل کدها و هم شامل گزارش مربوطه باشد.**
* **همفکری و همکاری در یافتن پاسخ سوال ها توصیه می شود**
* **در صورت کپی بودن یکی یا چند مورد از پاسخ ها، کل نمره تمرین برای طرفین کپی در نظر گرفته نمی شود.**
* **حتما کدهای ارائه شده با فرمت نوت بوک (با پسوند .ipynb) به همراه گزارش کار در یک فایل زیپ ارائه شود.**
* **از آن‌جایی که تکلیف بعدی پس از اتمام ددلاین ارسال می‌شود از دیرکرد در تحویل تکلیف خودداری کرده و سر موعد مقرر تحویل دهید.**
* **تمارین به صورت الکترونیکی و به ایمیل تدریسیار (**[**zobeiria1@gmail.com**](mailto:zobeiria1@gmail.com)**) ارسال شود**