به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی
(۸۳۰۴۳۶۸)
نیمسال اول ۱۴۰۴-۱۴۰۵

تمرین شماره ۲

تاریخ تحویل:
۱۴۰۴/۰۸/۱۶

دانشکده سامانه‌های هوشمند

۱. (۱۰ ٪) [نظری - بیز ساده و توزیع احتمالاتی]

هدف از این بخش، مطالعه مبحث‌های بیز ساده و توزیع احتمالاتی است.

۱. دسته‌بندی برای تفکیک ایمیل‌های غیر تبلیغاتی از تبلیغاتی طراحی کرده‌ایم، این دسته‌بند در ۹۰ درصد مواقع پاسخ درست می‌دهد. در صورتی که فرض کنیم تنها یک درصد ایمیل‌ها تبلیغاتی هستند، اگر ایمیل تبلیغاتی تشخیص داده شود، با چه احتمالی تبلیغاتی است؟

۲. یک پیشامد موفقیت یا شکست با احتمال موفقیت P مرتبا تکرار می‌شود. پیشامدها در صورت مشاهده دو موفقیت یا دو شکست پیاپی به پایان می‌رسند. احتمال دو موفقیت پیاپی به شرط آن‌که دو شکست پیاپی اتفاق نیفتد چیست؟

۳. در یک آزمایش پزشکی سه پارامتر x، y و z اندازه‌گیری می‌شود. اگر شخصی سالم باشد، توزیع احتمالی این سه پارامتر به صورت گوسی با واریانس‌های به ترتیب $1, 4, 9$ و میانگین‌های به ترتیب برابر $1, 4, 6$ است. برای شخص بیمار نیز توزیع احتمالی این سه پارامتر به صورت گوسی با واریانس‌های به ترتیب $1, 4, 9$ و میانگین‌هایی به ترتیب برابر با $6, 8, 10$ است. با فرض اینکه قصد داریم از روش بیز ساده (روشی که در آن فرض می‌شود سه پارامتر x، y و z مستقل از یکدیگر هستند) برای تفسیر نتایج این آزمایش استفاده کنیم، نشان دهید این روش معادل این است که حاصل عبارت

$$\alpha x + \beta y + \gamma z + \lambda$$

را به دست آورده و اگر این حاصل مثبت بود، فرد بیمار و در غیر این صورت فرد سالم در نظر گرفته می‌شود. با فرض اینکه احتمال بیمار بودن فردی که به طور تصادفی از جامعه انتخاب می‌گردد برابر $\frac{1}{2}$ است، مقادیر α، β، γ و λ را محاسبه کنید.

---

۲. (۱۰ ٪) [نظری-محاسباتی درخت تصمیم و انتخاب ویژگی]

هدف: ارزیابی عملیِ معیارهای Information Gain و Gain Ratio در ساختِ درختِ تصمیم و درکِ مفاهیمِ بایاس و بیش‌برازش.

سناریو: در یک کتابخانه قصد داریم با استفاده از یک مدل درختِ تصمیم پیش‌بینی کنیم که آیا یک کتابِ امانت‌داده‌شده «با تأخیر بازگردانده می‌شود» یا «به‌موقع بازگردانده می‌شود». برچسبِ دودوییِ مسئله به‌صورت $Y \in \{\text{دیر} = 1, \text{به‌موقع} = 0\}$ تعریف می‌شود.

ویژگی‌ها:

- موضوع: {ادبیات، علوم، کودک}
- روشِ دریافت: {حضوری، پستی}
- تمدید: {بله، خیر}
- شماره امانت: شناسه‌ای یکتا برای هر امانت (تعداد مقادیرِ متمایز بسیار زیاد).

