بسمه تعالی

تمرین چهارم **یادگیری آماری** نیمسال اول **۰۳**-۰۲

۱ Natural Cubic Splines

نمایش زیر را برای cubic splines با K نقطه درونی (knots) در نظر بگیرید:

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

۱٫۱ نشان دهید که شرایط خطی بودن تخمین در بازه‌های ابتدایی و انتهایی منجر به نتایج زیر می‌شود:

$$\beta_2 = \beta_3 = 0$$

$$\sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \xi_k \theta_k = 0$$

۱٫۲ (امتیازی) نشان دهید در صورتی که از متغیرهای زیر برای تخمین استفاده کنیم، شرایط ذکر شده حتما برقرار خواهند بود:

$$N_1(X) = 1$$

$$N_2(X) = X$$

$$N_{j+2}(X) = d_j(X) - d_{K-1}(X)$$

$$d_j(X) = \frac{\left(X - \xi_j\right)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_j}. \quad \forall j \in \{1.2 \ldots . K-1\}$$

۲ Weighted Natural Cubic Splines

تابع هزینه زیر را در نظر بگیرید:

$$RSS(f,\lambda) = \sum_{i=1}^{n} w_i \left(y^{(i)} - f\left(x^{(i)}\right)\right)^2 + \lambda \int \{f''(t)\}^2 \, dt$$

ما به‌دنبال پیدا کردن تابع $\hat{f}$ هستیم به طوری‌که تابع هزینه بالا را کمینه کند. فرض کنید می‌دانیم که این تابع به صورت ترکیبی خطی از Natural Cubic Splines می‌باشد؛ یعنی:

$$\hat{f}\left(x^{(i)}\right) = \sum_{j=1}^{n} N_j\left(x^{(i)}\right)\hat{\theta}_j$$

با فرض $w_i \geq 0$ فرم بسته‌ای برای f پیدا کنید. (مشخصا کافی است که فقط بردار $\widehat{\boldsymbol{\theta}}$ را بدست آورید.)

۳ Soft Margin SVM با نرم l_2

در این مسئله با تغییر در تابع هزینه مورد بهینه‌سازی سعی می‌کنیم الگوریتم بهینه‌سازی خود در مسئله SVM را تغییر دهیم؛ به طور خاص با تغییر نرم l_1 به نرم l_2 برای slack variables تفاوت‌های این دو مسئله را بررسی می‌نماییم.

این الگوریتم جدید با مسئله بهینه‌سازی زیر داده می‌شود: (به مربع ξ ها دقت نمایید)

$$\begin{array}{rl} \min_{w,b,\xi} & \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 \\ \text{s.t.} & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i,\ \ i = 1,\ldots,m \end{array}$$

با توجه به مسئله بالا، به سوالات زیر پاسخ دهید.

۳,۱ نشان دهید چرا شرط $\xi_i \geq 0$ از شروط مسئله حذف شده است و در واقع این حذف تاثیری بر مقدار بهینه مسئله بالا نخواهد داشت.

۳,۲ تابع لاگرانژین $\mathcal{L}\ (w, b, \xi, \alpha)$ مسئله بهینه‌سازی بالا را بدست آورید.

۳,۳ مشتقات نسبی تابع لاگرانژین را نسبت به w, b, α بدست آورید و با برابر قرار دادن آن‌ها با صفر، مقدار بهینه این متغیرها را بدست آورید.

۳,۴ با استفاده از نتایج بخش قبل، دوگان مسئله بهینه‌سازی معرفی‌شده در این سوال را بدست آورید وآن را با مسئله دوگان حالت مطرح شده در کلاس مقایسه نمایید.

