سلام بر تمام دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
۳. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده از کنید. شکل ها به طور واضح و در فرمت درست گزارش شوند.
۴. از بین سوالات شبیه سازی حتما به هر دو مورد پاسخ داده شود. حداکثر تا نمره ۱۱۰ (۱۰ نمره امتیازی) لحاظ خواهد شد.
۵. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر می‌شود.
۶. در صورت داشتن سوال، از طریق ایمیل banafshehkarimian@ut.ac.ir و a.rokni@ut.ac.ir، سوال خود را مطرح کنید.

سوال ۱)

به سوالات مفهومی زیر پاسخ دهید: (۱۵ نمره)

- اگر تک بعدی باشن داده‌ها کدام روش نان پارامتریک ساده ترین روش است؟
- خوبی و بدی پارزن و kNN به نسبت هم را بیان کنید؟
- به طور خلاصه توضیح دهید که چه زمانی MAP و ML یکسان اند.

سوال ۲) داده های زیر را در نظر بگیرید: (۱۵ نمره)

c1		c2		c3	
x	y	x	y	x	y
10	0	5	10	2	8
0	-10	0	5	-5	2
5	-2	5	5	10	-4

الف) فرض کنید تنها c1 و c2 در دست شما اند. حال برای الگوریتم نزدیک ترین همسایه مرز بین کلاس ها را در یک نمودار مشخص کنید (اطراف هر نقطه از هر کلاس در فضای دو بعدی محدوده ای رسم کنید که اگر نقطه دیگری در آن محدوده قرار گیرد و شما قصد کلاس بندی آن را داشته باشید به کلاس آن نقطه اختصاص داده شود)

ب) بخش الف را با پیدا کردن میانگین هر کلاس در نظر بگیرید که هر نقطه به کلاسی با نزدیک ترین میانگین اختصاص داده میشود. محدوده تصمیم را برای هر کلاس مشخص کنید (هر نقطه در هر محدوده به آن کلاس تعلق میگیرد)

ج) بخش الف و ب را برای حالتی که هر سه کلاس را داریم رسم کنید.

سوال ۳) نشان دهید برای حالت یک بعدی که: (۲۰ نمره)

$$P(wi) = 1/c$$

و

$$P(x|wi) = \begin{cases} 1 & 0 \le x \le \dfrac{cr}{c-1} \\ 1 & i \le x \le i+1-\dfrac{cr}{c-1} \\ 0 & o, w, \end{cases}$$

Bayes error rate برابر P*=r و nearest-neighbor error rate برابر آن یعنی P*=P است.

سوال ۴) مجموعه داده های (1,1)،(3,3) و (*,2) که * به معنای یک مقدار ویژگی نامعلوم است از یک توزیع دو بعدی جدایی پذیر با توزیع P(x1,x2)=P(x1)*P(x2) به دست آمده اند به طوری که: (۲۰ نمره)

$$p(x1) = \begin{cases} \frac{1}{\theta 1} e^{-\theta 1 x1} if\ x1 > 0 \\ 0\ otherwise \end{cases}$$

$$p(x2) = U\left(0, \theta 2\right) = \begin{cases} \frac{1}{\theta 2} & if\ 0 \le x2 \le\ \theta \\ 0 & otherwise \end{cases}$$

الف) به صورت تحلیلی مرحله E الگوریتم EM را برای گام اولیه زیر محاسبه کنید.

دقت کنید که نرمالیزیشن توزیع را لحاظ کنید.

$$\theta^0 = \binom{2}{4}$$

ب) مرحله M از الگوریتم EM را با بدست آوردن پارامترها حل کنید.

پ)داده ها را روی یک نمودار دو بعدی نمایش دهید و تخمین های جدید از پارامتر ها را نمایش دهید.