داده آموزش:

به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیمسال اول ۱۴۰۴–۱۴۰۵

تمرین شماره ۲

تاریخ تحویل: ۱۴۰۴/۰۸/۱۶

دانشکده سامانه‌های هوشمند

ردیف	موضوع	روشِ دریافت	تمدید	شماره امانت	y (برچسب)
۱	علوم	حضوری	خیر	۲۰۰۱	۱ (دیر)
۲	علوم	حضوری	خیر	۲۰۰۲	۱ (دیر)
۳	علوم	پستی	خیر	۲۰۰۳	۰ (به‌موقع)
۴	علوم	پستی	بله	۲۰۰۴	۱ (دیر)
۵	ادبیات	پستی	خیر	۲۰۰۵	۰ (به‌موقع)
۶	ادبیات	حضوری	بله	۲۰۰۶	۰ (به‌موقع)
۷	ادبیات	پستی	بله	۲۰۰۷	۰ (به‌موقع)
۸	کودک	حضوری	بله	۲۰۰۸	۱ (دیر)
۹	کودک	پستی	خیر	۲۰۰۹	۰ (به‌موقع)
۱۰	کودک	حضوری	خیر	۲۰۱۰	۰ (به‌موقع)

نکته طراحی: شکافتن روی ویژگی «موضوع» در گره ریشه، برخی شاخه‌ها را ناخالص می‌گذارد. با شکافتن‌های بعدی بر اساس «روشِ دریافت» و سپس «تمدید»، برگ‌های خالص به‌دست می‌آیند؛ بنابراین عمق درخت در بعضی مسیرها ۲ یا ۳ خواهد شد.

خواسته‌ها:

(الف) آنتروپی کل مجموعه‌داده $H(Y)$ را با لگاریتمِ پایه ۲ محاسبه کنید. سپس «کسب اطلاعات» (IG) برای گره ریشه را در دو حالت زیر به‌دست آورید: ۱) شکافِ چندشاخه بر اساس «موضوع». ۲) شکافِ چندشاخه بر اساس «شماره امانت» (هر مقدار یک شاخه مجزا).

$$H(Y) = -\sum_{\text{کلاس} \in \{0,1\}} p(\text{کلاس})\, \log_2 p(\text{کلاس})$$

$$\mathrm{IG}(X) = H(Y) - \sum_{j} \frac{|S_j|}{|S|}\, H(Y \mid X{=}j)$$

(ب) با تکیه بر نتایجِ (الف) توضیح دهید چرا IGنسبت به ویژگی‌هایی با تعدادِ مقادیرِ متمایزِ زیاد (مثل «شماره امانت») بایاس دارد. یک استدلالِ کوتاه و کمی بر اساس فرمول ارائه کنید.

(ج) «نسبتِ کسب اطلاعات» (Gain Ratio) را برای دو ویژگیِ «موضوع» و «شماره امانت» محاسبه و مقایسه کنید. بر اساس این معیار، کدام ویژگی برای گره ریشه انتخاب می‌شود؟ نتیجه را تفسیر کنید.

$$\mathrm{SplitInfo}(X) = -\sum_{j} \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|}$$

$$\mathrm{GR}(X) = \frac{\mathrm{IG}(X)}{\mathrm{SplitInfo}(X)}$$

(د) با فرضِ انتخابِ ویژگیِ ریشه طبق (ج)، درختِ تصمیم را تا رسیدن به برگ‌های خالص یا حداکثر عمقِ ۳ رشد دهید. در هر گرهٔ داخلی، از IG برای انتخاب بهترین شکاف میانِ «روشِ دریافت» و «تمدید» استفاده کنید. اگر IG برابر شد، ویژگی با تعدادِ مقادیرِ کمتر را برگزینید. ساختار نهاییِ درخت را گزارش کنید: ویژگی و قاعده هر گره و کلاسِ هر برگ.

(هـ) دقتِ آموزش (Training Accuracy) درخت نهایی را گزارش کنید. به‌اختصار توضیح دهید اگر در گره ریشه به‌جای «نسبتِ کسب اطلاعات» صرفاً از IG استفاده می‌کردید، خطرِ بیش‌برازش چگونه تغییر می‌کرد و چرا.

۳. (۵ ٪) [مفهومی-محاسباتی کاربرد رگرسیون لجستیک]

هدف: ارزیابی درک نحوه استفاده از یک مدل رگرسیون لجستیک آموزش‌دیده برای پیش‌بینی، یافتن مرز تصمیم و تفسیر پارامترها.

سناریو: یک فروشگاه آنلاین مدلی ساخته است تا پیش‌بینی کند آیا یک بازدیدکننده از سایت خریدی انجام می‌دهد یا خیر. ویژگی (x): تعداد دقایقی که بازدیدکننده در سایت صرف کرده است. برچسب (y): نتیجه بازدید ($1 =$) خرید انجام شد، ($0 =$) خریدی انجام نشد.)

