به نام خدا

تمرین سری ۴ ماشین لرنینگ

ایلدار صمدزاده طریقت ۴۰۰۳۳۰۴۴

1- تفاوت شاخص‌های ارزیابی(criterion) انتروپی و gini چیست؟

تفاوت اصلی بین شاخص‌های Gini و Entropy در ارزیابی گره‌ها در درخت تصمیم، در نحوه محاسبه نااطمینانی است .Entropy از فرمول اطلاعات شانون استفاده می‌کند و مقدار آن در بازه $[0, \log_2 C]$ برای C کلاس (تغییر می‌کند، در حالی که Gini impurity معمولاً ساده‌تر محاسبه می‌شود و مقدارش در بازه $[0, 0.5]$ برای دو کلاس است. از نظر عملکرد، تفاوت زیادی در دقت مدل ایجاد نمی‌کنند ولی Gini اغلب کمی سریع‌تر است چون محاسبه لگاریتم ندارد.

2- تاثیر حداکثر عمق (max depth) درخت بر واریانس و بایاس مدل چیست؟

افزایش حداکثر عمق درخت باعث کاهش بایاس و افزایش واریانس می‌شود، زیرا درخت عمیق‌تر می‌تواند الگوهای پیچیده‌تر را یاد بگیرد ولی در عوض بیشتر در معرض overfitting قرار می‌گیرد. عمق کم موجب بایاس بالا و underfitting می‌شود چون مدل نمی‌تواند روابط پیچیده را یاد بگیرد.

3- چگونه می‌توان رگرسیون لجستیک را برای مسأله‌های چندکلاسه (multiclass) تعمیم داد؟

برای تعمیم رگرسیون لجستیک به مسائل چندکلاسه از رویکردهایی مثل One-vs-Rest (OvR) یا One-vs-One (OvO) استفاده می‌شود. همچنین نسخه تعمیم‌یافته‌ای به نام Softmax Regression یا Multinomial Logistic Regression وجود دارد که احتمال تعلق به هر کلاس را به طور مستقیم مدل می‌کند.

4- پیچیدگی محاسباتی آموزش SVM چگونه با تعداد نمونه و ابعاد فضا رشد می‌کند؟

پیچیدگی محاسباتی آموزش SVM کلاسیک (با حل مسأله Quadratic Programming) در بدترین حالت برابر است با $O(n^3)$ نسبت به تعداد نمونه‌ها (n) و $O(n^2)$ از نظر حافظه. با افزایش ابعاد (d)، عملکرد تابع کرنل و تعداد ویژگی‌ها نیز به پیچیدگی کمک می‌کند ولی معمولاً تعداد نمونه تأثیر بیشتری دارد. برای داده‌های بزرگ، روش‌هایی مثل Linear SVM یا SMO استفاده می‌شوند.

5- مثالی از یک تابع Kernel متداول را نام ببرید و ویژگی‌اش را توضیح دهید.

یک تابع کرنل متداول، کرنل شعاعی پایه RBF) یا (Gaussian است:

$$K(x, x') = \exp(-\gamma\|x - x'\|^2)$$

این کرنل امکان نگاشت داده‌های غیرخطی به فضای ویژگی با ابعاد بالا را فراهم می‌کند و به خوبی می‌تواند مرزهای تصمیم پیچیده را مدل کند. پارامتر γ کنترل تأثیر نقاط داده را دارد.

6- چگونه SVM را برای مسائل چندکلاسه (multiclass) پیاده‌سازی می‌کنند؟

برای مسائل چندکلاسه در SVM از روش‌هایی مانند One-vs-Rest یا One-vs-One استفاده می‌شود. در OvR برای هر کلاس یک مدل ساخته می‌شود که آن کلاس را در برابر سایر کلاس‌ها تشخیص دهد. در OvO بین هر جفت کلاس، یک مدل جداگانه ساخته می‌شود و در نهایت با رأی‌گیری تصمیم‌گیری صورت می‌گیرد.

7- **مثالی از یک تابع Kernel متداول را نام ببرید و ویژگی‌اش را توضیح دهید.**

همان پاسخ سوال 5 تکرار شده است و کرنل RBF مناسب‌ترین مثال است.