سوال ۵) فرض کنید {xk} که k از 1 تا N است نمایانگر سمپل های مستقل از یکی از سه توزیع (rayleigh،exponential و beta) به فرم زیر اند. تخمین ML از تتا در هر یک از حالات زیر را بدست آورید. (۲۰ نمره)

1. $f(x_k\,;\,\theta) = \theta\, exp(-\theta x_k) \quad x_k \geq 0, \theta > 0 \quad (Exponential\ Density)$
2. $f(x_k;\,\theta) = \frac{x_k}{\theta^2}\, exp(-\frac{x^2{}_k}{2\,\theta^2}) \quad x_k \geq 0, \ \theta > 0 \ (Rayleigh\ Density)$
3. $f(x_k\,;\,\theta) = \sqrt{\theta}\, x_k{}^{\sqrt{\theta}-1} \quad 0 \leq x_k \leq 1, \theta > 0 \ (Beta\ Density)$

سوالات پیاده سازی: (در کل ۸۰ نمره)

سوال ۶) در ابتدا تابعی که با گرفتن ورودی‌های مشخص تعدادی داده از توزیع نرمال تولید کند را تکمیل کنید. (۱۰ نمره)

الف) برای تعداد bin برابر 3، 10، 30 و 300 هیستوگرام مربوطه را به همراه لیبل در یک پلات نمایش و تحلیل کنید.

ب) برای دو سری bin که یکی به صورت اعداد صحیح در وسط هر بخش و دیگری به صورت اعداد صحیح در ابتدای هر بخش نمودار هیستوگرام را بکشید و با استفاده از آن ها تحلیل کنید که هیستوگرام برای تشخیص توزیع چه معایبی دارد. به طور مثال اگر در حالت دوم لیست binها برابر [0, 3.33, 6.66, 10] باشد در حالت اول [1.66, 4.99, 8.33, 13.33] خواهد شد.

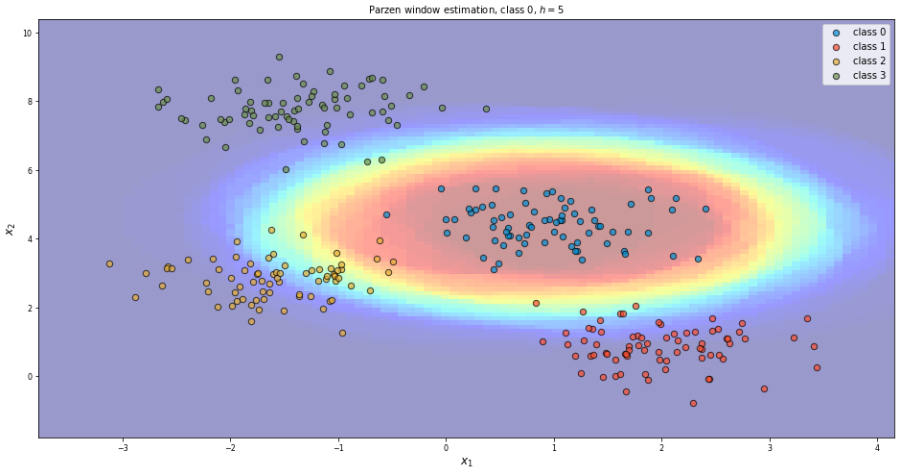
سوال ۷) در این سوال میخواهیم به پیاده سازی پارزن بپردازیم. (۲۰ نمره)

الف) در ابتدا در بخش مشخص شده دو تابع کرنل hypercube_kernel و gaussian_kernel را کامل کرده و سپس تابع parzen_window را جوری کامل کنید که pn(x) = kn/nVn را باز گرداند. در انتها تابع parzen را جوری تکمیل کنید که برای هر نقطه مشخص از فضای نمایش (mesh grid) که در X1 ذخیره شده کرنل ورودی را برای داده‌ی ورودی X فراخوانی و برای هر نقطه در X1 نتیجه کرنل را به فرم یک آرایه به خروجی بازگرداند.