مدل آموزش‌دیده: پس از آموزش مدل رگرسیون لجستیک، پارامترهای بهینه به صورت زیر به دست آمده‌اند: وزن: $w = 0.5$ بایاس: $b = -2$ بنابراین، فرمول پیش‌بینی احتمال خرید:

$$\hat{p} = \sigma(0.5\,x - 2), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

خواسته‌ها:

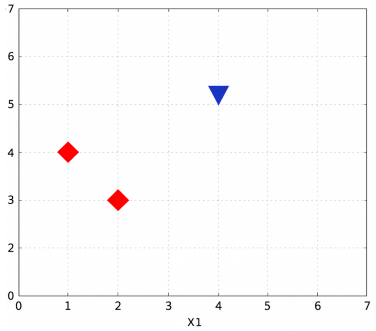
(الف) پیش‌بینی احتمال: یک بازدیدکننده جدید ۶ دقیقه در سایت وقت می‌گذراند ($x = 6$). با استفاده از مدل بالا، احتمال این‌که این بازدیدکننده خریدی انجام دهد چقدر است؟ (می‌توانید پاسخ را بر حسب e باقی بگذارید یا تقریب بزنید؛ مثلاً $e^{-1} \approx 0.37$.)

(ب) یافتن مرز تصمیم: مرز تصمیم نقطه‌ای است که در آن مدل احتمال خرید را دقیقا ۵۰٪ پیش‌بینی می‌کند (شانس خرید و عدم خرید برابر است). این چه وقتی رخ می‌دهد . بازدیدکننده باید چند دقیقه در سایت وقت بگذراند تا به این نقطه برسد؟

(ج) تفسیر وزن مدل: چون وزن ویژگی «زمان صرف‌شده در سایت» مثبت است ($w = 0.5$)، این موضوع چه چیزی درباره رابطه بین زمان ماندن در سایت و احتمال خرید می‌گوید؟ (پاسخ یک جمله کافی است.)

۴. (۱۰ ٪) [نظری – SVM]

۴-۱ می خواهیم یک طبقه بند ماشین بردار پشتیبان (SVM) را روی داده های زیر آموزش دهیم. در این شکل، داده با مقدار -1 (نمونه های قرمز) و داده با مقدار $+1$ (نمونه های آبی) نشان داده شده است.



شکل ۱: نمودار داده های آموزشی SVM

(الف) معادله خط تصمیم را بدست آورید و مقادیر m، w و b را محاسبه کنید.

(ب) نقاط بردار پشتیبان را روی تصویر مشخص کرده و خط تصمیم را رسم کنید.

۲-۴ ماشین های بردار پشتیبان را می توان با ترفند کرنل برای طبقه بندی غیرخطی به کار برد. SVM با حاشیهٔ سخت از کلاس را به خاطر بیاورید:

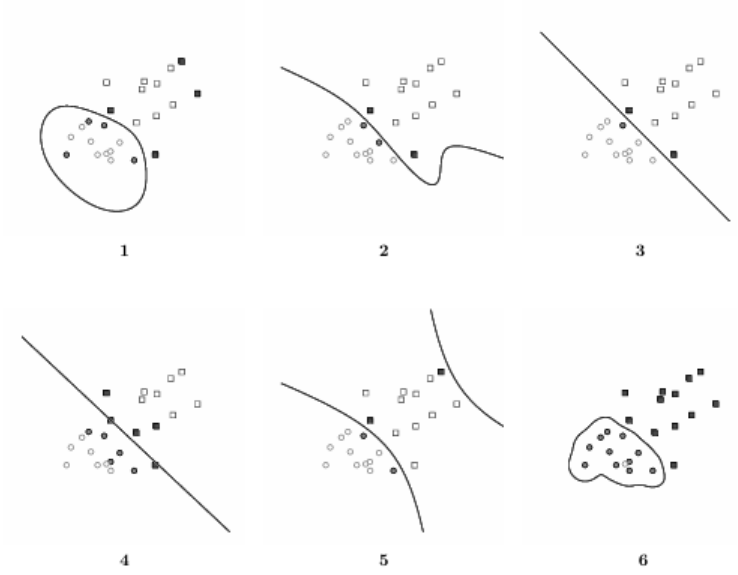
$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top\mathbf{w} \quad \text{s.t.} \quad y^{(i)}\left(\mathbf{w}^\top\mathbf{x}^{(i)}+b\right) \geq 1.$$

