

Homework4 – Programming Part

مقدمه

در این تمرین از داده‌های واقعی Heart Disease UCI استفاده می‌کنیم. هدف شما این است که چند مدل را از پایه پیاده‌سازی کنید و بعد آن‌ها را از نظر دقیق و شکل مرز تصمیم با هم مقایسه کنید.

در این تمرین یاد می‌گیرید:

- تفاوت‌های آماری و هندسی مدل‌ها را بشناسید.
- ببینید وقتی داده‌ها تغییر می‌کنند (مثل اضافه شدن نویز، حذف ویژگی یا اضافه شدن مقادیر پرت)
- عملکرد مدل‌ها چه تغییری می‌کند.

مدل‌هایی که بررسی می‌کنیم دو نوع هستند:

- مدل‌های مولد(Generative) : توزیع داده‌های هر کلاس را یاد می‌گیرند و با قانون بیز تصمیم می‌گیرند.
- مدل‌های متمایزکننده(Discriminative) : بدون مدل کردن توزیع داده، مستقیماً مرز جداکننده کلاس‌ها را پیدا می‌کنند.

مباحث پایه‌ای که باید پیاده‌سازی شوند

(QDA و LDA) Gaussian Classifier

- فرض می‌کنیم داده‌های هر کلاس از یک توزیع گاووسی چندبعدی می‌آیند.
- برای هر کلاس یک ماتریس کوواریانس جداگانه \rightarrow مرز تصمیم منحنی.
- یک ماتریس کوواریانس مشترک برای همه کلاس‌ها \rightarrow مرز تصمیم خطی.

Gaussian Naive Bayes

- فرض می‌کند ویژگی‌ها مستقل هستند.
- برای هر ویژگی و هر کلاس یک میانگین و واریانس محاسبه می‌شود.
- احتمال پسین با فرمول بیز بدست می‌آید.
- سریع و ساده، اما اگر ویژگی‌ها همبسته باشند دقت کاهش می‌یابد.

Homework4 – Programming Part

Logistic Regression

- احتمال تعلق به کلاس 1 را مدل می کند:

$$p(y = 1 \mid x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

- با حداکثر کردن لاغ-لایکلی هود آموزش داده می شود.
- از روش نیوتن برای به روزرسانی ضرایب استفاده می شود.
- مرز تصمیم خطی است.

Softmax Regression

- تعمیم Logistic Regression به مسائل چند کلاسه است.
- در حالت دو کلاسه به Logistic ساده کاهش می یابد.

ارزیابی و مقاومت

- معیارها: Accuracy, Precision, Recall, F1-score.
- آزمایش مقاومت: بررسی تغییر عملکرد با نویز، حذف ویژگی ها یا اضافه کردن آوتلایر.

مراحل تسکها

پردازش داده ها Task 1

- ستون های اضافی مثل id و dataset را حذف کنید.
- مقادیر گم شده را با میانگین یا مد پر کنید.
- نمودارهای هیستوگرام / KDE برای هر ویژگی بر اساس کلاس رسم کنید.
- دو ویژگی مفید انتخاب کنید (در این تمرين: chol و thalach).
- برای هر کلاس میانگین و کواریانس حساب کرده و کانتورهای توزیع گاوی را رسم کنید.
- با نمودارها بررسی کنید که فرض گاوی بودن چقدر منطقی است.

Homework4 – Programming Part

(Gaussian Classifier) مدل‌های مولد Task 2

- میانگین و کوواریانس هر کلاس را محاسبه کنید.
- مرز تصمیم را از طریق Log-Likelihood Ratio به دست آورید.
- LDA (کوواریانس مشترک) و QDA (کوواریانس جدگانه) را پیاده‌سازی و مقایسه کنید.
- مرز تصمیم هر دو رسم کنید و درباره اثر فرض اشتراک کوواریانس بحث کنید.

Gaussian Naive Bayes: Task 3

- برای هر ویژگی و هر کلاس میانگین و واریانس محاسبه کنید.
- با فرض استقلال ویژگی‌ها احتمال پسین را حساب کنید.
- مرز تصمیم را رسم و با LDA و QDA مقایسه کنید.
- توضیح دهید اگر ویژگی‌ها همبسته باشند چه اثری دارد.

Logistic Regression: Task 4

- تابع لاغ-لایکلی‌هود، گرادیان و هسیان را بنویسید.
- با روش نیوتون ضرایب را به روزرسانی کنید.
- نمودار همگرایی لاغ-لایکلی‌هود را رسم کنید.
- مرز تصمیم Logistic Regression را رسم و با مدل‌های Generative مقایسه کنید.

Task 5: ارزیابی و مقاومت مدل‌ها

- داده‌ها را با seed=42 به نسبت 20/80 به آموزش و آزمون تقسیم کنید.
- معیارهای Accuracy، Precision، Recall، F1-score را برای هر مدل محاسبه کنید.
- سه آزمایش مقاومت انجام دهید:
 1. افزودن نویز گاوی به داده‌ها.
 2. حذف ویژگی کلیدی (مثل chol).
 3. افزودن مقادیر پرت ($\pm 3\sigma$) به ویژگی‌ها.
- قبل و بعد از تغییرات مرز تصمیم و عملکرد را مقایسه کنید.

Homework4 – Programming Part

Task 6: مقایسه مولد و متمایزکننده

- مرز تصمیم Logistic Regression ، QDA و Naive Bayes را در یک نمودار رسم کنید.
- نتایج عددی معیارها را مقایسه کنید.
- مزایا و معایب هر رویکرد را توضیح دهید:
 - داده کم → بهتر / تفسیرپذیری بالاتر.
 - داده زیاد → عملکرد بهتر.

Task 7: سناریوی تصمیم‌گیری بالینی

- یک نمونه جدید تعریف کنید:

```
makefile  
CopyEdit  
age = 58  
chol = 245  
thalach = 140  
trestbps = 130
```

سایر ویژگی‌ها = میانگین هر ستون

- این نمونه را به هر مدل بدهید و احتمال کلاس 1 (بیماری قلبی) را محاسبه کنید.
- نتایج را عددی گزارش کرده و در نمودارها نمایش دهید.

Bonus ها

Bonus A: ویژگی غیرخطی و مرز تصمیم

- معمولی مرز تصمیم خطی دارد:

$$w_0 + w_1x_1 + w_2x_2 = 0$$

- اگر داده‌ها غیرخطی جدا شوند، ویژگی‌های غیرخطی اضافه کنید.
مثالاً:

$$\text{chol_squared} = (\text{chol})^2$$

Homework4 – Programming Part

مدل جدید:

$$p(y = 1 | x) = \frac{1}{1 + e^{-(w_0 + w_1 \text{chol} + w_2 \text{thalach} + w_3 \text{chol}^2)}}$$

در فضای اصلی مرز تصمیم دیگر خط نیست (منحنی می‌شود).
در فضای جدید (ویژگی‌های اصلی $+ \text{chol}^2$) مرز تصمیم همچنان خطی است.

Softmax Regression:Bonus B

برای چند کلاس:

$$p(y = k | x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^K \exp(w_j^T x)}$$

اگر فقط دو کلاس داشته باشیم:

$$p(y = 1 | x) = \frac{1}{1 + \exp(-(w_1 - w_0)^T x)}$$

پیاده‌سازی:

- وزن‌ها را برای هر کلاس جداگانه بگیرید.
- و لایکلیهود را بنویسید.
- گرادیان‌ها را محاسبه و وزن‌ها را به روزرسانی کنید.
- مرز تصمیم را رسم کنید و با Logistic Regression مقایسه کنید.