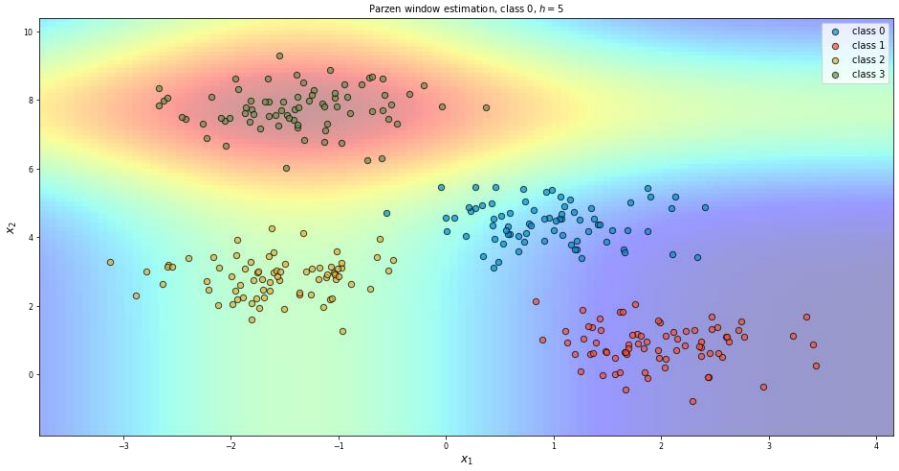
ب) تابع draw_point_distribution را جوری تکمیل کنید که آرایه پارزن هر کلاس را گرفته و مانند نمونه آن را نمایش دهد.

ج) با فرض کرنل hypercube_kernel و gaussian_kernel و h برابر عدد داده شده برا هر کلاس خروجی را نمایش دهید. (۸ نمودار خروجی)

نمونه برای خروجی هایپرکیوب:



نمونه برای خروجی گوسی:



دقت کنید که ظاهر خروجی لزوما نباید با نمونه برابر باشد و درستی خروجی شما مد نظر ماست.

سوال ۸) هدف سوال هشتم یادگیری روش کاان‌ان و تحلیل تاثیر معیار فاصله و عدد k بر روی جواب است. (۲۵ نمره)

الف) ابتدا داده‌ی cifar ۱۰ را خوانده (کد مربوط به دانلود در کولب داده شده در غیر این صورت از لینک میتوانید استفاده کنید) و مانند نمونه از هر کلاس ۷ نمونه را نمایش دهید.

ب) توابع distance_function _* را کامل کنید به طوری که برای هر داده ورودی نمونه فاصله را با داده‌ی آموزشی حساب کرده و به خروجی باز گرداند.

ج) تابع KNN را تکمیل کنید که با ورودی‌های گرفته شده y_pred مربوطه را از طریق روش kNN بدهد.

د) برای سه مقدار مختلف k نتیجه را برای هر تابع فاصله داده شده بدست آورده و مقایسه و تحلیل کنید.

سوال ۹) (۲۵ نمره)

پنج تابع چگالی احتمال گوسی N1(1, 0.11)، N2(1.6, 0.03)، N3(2.3,0.02)، N4(3.5, 0.01) و N5(4.1,0.04) را در نظر بگیرید. ابتدا با استفاده از چگالی‌های ذکر شده ۱۵۰۰ نمونه داده را بدین صورت تولید کنید که: ابتدا پنج نمونه دیتا را به ترتیب از چگالی‌های (1,2,3,4,5) تولید کنید و سپس پنج نمونه دیگر را به صورت معکوس ترتیب قبل، یعنی از چگالی‌های (5, 4, 3, 2, 1) تولید نمایید. با تکرار این روند ۱۵۰۰ نمونه دیتا را تولید نمایید.

حال برای داده‌های تولید شده تابع چگالی احتمالی را به صورت Mixture of Gaussians در نظر می‌گیریم:

$$\Sigma^{I}{}_{i=1}\, p_i\, N(\mu_i; \sigma^2{}_i)$$

الف) در نظر بگیرید $I = 5$. سپس با استفاده از الگوریتم EM و با استفاده از دادگان تولید شده، پارامترهای مجهول توزیع گوسی مخلوط را تخمین بزنید.

ب) بند (الف) را با در نظر گرفتن $I = 2$ پیاده سازی نمایید و نتیجه را مقایسه کنید.

پ) بند (الف) را با در نظر گرفتن $I = 3$ پیاده سازی نمایید و نتیجه را مقایسه کنید.

ج) بند (الف) را مجددا با استفاده از تنها ۵۰۰ نمونه از دادگان تولید شده تکرار کنید و به مقایسه‌ی نتایج بپردازید.