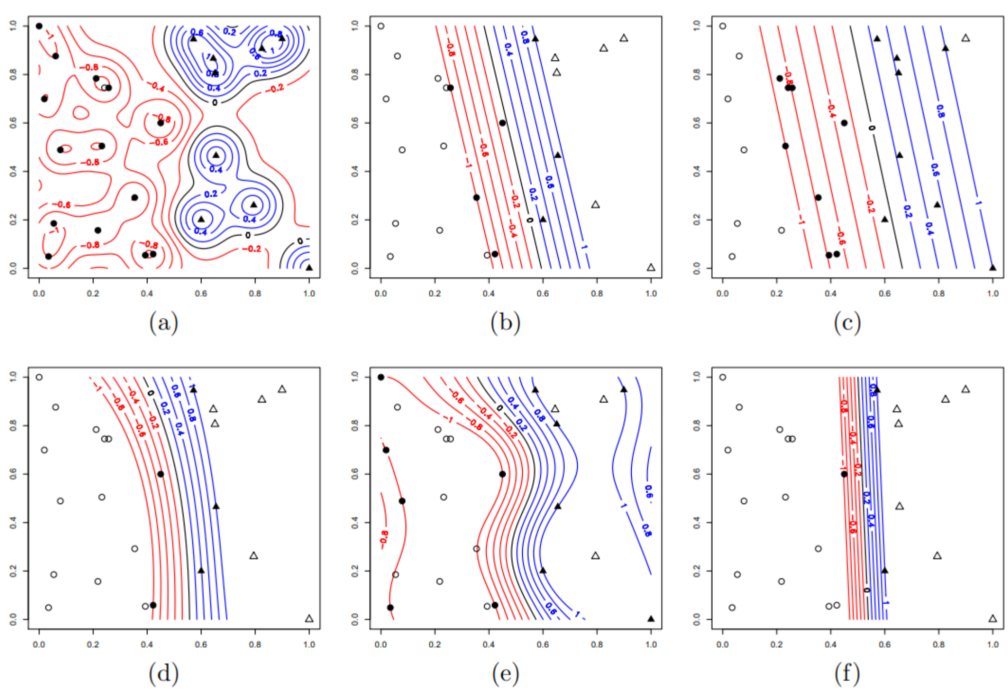
۴ SVM Hyperparameters

در شکل ۱ مرز تصمیم‌گیری و حاشیه متناظر برای چند مدل SVM که با ابرپارامترهای زیر بر روی یک دیتاست یکسان آموزش داده شده‌اند ترسیم شده است:

- الف) کرنل خطی - $C = 0.1$
- ب) کرنل خطی - $C = 1$
- پ) کرنل خطی - $C = 10$
- ت) کرنل RBF - $\gamma = 0.1$ و $C = 15$
- ث) کرنل RBF - $\gamma = 1$ و $C = 3$
- ج) کرنل RBF - $\gamma = 10$ و $C = 1$

$$\text{RBF kernel: } k(x, y) = \exp\left(-\gamma ||x - y||^2\right)$$

با ذکر توضیحی مختصر ابرپارامترهای متناظر با هر یک از نمودارهای زیر را مشخص نمایید.



شکل ۱: دایره‌ها و مثلث‌ها به ترتیب کلاس ۱ و ۲ را نشان می‌دهند و نمونه‌های support با نقاط توپر نمایش داده شده‌اند.

۵ کرنل معتبر

۵٫۱ ماتریس متقارن مثبت نیمه‌معین $A \in \mathbb{R}^{p\times p}$ و بردار ویژگی‌های $x \in \mathbb{R}^{p\times 1}$ در نظر بگیرید. نشان دهید تابع $k(x,x')$ یک تابع کرنل معتبر است:

$$k(x,x') = x^T A x$$

۵٫۲ دو تابع $k_1(x,y)$ و $k_2(x,y) \in \mathbb{R}^{p\times p}$ را کرنل‌های معتبری در نظر بگیرید. آیا هر ترکیب خطی این دو کرنل، یعنی

$$k_3(x,y) = \alpha k_1(x,y) + \beta k_2(x,y)$$

همچنان یک تابع کرنل معتبر است؟ در صورتی که شرطی برای معتبر بودن این تابع وجود دارد، آن را به طور صریح اثبات کنید.