8- **چرا در رگرسیون لجستیک از تابع سیگموید به‌جای تابع خطی استفاده می‌شود؟**

در رگرسیون لجستیک از تابع سیگموید استفاده می‌شود چون خروجی آن بین 0 و 1 قرار دارد و می‌تواند احتمال تعلق یک نمونه به کلاس مثبت را مدل کند. تابع خطی می‌تواند مقادیر منفی یا بزرگ‌تر از یک تولید کند که برای تفسیر احتمال مناسب نیست.

9- **اگر داده‌هایی که به صورت غیرخطی قابل تفکیک هستند داشته باشیم، آیا می‌توان از LDA استفاده کرد؟ چرا؟ چه راهکاری برای این شرایط پیشنهاد می‌کنید؟**

LDAفرض می‌کند که داده‌ها به صورت خطی قابل تفکیک هستند و توزیع هر کلاس گاوسی با کوواریانس یکسان دارد. بنابراین اگر داده‌ها غیرخطی تفکیک‌پذیر باشند، عملکرد LDA ضعیف می‌شود. در این شرایط می‌توان از Kernel LDA یا روش‌های غیرخطی مانند SVM با کرنل یا شبکه‌های عصبی استفاده کرد.

10- **در صورتی که ویژگی‌های ورودی به شدت به هم وابسته باشند (highly correlated features)، عملکرد Naive Bayes چگونه تغییر می‌کند؟ توضیح داده و مثال بزنید.**

در Naive Bayes فرض استقلال ویژگی‌ها وجود دارد. اگر ویژگی‌ها به شدت هم‌بسته باشند، این فرض نقض می‌شود و مدل دقت کمتری خواهد داشت. برای مثال اگر دو ویژگی دقیقاً یکسان باشند، Naive Bayesوزن آن‌ها را دو برابر در نظر می‌گیرد که باعث overconfidence می‌شود. راهکارهایی شامل انتخاب ویژگی، PCAیا استفاده از مدل‌هایی که فرض استقلال ندارند مانند Logistic Regression است.

11- **آیا LDA همیشه برای کاهش ابعاد بهتر از PCA عمل می‌کند؟ تحت چه شرایطی PCA می‌تواند نتیجه بهتری نسبت به LDA داشته باشد؟ پاسخ خود را تحلیل کنید.**

LDAنسبت به PCA از اطلاعات کلاس‌ها استفاده می‌کند و برای مسائل طبقه‌بندی عملکرد بهتری دارد. اما اگر توزیع کلاس‌ها از فرضیات LDA (گاوسی بودن و کوواریانس مشابه) پیروی نکنند یا اگر کلاس‌ها به صورت پیچیده‌تری تفکیک شوند، PCAممکن است عملکرد بهتری داشته باشد چون بر اساس بیشینه واریانس کل داده‌ها عمل می‌کند و محدود به کلاس‌ها نیست.

12- **فرض کنید داده‌هایی بسیار نامتوازن (Imbalanced) در اختیار شما قرار داده شده‌اند. استفاده از Random Forest چه مشکلاتی می‌تواند ایجاد کند؟ چه راهکارهایی برای بهبود پیشنهاد می‌کنید؟**

در داده‌های نامتوازن، Random Forest ممکن است به سمت کلاس غالب متمایل شود چون درخت‌ها بیشتر بر اساس آن آموزش می‌بینند. این مسئله باعث کاهش دقت در کلاس اقلیت می‌شود. راهکارهایی مانند استفاده از undersampling، oversampling، SMOTE، استفاده از وزن‌دهی کلاس یا الگوریتم‌هایی مثل Balanced Random Forest مفید هستند.

13- در Bagging چرا استفاده از نمونه‌گیری با جایگزینی (Bootstrapping) بهتر از نمونه‌گیری بدون جایگزینی است؟ چه تاثیری روی واریانس و بایاس مدل دارد؟

Bootstrappingبا جایگزینی تنوع بیشتری به مجموعه‌های داده می‌دهد و این باعث کاهش واریانس مدل در Bagging می‌شود. اگر از نمونه‌گیری بدون جایگزینی استفاده شود، تنوع کمتری بین مدل‌های پایه ایجاد می‌شود و کاهش واریانس به اندازه مطلوب رخ نمی‌دهد. اما بایاس معمولاً تفاوت زیادی نمی‌کند.

14- اگر در مسئله‌ای داده‌های زیاد، نویزی و با کلاس‌های متعدد داشته باشیم، کدام یک از classifierهای LDA، Naive Bayes و یا Random Forest بهتر عمل می‌کنند؟ دلیل انتخاب خود را توضیح دهید.

