دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

تمرین دوم درس یادگیری ماشین

Bayesian Decision Theory، Parametric Methods و Non-Parametric Methods

استاد محترم درس:

جناب آقای دکتر قادری

دستیاران آموزشی:

آبتین مفید

نیلوفر مقدس

نیمسال اول ۱۴۰۴-۱۴۰۳

نکات مربوط به تحویل پروژه

- ✓ پروژه فقط با زبان برنامه نویسی پایتون قابل قبول می‌باشد.
- ✓ فایل تحویلی شما، یک فایل زیپ شده‌ی نهایی شامل گزارش کار (فایل pdf و word) و فایل کد پایتون با پسوند ipynb (Jupiter Notebook) می‌باشد. لطفا آن را به صورت زیر نام‌گذاری و ارسال نمایید.
- ✓

HW_۲_[LastName]_[FirstName]

- ✓ گزارش کار خود را در یک فایل pdf و word تحویل دهید و از گذاشتن صرفا اسکرین شات‌های پشت سرهم از کد در گزارش کار خودداری کنید.
- ✓ توجه داشته باشید که در فایل ارسالی پایتون، خروجی هر سلول (شامل نمودار، خروجی عددی و غیره) حتما ذخیره شده و قابل مشاهده باشد.
- ✓ لازم است حتما نتایج بدست آمده را گزارش و تحلیل کنید.
- ✓ علاوه بر مهارت حل سوالات، نوشتن پاسخ مینی پروژه‌ها در فرمت گزارش فنی (فصل بندی و صفحه بندی مناسب، رعایت اصول نگارش، درج زیرنویس برای شکل‌ها و بالانویس برای جداول و اشاره به شماره شکل یا جدول در متن و ...) برای دانشجویان تحصیلات تکمیلی اهمیت دارد، این مورد نیز در ارزشیابی لحاظ می‌شود.
- ✓ در صورت فراموشی در ارسال کد پایتون، هیچ نمره‌ای به شما تعلق نخواهد گرفت.
- ✓ در صورت مشاهده تشابه در هر بخش از انجام پروژه، نمره هر دو نفر صفر لحاظ می‌گردد.

- ✓ **ایمیل دستیار آموزشی: Abtin۱۳۷۸۱۳۷۸@gmail.com**

مقدمه

توضیح اولیه و هدف پروژه

هدف این پروژه، مقایسه عملکرد مدل‌های پارامتریک و ناپارامتریک در دسته‌بندی داده‌ها است. در این پروژه، دانشجویان با استفاده از مجموعه داده‌ای شامل ویژگی‌های شیمیایی شراب، دو مدل پارامتریک و ناپارامتریک را ارزیابی و با یکدیگر مقایسه می‌کنند. همچنین، در انتها با اصول مدل بیزین ساده آشنا می‌شوند و عملکرد آن را بررسی می‌کنند.

مجموعه داده

مجموعه داده‌ی مورد استفاده، **Wine** از کتابخانه‌ی SciKit-Learn است که شامل ۱۳ ویژگی شیمیایی از انواع مختلف شراب است. این مجموعه داده ۳ کلاس مختلف شراب را شامل می‌شود و ویژگی‌هایی مانند میزان الکل و فلانوئیدها برای هر شراب در آن موجود است.

معیارهای ارزیابی

در این پروژه، مدل‌ها با استفاده از معیارهای ارزیابی زیر تحلیل و مقایسه می‌شوند:

۱. **صحت**[۱]: صحت، نسبت تعداد نمونه‌های درست دسته‌بندی‌شده به کل نمونه‌ها است و به عنوان یک معیار کلی برای اندازه‌گیری عملکرد مدل‌ها استفاده می‌شود.

۲. **گزارش طبقه‌بندی**[۲]: این گزارش شامل معیارهای **دقت**، **فراخوانی** و **امتیاز F۱** برای هر کلاس است و به مقایسه عملکرد مدل‌ها در دسته‌بندی کلاس‌های مختلف کمک می‌کند:

- **دقت**[۳]: نشان‌دهنده درصد مواردی است که مدل به درستی به کلاس مورد نظر اختصاص داده است.
- **فراخوانی**[۴]: درصد نمونه‌های درست پیش‌بینی‌شده از کل نمونه‌های واقعی آن کلاس.
- **امتیاز F۱**[۵]: میانگین هارمونیک دقت و فراخوانی که در مواقعی که توزیع کلاس‌ها نامتعادل باشد، به خصوص اهمیت دارد.

۳. **ماتریس درهم‌ریختگی**[۶]: این ماتریس نشان‌دهنده‌ی تعداد نمونه‌های درست و نادرست طبقه‌بندی‌شده در هر کلاس است و به ارزیابی دقیق‌تر عملکرد مدل و بررسی موارد خاص کمک می‌کند.