دوگانهٔ این مسئله را می توان به صورت زیر برای یادگیری یک رده بند خطی نوشت:

$$f(\mathbf{x}) = \sum_{i=1}^{N} y_i \left(\mathbf{x}_i^\top \mathbf{x}\right) + b.$$

اکنون می توان ضرب داخلی $\mathbf{x}_i^\top\mathbf{x}$ را با یک کرنل $k(\mathbf{x}_i,\mathbf{x})$ جایگزین کرد تا مرز تصمیم غیرخطی به دست آید.

در شکل زیر چند SVM با شکل ها و الگوهای متفاوت مرز تصمیم نشان داده شده اند. داده های آموزشی با برچسب های $y_i \in \{+1,-1\}$ داده شده اند و به ترتیب با دایره ها و مربع ها نمایش یافته اند. بردارهای پشتیبان با دایره های پُر مشخص شده اند.



شکل ۲: نمونه هایی از مرز تصمیم در SVM های خطی و کرنلی

سناریوهای زیر را با یکی از شش نمودار تطبیق دهید (توجه کنید که یکی از نمودارها با هیچ سناریویی تطابق ندارد). هر سناریو باید به یک نمودار یکتا نسبت داده شود و برای هر سناریو در کمتر از دو جمله توضیح دهید چرا چنین است.

۱. یک SVM خطی با حاشیه نرم با $C = 0.02$

۲. یک SVM خطی با حاشیه نرم با $C = 20$

۳. یک SVM کرنلی با حاشیه سخت با $k(u,v) = u\cdot v + (u\cdot v)^2$

۴. یک SVM کرنلی با حاشیه سخت با $k(u,v) = \exp(-5\|u-v\|^2)$

۵. یک SVM کرنلی با حاشیه سخت با $k(u,v) = \exp\left(-\frac{1}{5}\|u-v\|^2\right)$

---

۵. (۴۰ ٪) [عملی-پیاده سازی] [تحلیل احساس، استخراج ویژگی متنی، Naive Bayes، Logistic Regression، KNN]

هدف: پیاده سازی و مقایسه سه مدل کلاسیک یادگیری ماشین (Naive Bayes, Logistic Regression, KNN) برای مسئله تحلیل احساس روی داده‌های واقعیِ نظرات فارسی در حوزه سفارش آنلاین غذا؛ شامل:

- طراحی و پیاده سازی توکنایزرهای قاعده مبنا و لمه محور برای زبان فارسی،
- ساخت نمایش‌های برداری از متن (Unigram & Bigram Bag-of-Words)،
- پیاده سازی سه مدل طبقه بندی از صفر (بدون استفاده از کتابخانه‌های آماده)،
- ارزیابی و مقایسه بر اساس معیارهای استاندارد.

سناریو و داده‌ها:

- داده‌ها: مجموعه‌ای از نظرات کاربران برای پلتفرم‌های سفارش آنلاین غذا با برچسب احساسی.
- فایل‌ها: داده‌های آموزش و آزمون به صورت تفکیک شده ارائه شده‌اند.
- ستون‌ها: comment (متن نظر) و label_id (برچسب دودویی: ۰ = منفی، ۱ = مثبت).

مراحل تمرین

مرحله ۱: پردازش متن و استخراج ویژگی

۱. طراحی توکنایزر (یکی را انتخاب و پیاده سازی کنید):
   - مسیر A: توکنایزر قاعده مبنا (Regex-based) با الگوی جداسازی بر اساس حروف فارسی/انگلیسی و اعداد.
   - مسیر B: توکنایزر لمه محور (Lemma-based) با ابزارهای فارسی مانند Hazm، Parsivar یا Stanza.
۲. واژه نامه را فقط از روی داده آموزش بسازید؛ کلماتی از آزمون که در واژه نامه نیستند نادیده گرفته شوند.
۳. نمایش‌های ویژگی را برای هر دو مجموعه آموزش و آزمون بسازید:
   - Unigram BOW: بردار حضور/عدم حضور واژه‌های منفرد.
   - Bigram BOW: بردار حضور/عدم حضور دوواژه‌ای‌های متوالی.