در شرایطی که داده‌ها زیاد، نویزی و کلاس‌ها متعدد باشند، Random Forestعملکرد بهتری نسبت به LDA و Naive Bayesدارد چون مقاوم به نویز است، فرضیاتی در مورد توزیع ندارد و می‌تواند روابط غیرخطی را مدل کند .LDA در صورت نقض فرضیات گاوسی بودن و کوواریانس برابر ضعیف عمل می‌کند، و Naive Bayes با فرض استقلال ویژگی‌ها نسبت به نویز حساس‌تر است.

1- برای حداکثر عمق 2و3 و 4 در الگوریتم **Decision Tree**، میزان دقت و صحت الگوریتم درخت تصمیم را بر روی دیتاست **diabetes** مقایسه کنید وعلت اختلاف آن‌را بررسی کنید.

عمق 2 و 3 عملکرد مشابهی دارند: دقت کل، precisionو f1-score برای هر دو تقریباً یکسان است. به احتمال زیاد، با توجه به min_samples_split=100 ، مدل عملاً با عمق 2 و 3 تفاوت ساختاری چندانی ندارد، یا اینکه تقسیم‌های جدید اطلاعات زیادی اضافه نمی‌کنند.

در عمق 4 دقت افت کرده است: با اینکه Recall برای کلاس 1 در عمق 4 بهتر شده (یعنی مدل بیماران مبتلا را بیشتر تشخیص می‌دهد)، اما این به قیمت افزایش خطا در کلاس 0) افزایش (FP تمام شده است که دقت کلی را کاهش داده.

علت اختلاف عملکرد: با افزایش عمق، درخت تصمیم می‌تواند داده‌ها را دقیق‌تر تقسیم کند ولی اگر داده‌ها نویزی باشند یا تعداد نمونه‌ها برای تقسیم‌های دقیق کم باشد) به‌ویژه با(min_samples_split=100 ، مدل دچار overfitting روی ویژگی‌های غیرعمومی می‌شود، که منجر به افت دقت کلی در داده‌های تست می‌شود.

در این تنظیمات خاص (با min_samples_split=100 و criterion=gini)، مدل با عمق 2 یا 3 عملکرد بهتری دارد و افزایش عمق به 4 باعث کاهش دقت می‌شود. بهترین انتخاب بین عمق 2 و 3 وابسته به اولویت شما بین Recall یا Precision برای کلاس 1 است، ولی تفاوت آن‌ها در اینجا ناچیز است.

2- الگوریتم SVM را با دو kernel مختلف خطی و radial basis function پیاده سازی کنید و خروجی ها را با یکدیگر مقایسه کنید و تفسیر خود را از اختلاف آن‌ها بنویسید.

عملکرد بهتر kernel خطی در این داده‌ها: دقت کلی و F1-score برای هر دو کلاس با kernel خطی بالاتر است. این نشان می‌دهد که داده‌ها تا حد زیادی در فضای ویژگی به صورت خطی قابل جداسازی هستند یا حداقل جداسازی خطی بهتر تعادل بین precision و recall را حفظ کرده است.

کاهش recall در کلاس 1 با kernel RBF: SVM با kernel RBF حساسیت (recall) پایین‌تری برای کلاس 1 دارد. به بیان دیگر، موارد مثبت (مثلاً بیماران دیابتی) را بیشتر از دست می‌دهد که از نظر کاربردی ممکن است نامطلوب باشد.

انعطاف‌پذیری RBF و خطر overfitting: kernel شعاعی (RBF) به مدل اجازه می‌دهد مرزهای تصمیم پیچیده‌تری ترسیم کند. اگر داده‌ها ساختار خطی داشته باشند یا ویژگی‌های کافی برای نمایش غیرخطی نداشته باشند، استفاده از RBFممکن است منجر به overfitting یا تخمین ضعیف در داده‌های تست شود.

مدل خطی ساده‌تر و عمومی‌تر است kernel :خطی با پارامترهای کمتر و پیچیدگی پایین‌تر، در بسیاری از مسائل واقعی عملکرد بهتری دارد، مخصوصاً زمانی که تعداد ویژگی‌ها نسبت به تعداد نمونه‌ها زیاد باشد.

