دانشگاه تهران

به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری دوم یادگیری ماشین

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. 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.
3. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.
4. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
5. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.
6. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل های زیر سوال خود را مطرح کنید.

سوالات ۴ و ۵ و ۸ : alireza.javid84@ut.ac.ir

سایر سوالات : arhosseini77@ut.ac.ir

سوال ۱ : (۶ نمره)

۱-۱. مفهوم bias-variance trade off را با توجه به h_n در روش پارزن و k_n در روش KNN توضیح دهید.

۱-۲. نشان دهید که مدل KNN برای ($K \neq 1$) تابع توزیع نامناسبی را تعریف میکند که انتگرال آن در تمام فضا واگرا میباشد.

سوال۲ : (۱۶ نمره)

توزیع یکنواخت $p(x)$ و پنجره پارزن $\varphi(x)$ به صورت زیر تعریف شده است.

$$p(x) \sim U(0,a)$$

$$\varphi(x) = \begin{cases} e^{-x} ; x > 0 \\ 0 ; x \leq 0 \end{cases}$$

۲-۱. نشان دهید که میانگین چنین تخمینی از پنجره پارزن به صورت زیر میشود.

$$\bar{p}_n(x) = \begin{cases} 0 ; x < 0 \\ \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right); 0 \leq x \leq a \\ \frac{1}{a}\left(e^{\frac{a}{h_n}} - 1\right) e^{-\frac{x}{h_n}} ; a \leq x \end{cases}$$

۲-۲. $\bar{p}_n(x)$ را بر حسب x برای $a = 1$ و $h_n = \{1 \ , \frac{1}{4} \ , \frac{1}{16}\}$ رسم کنید.

۲-۳. h_n چه قدر باشد تا در بازه $0 < x < a$ مقدار بایاس کمتر از ۱ درصد باشد.

۲-۴. در شرایط h_n بخش ۳-۵ و مقدار $a = 1$ ، $\bar{p}_n(x)$ را در بازه $0 < x < 0.05$ رسم کنید.

سوال ۳ : (۱۶ نمره)

توزیع نرمال $p(x) \sim N(\mu, \sigma^2)$ و پنجره پارزن $\varphi(x) \sim N(0,1)$ را در نظر بگیرید. نشان دهید که تخمین پنجره پارزن $p(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\varphi(\frac{x-x_i}{h_n})$ برای h_n های کوچک دارای ویژگی های زیر است :

- $p_n^{\sim}(x) \sim N(\mu, h_n^2 + \sigma)$
- $p_n(x) - p_n^{\sim}(x) \cong \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2 \left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right] p(x)$
- $var[p_n(x)] \cong \frac{1}{2nh_n\sqrt{\pi}} p(x)$

سوال ۴ : (۱۶ نمره)

در هر ۲ سوال زیر متغیر های X و θ به ترتیب بیانگر نمونه های مشاهده شده و پارامتر های مسئله می باشند.

۴-۱. فرض کنید که توزیع پارامتر P(θ) می باشد. مراحل expectations و maximization را برای بیشینه کردن P(θ|X) بنویسید. (محاسبات را تنها به صورت پارامتری بنویسید.)

راهنمایی: از نامساوی جنسن می دانیم:

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta) \propto \left(\sum_z Q(z)\frac{p(x,z \mid \theta)}{Q(z)}\right)p(\theta)$$

$$ln(\sum_z Q(z)\frac{p(x,z \mid \theta)}{Q(z)}) \geq \sum_z Q(z)\{\ln(p(x,z \mid \theta)) - \ln(Q(z))\}$$

۴-۲. متغیرتصادفی X با ۴ حالت طبق جدول زیر مفروض است. فرض کنید θ یک عدد حقیقی در بازه [0,1] و احتمال هر حالت مطابق زیر میباشد.

State	Probability
A	$\frac{1}{3}$
B	$\frac{1}{3}(1-\theta)$
C	$\frac{2}{3}(\theta)$
D	$\frac{1}{3}(1-\theta)$

با فرض انجام n آزمایش روی X ، حالت های A, B, C, D به تعداد n_a, n_b, n_c, n_d بار به دست آمده است. متاسفانه مقدار متغیر های n_a و n_c ناشناخته است. فرض کنید که تابع توزیع θ در ابتدا به صورت زیر نوشته شده است. مراحل E و M را نوشته و θ را بدست آورید.

$$p(\theta) = \frac{\Gamma_{(v_1+v_2)}}{\Gamma_{(v_1)}\Gamma_{(v_2)}}\theta^{v_1-1}(1-\theta)^{v_2-1}$$

راهنمایی: توجه کنید که داده های با لیبل های B و D پس از مشاهده مشخص است.

