تکلیف شماره:

۹

آخرین زمان تحویل:

جمعه، ۵ دی ۱۴۰۴، ساعت ۸ صبح

با آرزوی موفقیت برای شما

توضیح:

این تکلیف شامل دو قسمت ۱) تحلیلی/تحقیقی و ۲) برنامه‌نویسی می‌باشد.

برای قسمت اول سعی شده که دانش قبلی شما ارزیابی شود.

هر سوال را در یک صفحه جداگانه جواب بدهید (به عبارت دیگه، دو پاسخ همزمان در یک صفحه نباشند).

برای نوشتن متن فارسی از فونت B Nazanin (14) و برای متن انگلیسی از فونت Times New Roman (12) استفاده کنید.

سوالات تحلیلی/تحقیقی

سوال ۱: تفاوت بین Correlation و Covariance چیست؟

سوال ۲: تفاوت بین Correlation و Dependency چیست؟

سوال ۳: تفاوت بین Classification و Regression چیست؟

سوال ۴: اثر نویز بر bias-variance چیست؟

سوال ۵: عدم قطعیت چیست؟ عدم قطعیت مدل متاثر از چه چیزیست؟ عدم قطعیت داده چیست؟

سوال ۶: تفاوت بین قدم قطعیت و High Dimension چیست؟ چه زمان این دو به هم مرتبط هستند و چه زمان نیستند؟

سوال ۷: تفاوت بین نویز سفید با نویز رنگی چیست؟

سوال ۸: تفاوت بین نویز و outlier در چیست؟

سوال ۹: تفاوت بین نویز و anomaly در چیست؟

سوال ۱۰: تفاوت بین outlier و anomaly در چیست؟

سوالات برنامه‌نویسی

تمرین شماره ۱

دیتاست مصنوعی بسازید که رابطه واقعی آن:

$$y = 2 - x_1 + 3x_2 + \mathcal{N}(\mu, \sigma^2)$$

جاییکه $x_1 = x^3, x_2 = \sin(x)$ هست و تعداد داده‌ها ۱۰۰۰ نمونه داده و x بین بازه 0 تا 2π باشه.

الف) مدل اول: با استفاده از تنها فیچر x^2 و بایاس $b = 2$ ،یک مدل خطی را در شرایطی که $\mu = 1, \sigma^2 = 0.2$ هست استفاده کنید و یک مدل Linear Regression را بسازید. خطای MSE, RMSE, NDEI را بر روی دیتا Train و Test محاسبه و گزارش دهید. (در هر دو حالت بدون نرمال سازی و همراه با نرمالسازی انجام دهید).

ب) مدل دومی در شرایطی که هر دو فیچر x_1 , x_2 حضور دارند ولی بایاس b در مدل ملاحظه نشده باشد.

ج) مدل سومی در حالتی که شرایط حالت (ب) وجود دارد و اینبار تخمین پارامتر b در مدلسازی خطی دیده شده باشد.

د) تغیرات پارامتر نویز را در مدلسازی لحاظ کنید. (افزایش و کاهش مقدار میانگین و واریانس نویز)

راهنمایی ساخت مدل Linear Regression:

لطفا برای ساخت مدل علاوه بر مدل Linear Regression موجود در ScikitLearn از مدل Least Square هم استفاده کنید. این مدل به صورت زیر قابل تعریف در محیط ماتریسی پایتون میباشد (شما میتوانید از پکیجهای موجود هم برای محاسبه پارامتر هایLS استفاده کنید. در حالیکه شما میتوانید با تحقیق و جستجوی ریز متوجه نحوه نوشتن این تخمین در محیط پایتون شوید، همچنان میتوانید با کمک از هوش مصنوعی نحوه نوشتن رابطه زیر را بدست بیاورید):

$$\theta_{LS} = (X^T X)^{-1} X^T Y$$

تمرین شماره ۲:

هدف: تخمین پذیرش دوباره بیماران دیابتی در بازه ۳۰ روزه (یعنی با چه احتمالی بیمار بعد از مراجعه به پزشک، باز هم در بازه ۳۰ روزه بعد از مراجعه قبلی به دکتر مراجعه میکند!)

(انجام این تسک باعث میشه که یکبار دیگه از دانش پایه پایتون (مثل تابع، کار با داده هاش Nan و داده های خراب) استفاده کنید، بحث ارزش فیچرها رو بررسی کنین، بحث Feature Selection رو انجام بدین، دو مدل Random Forest و XGBoost را برای ساخت مدل استفاده کنید، و نهایتا بحث اثر imbalance در داده ها را بررسی کنید. بدلیل اینکه قسمت اصلی این پروژه از ابتدا مبتنی بر زبان انگلیسی بوده، همان حالت بدون تغییر مونده، یسری سوال هم به تسک اصلی اضافه شده در جهت راهبردی هدفمند سوال برای شما کاروندان)

Dataset: UCI "Diabetes 130-US hospitals for years 1999–2008"
• Link: https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008
• Size: ~100,000 records

Tasks:

1. Load and clean the data (handle Missing values marked "?")
2. Encode ICD-9 text descriptions.
3. Build a model to predict whether a patient is readmitted within 30 days.
   - Use two models, including Random Forest and XGBoost to predict the possibility of readmission.
   - Check the sensitivity for two different numbers of leafs, for example 100 and 300.
   - Use two main features, including 1) all features and 2) most important features (like, gender, age, ICD information, and other possible main features for your model discrimination.
4. Deliver code, brief report, and model artifacts.
5. Evaluation Questions & Preferred Answers:
   - How would you measure model accuracy for this task?
   - How would you check for bias in your model?
     - Evaluate the effect of imbalance data in the model performance. How can you tackle this challenge?
     - Write a constructive comparison between your findings and report them exhaustively.
   - How would you make your model's predictions explainable to clinicians?