در این دیتاست خاص (diabetes)، استفاده از kernel خطی برای الگوریتم SVM نتیجه بهتری نسبت به kernel RBF داده است. دلیل آن می‌تواند ساختار تقریباً خطی داده‌ها یا عدم نیاز به مرز تصمیم پیچیده باشد. بنابراین، در این‌جا مدل ساده‌تر (linear SVM) نه تنها کارایی بهتری دارد، بلکه خطر overfitting کمتری نیز دارد.

3- الگوریتم SVM را با دو مقدار 1 و 100 برای c امتحان کرده و خروجی را تفسیر کنید.

تأثیر C در kernel خطی:

با تغییر مقدار C از 1 به 100 در kernel خطی، عملکرد مدل تقریباً تغییری نکرده است. این نشان می‌دهد که داده‌ها به خوبی با یک مرز خطی قابل جداسازی هستند و تغییر در جریمه‌ی margin که با C کنترل می‌شود (تأثیر معناداری روی مدل نگذاشته است.

تأثیر C در kernel RBF:

در kernel شعاعی، افزایش مقدار C منجر به کاهش دقت (Accuracy) و همچنین کاهش precision و f1-score در کلاس 1 شده است. علت این موضوع آن است که با افزایشC ، مدل نسبت به اشتباهات در آموزش سخت‌گیرتر می‌شود و سعی می‌کند مرز تصمیم را دقیق‌تر از حد لازم رسم کند، که ممکن است باعث overfitting روی داده‌های آموزشی شود و در نتیجه عملکرد در داده‌های تست کاهش یابد.

رفتار پارامتر C به‌صورت کلی:

پارامتر C نقش یک ضریب جریمه برای خطاهای طبقه‌بندی را دارد:

Cکوچک‌تر: جریمه‌ی کمتری برای خطاها، در نتیجه مرز تصمیم صاف‌تر ولی با خطای بیشتر (مدل ساده‌تر).

Cبزرگ‌تر: تلاش برای خطای کمتر، ولی مرز تصمیم پیچیده‌تر و خطر overfitting بیشتر.

در این دیتاست (دیابت)، مدل با kernel خطی نسبت به RBF بهتر عمل می‌کند و تغییر مقدار C از 1 به 100 در kernel خطی تقریباً بی‌تأثیر است. اما درkernel RBF ، افزایش C باعث افت عملکرد مدل می‌شود. بنابراین، برای این مسئله: استفاده از kernel خطی با C=1 مناسب‌ترین انتخاب است.

افزایش C در kernel غیرخطی باعث کاهش توان تعمیم مدل شده است.

4- در نمودار توزیع کلاس‌ها (در زیربخش balance)، تعیین کنید آیا توزیع داده‌ها نامتقارن است؟ اگر داده‌ها نامتقارن باشند، چه راهکاری برای رفع این مشکل پیشنهاد میکنید؟

بله، توزیع داده‌ها نامتقارن (imbalanced) است. کلاس 0 دارای 500 نمونه و کلاس 1 دارای 268 نمونه است. نسبت بین دو کلاس تقریباً 2 به 1 است که نشان می‌دهد داده‌ها به نفع کلاس 0 متمایل‌اند. این عدم توازن می‌تواند باعث شود که

مدل یادگیری به سمت پیش‌بینی کلاس غالب (کلاس 0) متمایل شود و عملکرد ضعیفی در تشخیص کلاس اقلیت (کلاس 1) داشته باشد، به خصوص اگر معیار ارزیابی فقط دقت (accuracy) باشد.

راهکارهای پیشنهادی برای رفع عدم توازن داده‌ها:

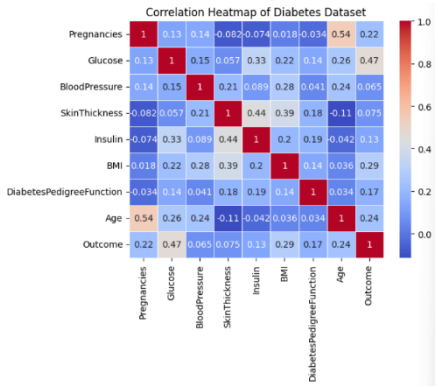
- Oversamplingکلاس اقلیت با استفاده از تکنیک‌هایی مثل:
  SMOTE (Synthetic Minority Over-sampling Technique) که داده‌های مصنوعی برای کلاس اقلیت ایجاد می‌کند.
  Random Oversamplingکه داده‌های کلاس اقلیت را به صورت تصادفی تکرار می‌کند.
- Undersamplingکلاس غالب: کاهش داده‌های کلاس 0 برای متوازن‌سازی تعداد نمونه‌ها، البته با خطر از دست رفتن اطلاعات.
- استفاده از الگوریتم‌های مقاوم در برابر عدم توازن:
  مثل درخت تصمیم، رندوم فارست یا XGBoost با پارامترهایی مثل 'class_weight='balanced یا تنظیم وزن دستی برای کلاس‌ها.
- استفاده از معیارهای ارزیابی مناسب:
  به جای دقت(accuracy) ، از معیارهایی مثلF1-score ، AUC-ROC، precisionو recall استفاده شود، مخصوصاً برای کلاس اقلیت.
- ترکیب چند روش:(ensemble sampling)
  گاهی ترکیب oversampling و undersampling (مثلاً SMOTE + Tomek Links) می‌تواند بهتر عمل کند.