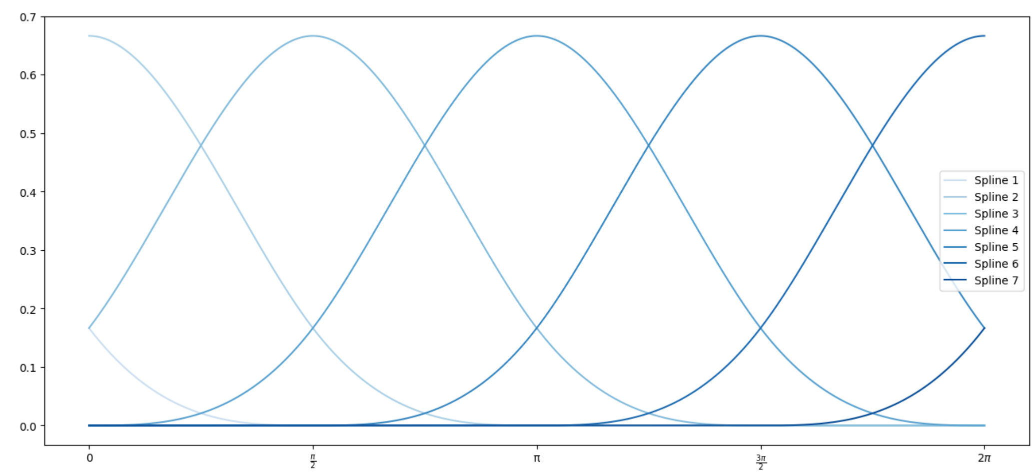
۶ تمرین پایتون

قسمت اول

در این قسمت ابتدا تعدادی داده یک بعدی (یک ویژگی) تولید خواهید کرد و سپس قرار است تا با استفاده از توابع مختلف، ویژگی‌های جدیدی بدست آورید و نهایتا با اعمال روش رگرسیون خطی بر روی ویژگی‌های بدست آمده، منحنی اصلی تابع در دامنه داده را تخمین بزنید.

برای اینکار در بازه $0 \leq x \leq 2\pi$ ابتدا به تابع $\sin(x)$ نویز نرمال اضافه نمایید (برای یکسان بودن نتایج، مقدار ۰٫۲۵ برابر نویز نرمال را به خروجی تابع اضافه نمایید) و ۷۰ نقطه را به عنوان داده آموزش و ۳۰ نقطه را به عنوان داده تست انتخاب نمایید.

۱.در این بخش قرار است تا با استفاده از B_Spline تابع $\sin(x)$ تخمین بزنید. برای اینکار لازم است تا SplineTransformer را از کتابخانه sklearn اضافه کنید. ابتدا با جستجو در اینترنت درباره آن مطالعه کنید. سپس پارامترهای ورودی n_knots وdegree را مطابق پیش فرض قرار دهید و B_Splineها را در بازه $0 \leq x \leq 2\pi$ رسم نمایید (مشابه شکل ۱). همچنین توضیح دهید که چه رابطه‌ای میان تعداد نات‌ها، درجه B_Splineها و تعدادشان وجود دارد.



شکل ۱

۲.در این قسمت با اعمال رگرسیون خطی به ویژگی‌هایی که در قسمت قبل بدست آوردید، ضرایب رگرسیون را بدست آورید و این ضرایب را به همراه MSE داده‌های آموزش و تست گزارش نمایید و در آخر نمودار تابع بدون نویز $\sin(x)$، تابع تخمینی بدست آمده و داده‌ها را در یک نمودار رسم کنید.

۳.در این قسمت با استفاده از PolynomialFeatures ویژگی‌های جدید را بدست آورید (درجه ۳) و این ویژگی‌ها را در بازه $0 \le x \le 2\pi$ ترسیم نمایید و سپس با اعمال رگرسیون خطی به ویژگی‌هایی جدید، ضرایب رگرسیون را بدست آورید و این ضرایب را به همراه MSE داده‌های آموزش و تست گزارش نمایید و در آخر نمودار تابع بدون نویز $\sin(x)$، تابع تخمینی بدست آمده و داده‌ها را در یک نمودار رسم کنید.

