پروژه درس هوش مصنوعی

دسته بندی مشتریان بانک

دستیاران آموزشی: بهداد نادری فرد، سجاد مهرپیما، نفیسه احمدی

این پروژه به پیاده‌ســـازی و مقایســـه الگوریتم های یادگیری ماشـــین (K-Nearest Neighbors (KNN - ، درخت تصـــمیم (Decision Tree)، جنگل تصادفی(Random Forest) و Ada Boost می‌پردازد. هدف اصلی، ارزیابی عملکرد این الگوریتم ها در یک مسئله طبقه‌بندی دوتایی است.

ساختار پروژه

پروژه از چندین بخش اصلی تشکیل شده است:

1. **پیاده‌سازیKNN**

کلاس KNNClassifier با سه متد اصلی:

- __init__(k): مقداردهی اولیه با پارامترk
- fit(X, y): آموزش مدل با داده‌های ورودی
- predict(X): پیش‌بینی برچسب‌های نمونه‌های جدید

2. **پیش پردازش داده**

تابع load_and_preprocess_data مراحل زیر را انجام می‌دهد:

- بارگذاری داده‌ها
- نمونه‌برداری متوازن از هر کلاس
- کدگذاری متغیرهای کیفی
- تقسیم داده‌ها به سه مجموعه آموزش، اعتبارسنجی و آزمون
- نرمال‌سازی داده‌ها

3. **آموزش مدل‌ها**

دو تابع اصلی برای آموزش مدل‌ها:

train_decision_tree

- آموزش درخت تصمیم با عمق maximum depth برابرd

train_knn

- آموزش مدل KNN با پارامترk

- آموزش مدل جنگل تصادفی با پارامترهای:
- n_estimators: تعداد درختان در جنگل
- max_depth: حداکثر عمق هر درخت

train_adaboost

- آموزش مدل AdaBoost با پارامترهای:
- n_estimators: تعداد طبقه‌بندهای پایه

**** ستون لیبل‌ها، ستون bad_loans است.**

4. مقایسه مدل‌ها

تابع compare_models برای مقایسه بصری عملکرد دو مدل با استفاده از نمودار میله‌ای

5. تابع اصلی(main)

در تابع main، وظایف زیر انجام می‌شود:

- بارگذاری و پردازش داده‌ها
- بهینه‌سازی ابرپارامترها :
  - یافتن بهترین مقدار k برایKNN
  - یافتن بهترین عمق برای درخت تصمیم
- آموزش مدل‌های نهایی
- گزارش دقت روی مجموعه آزمون
- نمایش ساختار درخت تصمیم

نکات پیاده‌سازی

الف. پیاده‌سازیKNN

1. محاسبه فاصله اقلیدسی بین نقاط
2. یافتن k همسایه نزدیک
3. رأی‌گیری اکثریت برای تعیین برچسب

ب. تنظیم ابرپارامترها

1. برای KNN:
- آزمایش مقادیر مختلف
- استفاده از مجموعه اعتبارسنجی برای انتخاب بهترین k

2. برای درخت تصمیم :
- آزمایش عمق‌های مختلف

- جلوگیری از over-fitting با محدود کردن عمق

3. برای Random Forest:
   - بهینه‌سازی تعداد درختان (n_estimators)
   - تنظیم حداکثر عمق درختان برای جلوگیری از over-fitting
4. برای AdaBoost:
   - تنظیم تعداد طبقه‌بندهای پایه (n_estimators)

**** توجه کنید برای یافتن بهترین هایپر پارامترها برای Random Forest از GridSearch استفاده کنید.**

خروجی‌های مورد انتظار

1. نمودار مقایسه‌ای دقت مدل‌ها
2. نمایش گرافیکی ساختار درخت تصمیم
3. گزارش دقت نهایی روی مجموعه آزمون

معیارهای ارزیابی

- دقت (Accuracy) به عنوان معیار اصلی ارزیابی
- مقایسه زمان اجرا و پیچیدگی محاسباتی