5- با توجه به نمودار انتهایی در زیربخش Comparison و Confusion matrixهای رسم شده برای هر مدل، دقت،صحت،درستی و f1-score مدل‌ها را به صورت خلاصه با یکدیگر مقایسه و نتیجه‌گیری کنید.

مدل	*Accuracy*	*Precision (class 1)*	*Recall (class 1)*	*F1-score (class 1)*
Decision Tree	0.7727	0.73	0.58	0.65
Naive Bayes	0.7662	0.66	0.71	0.68
SVM	0.7597	0.67	0.65	0.66
LDA	0.7597	0.66	0.67	0.67
Random Forest	0.7208	0.61	0.62	0.61
Logistic Regression	0.7532	0.65	0.67	0.66
Bagging (Ensemble)	0.7013	0.58	0.58	0.58

مدل Naive Bayes با بالاترین F1-score و recall برای کلاس 1 بهترین عملکرد را در تشخیص موارد مثبت دارد.
مدل‌های LDA و Logistic Regression نیز عملکرد متعادل و قابل قبولی ارائه می‌دهند.
Bagging و Random Forest در تشخیص کلاس 1 ضعیف‌تر ظاهر شده‌اند و برای کاربردهای حساس به کلاس اقلیت توصیه نمی‌شوند.

6- با کمک رسم ماتریس همبستگی، پنج ویژگی بهتر داده‌ها را مشخص کنید. دلیل انتخاب‌های خود را توضیح دهید.



به ترتیب Gloucose(0.47), BMI(0.29), Age(0.24), Pregnancies(0.22) & DiabetesP.Function (0.17) فیچرهایی با بیشترین همبستگی هستند.

ویژگی	ضریب همبستگی با Outcome	توضیح انتخاب
Glucose	0.47	Outcome بیشترین همبستگی مثبت با را دارد. این یعنی افزایش قند خون رابطه مستقیم با دیابت دارد.
BMI	0.29	وزن/قد بالا به‌وضوح با احتمال دیابت مرتبط است.
Age	0.24	با افزایش سن، احتمال دیابت بیشتر می‌شود. همبستگی نسبتاً قابل‌توجهی دارد.
Pregnancies	0.22	تعداد بارداری‌ها در زنان با افزایش ریسک دیابت ارتباط دارد.
DiabetesPedigreeFunction	0.17	نشان‌دهنده زمینه ژنتیکی دیابت است؛ همبستگی متوسط ولی مهم.

7- حال تمامی مدل‌ها را این بار تنها با استفاده از پنج ویژگی انتخاب شده خود آموزش دهید.

```
[128] X = df.drop(['Outcome','BloodPressure', 'SkinThickness', 'Insulin'], axis=1)
      Y = df['Outcome']
```

```
Decision Tree Accuracy: 0.7727272727272727
Naive Bayes Accuracy: 0.7662337662337663
SVM Accuracy: 0.7597402597402597
LDA Accuracy: 0.7597402597402597
Random Forest Accuracy: 0.7207792207792207
Logistic Regression Accuracy: 0.7532467532467533
Bagging (Ensemble) Accuracy: 0.7012987012987013
```

8- نتیجه بخش 7 را با حالت قبل مقایسه و تحلیل کنید.

```
Decision Tree Accuracy: 0.7727272727272727
Naive Bayes Accuracy: 0.7532467532467533
SVM Accuracy: 0.7662337662337663
LDA Accuracy: 0.7532467532467533
Random Forest Accuracy: 0.7857142857142857
Logistic Regression Accuracy: 0.7532467532467533
Bagging (Ensemble) Accuracy: 0.7467532467532467
```