۴.در این قسمت قرار است تا ویژگی‌های جدید را با استفاده از توابع گوسی بدست آوریم. برای این کار ۵ تابع گوسی تعریف کنید که واریانسشان برابر ۱ و میانگین آن‌ها به ترتیب برابر *0*، $\pi/2$، π، $3\pi/2$، 2π باشد. سپس با اعمال رگرسیون خطی به ویژگی‌هایی جدید، ضرایب رگرسیون را بدست آورید و این ضرایب را به همراه MSE داده‌های آموزش و تست گزارش نمایید و در آخر نمودار تابع بدون نویز $\sin(x)$، تابع تخمینی بدست آمده و داده‌ها را در یک نمودار رسم کنید. حال این سه روش را با یکدیگر مقایسه نمایید.

قسمت دوم

در این قسمت قرار است تا با استفاده از کتابخانه symfit (هر روش دیگری هم قابل قبول است) به داده‌های تولید شده در قسمت قبل، Spline درجه سوم فیت کنیم. با این تفاوت که ابتدا نات‌ها را مشخص می‌کنیم و به صورت گام به گام شروط پیوستگی را اعمال می‌کنم. در حالت اول فرض می‌کنیم که فقط شرط پیوستگی در نات‌ها برقرار است. در حالت بعدی فرض می‌کنیم که علاوه بر حالت قبل، شرط پیوستگی مشتق اول هم برقرار است. در حالت سوم فرض می‌کنیم که علاوه بر حالات قبل، مشتق دوم هم پیوسته باشد و در حالت چهارم فرض می‌کنیم که قطعه اول و آخر خطی باشند (Natural Cubic Spline).

برای اینکار ابتدا ماژول‌های زیرا را وارد کنید:

```
from symfit import parameters, variables, Fit, Piecewise, Eq, Model
```

در این قسمت فرض کنید که سه نات در $\pi/2$، π و $3\pi/2$ قرار دارد که در نتیجه چهار بازه داریم که باید برای هر بازه یک تابع چند جمله‌ای درجه سوم تعریف کنیم.

۱.ابتدا با استفاده از variables متغیرهای x و y را تعریف کنید و سپس با استفاده از parameters، پارامترهای توابع چندجمله‌ای را تعریف نمایید.

۲.حال با استفاده از پارامترهای بالا، توابع درجه سه را تعریف نمایید و با استفاده از Piecewise تابع تکه‌ای کل را تعریف کنید.

$$f(x)=\begin{cases} a_{10}+a_{11}x+a_{12}x^2+a_{13}x^3, x<\pi/2 \\ a_{20}+a_{21}x+a_{22}x^2+a_{23}x^3, \pi/2\le x<\pi \\ a_{30}+a_{31}x+a_{32}x^2+a_{33}x^3, \pi\le x<3/2\pi \\ a_{40}+a_{41}x+a_{42}x^2+a_{43}x^3, 3\pi/2\le x \end{cases}$$

که در آن a_{ij} همان پارامترهایی است که تعریف کردید.

۳.حال با استفاده از Eq چهار شرط بالا(شروط پیوستگی و خطی بودن قطعه اول و آخر) را تعریف کنید(برای محاسبه مشتق می‌توانید از diff از کتابخانه sympy استفاده کنید) و چهار مدل بدست بیاورید.

۴.حال داده‌ها را به چهار مدل بالا فیت نمایید و برای مدل‌های بالا، توابع بدست آمده در بازه $0\le x\le 2\pi$ رسم کرده و مقدار MSE را برای داده‌های آموزش و تست گزارش نمایید.