۴. نکته محاسباتی: برای مدیریت زمان، می‌توانید مسیر B را روی یک زیرمجموعه تصادفی (حداقل ٪۵۰) از داده آموزش اجرا کنید و اثر آن را در گزارش تحلیل کنید.

مرحله ۲: پیاده سازی Naive Bayes (از صفر)

- مدل Bernoulli Naive Bayes با هموارسازی لاپلاس.
- پیشین کلاس‌ها:

$$\hat{P}(y) = \frac{\text{تعداد اسناد کلاس } y}{\text{تعداد کل اسناد}}$$

- احتمال شرطی ویژگی‌ها (برای واژه w در کلاس y):

$$\hat{P}(w \mid y) = \frac{c(w,y) + \alpha}{n_y + 2\alpha}$$

که در آن $c(w,y)$ تعداد اسنادِ کلاس y است که w در آن‌ها حضور دارد و n_y تعداد کل اسنادِ کلاس y است.

- امتیاز پسین و قاعده تصمیم:

$$s_y(x) = \log P(y) + \sum_{w \in x} \log P(w \mid y), \quad \hat{y} = \arg\max_y s_y(x)$$

خواسته‌های این بخش:

۱. اندازه واژه نامه ($|V|$) را برای Unigram و Bigram گزارش کنید.

۲. برای هر کلاس (مثبت/منفی)، ۱۰ ویژگی با بزرگ‌ترین $P(w \mid y)$ را فهرست و کوتاه تفسیر کنید.

۳. حساسیت به α: مدل را برای $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ اجرا و نمودار دقت آزمون بر حسب α را رسم کنید.

۴. یک نمونه آزمون که به اشتباه طبقه بندی شده است را تحلیل کنید.

مرحله ۳: پیاده سازی Logistic Regression (از صفر)

- مدل احتمال: $\hat{p}(x) = \sigma(\theta^\top x)$، تابع $\sigma(z) = \frac{1}{1+e^{-z}}$.
- هزینه Cross-Entropy:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big]$$

- گرادیان: $\nabla_\theta J = \frac{1}{m} X^\top (\hat{p} - y)$.
- آموزش با نرخ یادگیری ثابت و تعداد مشخص Epoch؛ آستانه تصمیم 0.5.

مرحله ۴: پیاده سازی K-Nearest Neighbors (KNN) (از صفر)

- معیار شباهت: شباهت کسینوسی (Cosine Similarity) روی بردارهای باینری.

- انتخاب k: عملکرد برای $k \in \{1, 3, 5, 7, 9\}$ و انتخاب بهترین با استدلال.
- قانون تصمیم: رأی اکثریت بین k همسایه نزدیک.

مرحله ۵: ارزیابی، مقایسه و تحلیل نتایج

- معیارها: Accuracy, Precision, Recall, F1-Score, ROC-AUC.
- جدول مقایسه: نتایج هر ترکیب «مدل × ویژگی» (۶ حالت) را در یک جدول یکجا گزارش کنید.
- تحلیل: کدام مدل و کدام نمایش ویژگی بهترین بود؟ بر اساس نمودار حساسیت Naive Bayes، بهترین α چیست؟ در KNN چرا k منتخب مناسب است؟ اگر مسیر B را روی داده کمتر اجرا کردید، اثر آن بر عملکرد را توضیح دهید.

محدودیت ها:

- استفاده از پیاده سازی های آماده مدل ها (مثل `scikit-learn.naive_bayes`) مجاز نیست.