سوال ۵ : (۱۶ نمره)

فرض کنید K بازیکن در روز t وجود دارد. یکی از آنها به تعداد m_t بار بازی می کند و تعداد w_t بازی را می برد. شما تنها تعداد کل این افراد, تعداد کل راند های بازی شده و تعداد بازی های برده شده توسط بازیکن را می دانید اما نمیدانید کدام یک از K بازیکن در کدام روز بازی کرده است. شما می خواهید از یادگیری ماشین برای حل این مسئله استفاده کنید. برای هر یک از K بازیکن شما یک مدل احتمالی می سازید که در آن فرد با احتمال p_k بازی را می برد. بنابرین در روز t اگر بازیکن i ام به تعداد m_t بار بازی کرده باشد ، احتمال آن که w_t بازی را ببرد توسط یک توزیع دوجمله ای بیان می شود. (I)

در این مسئله شما باید از یک مدل ترکیب شده با K متغیر تصادفی دوجمله ای با پارامترهای $p_1, p_2, \dots, p_k$ استفاده کنید و برای N روز داده شده به صورت $(m_1, w_1), \dots (m_n, w_n)$ بکار ببرید. به این صورت که در ابتدا در روز t ما ابتدا یک بازیکن از کل آنها با احتمال π انتخاب می کنیم (c_t). در مرحله بعد در روز t این بازیکن m_t بار بازی می کند و با دانستن اینکه بازیکن c_t است, تعداد برد w_t با یک متغیر دوجمله ای توصیف می شود. (II)

۵-۱. روابط توصیف شده (I) و (II) را بنویسید.

۵-۲. مرحله E را برای بروزرسانی Q با توجه به پارامتر های مرحله قبل بدست آورید.(نشان دهید در دور i ام $Q_t^{(i)}[k]$ چیست)

۵-۳. برای هر مدل ترکیبی مرحله M برای π در دور i ام از رابطه زیر محاسبه می شود.

$$\pi^i[k] = \frac{\sum_{t=1}^{n} Q_t^i[k]}{n}$$

مرحله M را برای بروزرسانی پارامتر های مدل $p^i{}_1, p^i{}_2, \dots, p^i{}_k$ در دور i ام برحسب داده و مقادیر $Q_t^{(i)}$ بدست آورید. در ابتدا مرحله maximization را برای پارامتر ها نشان داده و مسئله بهینه سازی را حل کنید.

راهنمایی: در واقع بیشینه تابع $L = \sum_N \sum_K Q_t[k]^i \log(p(w_t \mid p_k))$ را بدست آورید.

سوال ۶: (**شبیه سازی**، ۱۵ نمره)

در این سوال هدف پیاده سازی الگوریتم KNN و استفاده از آن به عنوان طبقه بند میباشد.

۶-۱. همانطور که میدانید الگوریتم KNN ساده و شهودی است، هنگام پیش‌بینی، فاصله بین هر یک از نقاط داده موجود را محاسبه می‌کند و آن را همانند نزدیک‌ترین کلاس به آن طبقه‌بندی می‌کند. یک کلاس KNN ساخته و با استفاده از کتابخانه numpy این الگوریتم را پیاده سازی کنید.

۶-۲. مجموعه داده[1] iris را لود کرده و اطلاعات کلی دیتاست شامل تعداد کلاس و تعداد سمپل ها و فرمت داده ها و ... بیان کنید.

۶-۳. Scatter plot مجموعه داده iris را رسم کنید.

۶-۴. مجموعه داده iris را به دو دسته آموزش و ارزیابی تقسیم کنید.

۶-۵. به کمک کلاس KNN که در بخش ۱ پیاده سازی کردید ، مدلی بر روی داده های آموزش به ازای k برابر با ۵ آموزش داده سپس دقت مدل را بر روی داده های آموزش و ارزیابی گزارش کنید.

۶-۶. بخش ۵ را به ازای k های متفاوت (۱ تا ۱۰) تکرار کنید و نمودار دقت بر روی داده های ارزیابی به ازای k های متفاوت را رسم کرده و بهترین k را بر اساس آن گزارش کنید.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

سوال ۷: (**شبیه سازی**، ۱۰ نمره)

۷-۱. به کمک دستور زیر مجموعه داده X را تولید کنید.

```
import numpy as np
N = 1000
np.random.seed(1)
X = np.concatenate((np.random.normal(0, 1, int(0.3 * N)),
np.random.normal(5, 1, int(0.7 * N))))[:, np.newaxis]
```

۷-۲. توزیع دیتا X را با استفاده از روش پنجره پارزن با کرنل گوسی بدست آورید.

۷-۳. تاثیر اندازه پنجره پارزن را روی توزیع تخمین زده شده بررسی کنید (حداقل ۳ اندازه مختلف مثلا : ۱۰ و ۱ و ۰.۱)

سوال ۸: (**شبیه سازی**، ۱۵ نمره) (در بخش 2 و 3 استفاده از کتابخانه های آماده مجاز نمی باشد)

۸-۱ ابتدا دیتاست زیر را با استفاده از قطعه کد زیر ایجاد کنید.

```
from sklearn import cluster, datasets, mixture
noisy_moons=datasets.make_moons(n_samples=500, noise=0.11)
```

۸-۲. یک بار هر کلاس را با توزیع نرمال تقریب بزنید و پارامترهای آن را به دست آورده و کانتورهای مربوطه را رسم نمایید.

۸-۳. این بار از روش GMM استفاده کنید. روش GMM را با تعداد مولفه های ۱ تا ۱۶ تست کنید و شکل داده ها و کانتورها را برای تعداد مولفه برابر با ۳ و ۸ و ۱۶ بدست بیاورید.

۸-۴. تعداد مولفه های بهینه را با توجه به متریک های AIC و BIC به دست بیاورید.