✓ **برای درک بهتر از معیارهای ارزیابی متداول با مثال، می‌توانید به این لینک مراجعه فرمایید.**

---

۱ Accuracy
۲ Classification Report
۳ Precision
۴ Recall
۵ F۱-score
۶ Confusion Matrix

کتابخانه‌های مورد نیاز

برای انجام این پروژه از کتابخانه‌های pandas، matplotlib و scikit-learn استفاده می‌شود. جهت بارگذاری داده‌ها، پردازش، ترسیم نمودارها و پیاده‌سازی مدل‌های یادگیری ماشین، این کتابخانه‌ها ضروری هستند. برای نصب کتابخانه‌ها مطابق دستور زیر می‌توانید از دستور pip install استفاده کنید.

```
pip install pandas matplotlib scikit-learn
```

کدهای مورد نیاز برای بارگذاری و آماده‌سازی مجموعه داده

کدهای شکل ۱ برای بارگذاری و تقسیم داده‌ها به مجموعه آموزشی و تست (٪۸۰ آموزش و ٪۲۰ تست) استفاده می‌شوند:

```python
import pandas as pd
from sklearn.datasets import load_wine
from sklearn.model_selection import train_test_split

# بارگذاری مجموعه داده
wine = load_wine()
data = pd.DataFrame(wine.data, columns=wine.feature_names)
data['target'] = wine.target

# تقسیم داده‌ها به ٪۸۰ آموزشی و ٪۲۰ تست
X_train, X_test, y_train, y_test = train_test_split(data[wine.feature_names], data['target'], test_size=0.2, random_state=42)

# نمایش نمونه‌ای از داده‌ها
print(data.head())
```

شکل ۱ بارگذاری مجموعه داده و تقسیم داده ها

سوالات پروژه

سوال ۱: آماده‌سازی داده‌ها و تحلیل اولیه

سوال ۱.۱:

ابتدا مجموعه داده‌ی Wine را بارگذاری کرده و ویژگی‌های آن را به تفکیک کلاس‌ها بررسی کنید. برای هر ویژگی، یک نمودار هیستوگرام رسم کنید تا توزیع هر ویژگی بین کلاس‌ها مشخص شود.

سوال ۱.۲:

با توجه به توزیع ویژگی‌ها و نمودارهای رسم‌شده، فرضیاتی در مورد جداپذیری کلاس‌ها بیان کنید. آیا به نظر می‌رسد که مدل‌های پارامتریک مانند LDA برای دسته‌بندی این داده‌ها مناسب باشند؟ دلایل خود را توضیح دهید.

سوال ۲: پیاده‌سازی و ارزیابی مدل‌های پارامتریک و ناپارامتریک

سوال ۲.۱:

مدل پارامتریک LDA و مدل ناپارامتریک K-Nearest Neighbors با $k=۵$ را با استفاده از scikit-learn پیاده‌سازی کرده و هر یک را روی داده‌ها اعمال کنید. از روش ارزیابی ٪۸۰ داده‌های آموزشی و ٪۲۰ داده‌های تست استفاده کنید و دقت هر مدل را محاسبه کنید. برای ارزیابی دقیق‌تر، گزارش طبقه‌بندی و ماتریس درهم‌ریختگی هر مدل را نیز ارائه دهید.

سوال ۲.۲:

مقدار kرا در مدل K-Nearest Neighbors برای مقادیر ۱، ۳، ۵، ۱۰، ۱۵ و ۲۰ تغییر دهید و برای هر مقدار k، دقت مدل را محاسبه کنید. سپس نتایج را در نموداری نمایش دهید تا نحوه تغییر دقت با تغییر k مشخص شود. تحلیل کنید که در چه مقادیری از k مدل عملکرد بهتری داشته است و دلیل آن را توضیح دهید.

سوال ۳: پیاده‌سازی و تحلیل مدل بیزین

سوال ۳.۱:

یک مدل بیزین ساده (Naive Bayes) را با استفاده از scikit-learn برای دسته‌بندی داده‌های شراب پیاده‌سازی کنید. سپس دقت، گزارش طبقه‌بندی و ماتریس درهم‌ریختگی این مدل را با مدل‌های LDA و KNN مقایسه کنید.

سوال ۳.۲:

تحلیل کنید که در چه شرایطی مدل بیزین می‌تواند بهتر از مدل‌های پارامتریک و ناپارامتریک عمل کند. برای مثال، شرایطی مانند مستقل بودن ویژگی‌ها را در نظر بگیرید و توضیح دهید که چگونه این شرایط بر عملکرد مدل بیزین تاثیر می‌گذارند.