قسمت سوم

در این قسمت قرار است تا با استفاده از csaps، Smoothing Spline را پیاده‌سازی کنیم. این الگوریتم سعی دارد تا با جریمه کردن مجموع مربعات با استفاده از انتگرال مشتق دوم (مشابه همان چیزی که در اسلایدهای درس دیدید) تابع تخمینی را تا حد قابل قبولی نرم کند.

$$\mathrm{Cost} = p\sum_{i=1}^{n}\left(y_i-\hat{f}(x_i)\right)^2+(1-p)\int\left|\hat{f}''(t)\right|^2 dt$$

پارامتر p از صفر تا یک تغییر خواهد کرد و اگر صفر باشد، خروجی ما همان رگرسیون خطی خواهد شده و اگر یک باشد آنگاه خروجی‌اش Cubic Spline خواهد شد.

```
y_points = csaps(x_train, y_train, x_points, smooth=p)
```

حال به ازای پارامترهای p مختلف تابع قسمت اول را تخمین بزنید و مقدار MSE را برای داده‌های آموزش و تست گزارش کنید و نهایتا نمودار تابع بدون نویز $\sin(x)$، تابع تخمینی بدست آمده و داده‌ها را در یک نمودار رسم کنید.

قسمت چهارم

در این قسمت می‌خواهیم چندین روش را بر روی دادهای واقعی اعمال کنیم و آن‌ها را باهم مقایسه کنیم. دیتاستی که قرار است تا بررسی نمایید شامل ستون BMD[1]، سن و جنسیت است. می‌خواهیم منحنی BMD بر حسب age را هم برای خانم‌ها و هم برای آقایان تخمین بزنیم. ابتدا داده‌های آقایان و خانم‌ها را از یکدیگر جدا نمایید و داده‌های آموزش و تست را به نسبت ۰,۷ به ۰,۳ برای هر کدام مشخص نمایید.

۱.با استفاده از روش Ridge، به هر کدام از دیتاست‌ها (آقایان و خانم‌ها) چند جمله‌ای با درجه مناسب فیت نمایید (پارامتر Ridge و همچنین درجه چند جمله‌ای را با استفاده از cross-validation مشخص نمایید). در نهایت درجه چند جمله‌ای بدست آمده برای هر کدام را گزارش نمایید و همچنین مقدار MSE و R^2 را برای داده‌های تست و آموزش گزارش کنید و در نهایت منحنی چندجمله‌ای بدست آمده برای آقایان و خانم‌ها را به همراه داده‌های دیتاست، در یک نمودار رسم نمایید.

۲.با استفاده از روش Ridge و SplineTransformer (degree=3, n_knots=6) مدل بدست آمده را به داده‌های آموزش فیت نمایید و مقدار MSE و R^2 را برای داده‌های تست و آموزش گزارش کنید و در نهایت منحنی بدست آمده برای آقایان و خانم‌ها را به همراه داده‌های دیتاست، در یک نمودار رسم نمایید.

۳.در این قسمت قرار است تا با استفاده از Smoothing Spline منحنی مربوط به آقایان و خانم‌ها را تخمین بزنیم. برای این کار از csaps استفاده خواهیم کرد. اما مشکلی که وجود دارد این است که اگر nتا داده داشته باشیم، داده‌های که به csaps داده می‌شود باید به صورت زیرباشند:

$$x_1 < x_2 < \ldots < x_n$$

اما برخی از داده‌های ما age یکسانی دارند. برای اینکه بتوانیم از csaps استفاده نماییم، به ستون age داده‌ها مقدار بسیار کوچکی نویز اضافه نمایید تا بتوانیم از این ماژول استفاده نماییم. حال پس از انجام این کار پارامتر مناسب p را انتخاب نمایید و مقدار این پارامتر برای آقایان و خانم‌ها را گزارش نمایید و همچنین مقدار MSE و

[1] Bone Mineral Density

R^2 را برای داده‌های تست و آموزش گزارش کنید و در نهایت منحنی بدست آمده برای آقایان و خانم‌ها را به همراه داده‌های دیتاست، در یک نمودار رسم نمایید.

حال این سه روش را با یکدیگر مقایسه نمایید و درباره مزایا و معایبشان توضیح دهید.