---

۶. (۱۰٪) [پیاده سازی: برچسب زنی اجزای کلام با درخت تصمیم]

هدف این تمرین پیاده سازی برچسب زن اجزای کلام برای زبان فارسی با استفاده از روش درخت تصمیم است. در این تمرین از داده های ارائه شده به همراه تمرین استفاده کنید. داده های داده شده شامل یک فایل آموزش با اسم `POST-Persian-Corpus-Train.txt` (حدود ۳۰۰۰ جمله) و یک فایل آزمون با اسم `POST-Persian-Corpus-Test.txt` (حدود ۵۵۰ جمله) است که تمام کلمات آن برچسب گذاری دستی (با تعداد ۲۱ برچسب) شده اند. برای ساخت درخت از داده فایل آموزش و برای ارزیابی کار خود و تست کردن کیفیت سیستم ساخته شده، از داده فایل آزمون استفاده شود. ویژگی های مورد استفاده برای حل این سوال را بررسی کرده و در پاسخ خود بیان کنید که از چه ویژگی هایی استفاده کرده اید. برای فاز آزمون، از معیار F1 استفاده کنید و همچنین ماتریس Confusion را رسم کنید.

---

۷. (۱۵ ٪) [عملی–پیاده سازی: طبقه بندی با SVM برای پیش بینی جنسیت گربه ها]

در این سؤال می خواهیم با استفاده از مدل طبقه بندی بردار پشتیبان، جنسیت گربه ها را با داشتن اطلاعات زیستی آن ها طبقه بندی کنیم. اطلاعات این مجموعه داده به صورت زیر است:

Gender: جنسیت Bwt: وزن بدن Hwt: وزن قلب

چند سطر نمونه از داده:

Gender	Bwt	Hwt
F	2	7
F	2	7.4
M	2.8	10.2
M	2.9	11.3

(الف) داده ها را از فایل cats.csv بخوانید؛ تمام مقادیر را عددی کنید، سپس داده را به بخش های آموزش و تست با نسبت 20%/80% تقسیم نمایید.

(ب) داده ها را با استفاده از روش StandardScaler نرمال کنید و آن را بر روی یک صفحه دوبعدی، یک بار قبل از نرمال سازی و یک بار بعد از نرمال سازی نمایش دهید. محور افقی مقادیر Hwt و محور عمودی مقادیر Bwt باشد. مشاهدات خود را در مورد تغییر مقیاس و شکل توزیع داده ها گزارش کنید.

(ج) با استفاده از تابع np.logspace(start, stop, num) آرایه های زیر را ایجاد کنید:

- آرایه C شامل ۸ عدد در بازه $[-2, 5]$
- آرایه Gamma شامل ۱۶ عدد در بازه $[-6, 1]$

(د) با استفاده از تابع SVC در کتابخانه scikit-learn یک طبقه بند بردار پشتیبان خطی طراحی کنید. به کمک روش Grid Search، بهترین مقادیر C و Gamma را به دست آورید و مقادیر بهینه را همراه با دقت مدل گزارش کنید.

(هـ) با استفاده از مدل SVC در scikit-learn، سه مدل مختلف SVM را روی داده های آموزش، با کرنل های زیر آموزش دهید:

- تابع پایه شعاعی (RBF)
- سیگموید (Sigmoid)
- چندجمله ای (Polynomial)

برای هر کرنل:

۱. بهترین پارامترها (C،Gamma و در صورت نیاز درجه چندجمله ای) را با استفاده از Grid Search روی داده آموزش بیابید.

۲. دقت نهایی مدل را روی داده آموزش گزارش کنید.

در پایان، نتایج هر چهار مدل (شامل مدل خطی و سه مدل غیرخطی) را در یک جدول مقایسه ای ارائه دهید، به طوری که دقت نهایی و پارامترهای بهینه هر مدل در ستون های جداگانه آورده شود.

(و) برای هر یک از چهار مدل به دست آمده در بخش (هـ):

۱. مرز تصمیم (Decision Boundary) و ناحیه جداسازی کلاس ها را بر روی داده های آموزش رسم کنید. نقاط داده های آموزش باید بر روی همان نمودار نمایش داده شوند و هر کلاس با رنگ متفاوت مشخص شود.

۲. همان مدل ها را بر روی داده های تست اعمال کنید و دقت (Accuracy) هر مدل را گزارش نمایید.

۳. در پایان، با مقایسه نمودارها و نتایج به دست آمده، توضیح دهید که تفاوت عملکرد چهار کرنل (linear, rbf, sigmoid, polynomial) در چیست. سپس بر اساس مشاهدات خود بیان کنید که کدام کرنل در این داده ها عملکرد بهتری دارد و چرا (از نظر شکل مرز تصمیم و دقت عددی).