دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر
آمار و احتمال مهندسی

تمرین کامپیوتری سوم ـ MSE، رگرسیون، قضیه حد مرکزی و Sampling

طراح: نوید اخوان عطار
سوپروایزر: علی محمدی

تاریخ تحویل: ۱۴۰۲/۱۰/۲۷

- **نکته ۱:** در ابتدای کدهای خود جهت یکسان بودن مقادیر رندوم از کد زیر استفاده کنید.

```
import numpy as np, random
def set_seed(seed):
    np.random.seed(seed)
    random.seed(seed)
set_seed(810109203)
```

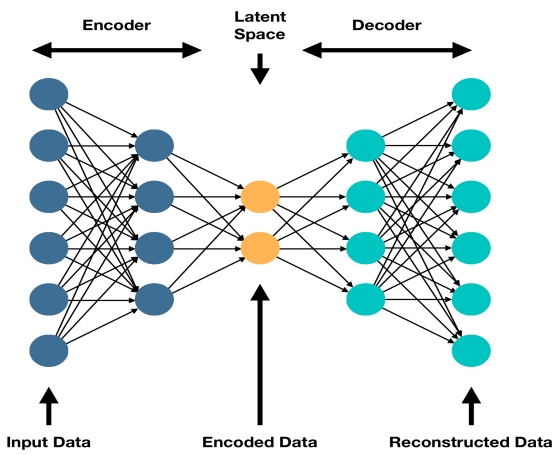
- **نکته ۲:** در تمامی سوالات سطح اطمینان مورد نیاز برای انجام آزمون‌های فرضیه آماری را 95 درصد در نظر بگیرید.

۱. Mean Squared Error (۳۵) نمره

مدل‌های «خود رمزگذار» (Autoencoder) یکی از مهم‌ترین مدل‌های حوزه یادگیری ماشین و یادگیری عمیق (Deep Learning) هستند. اتوانکودرها یک نوع شبکه عصبی هستند که به طور خاص برای فشرده‌سازی و بازسازی داده‌ها طراحی شده‌اند. توانایی اصلی یک اتوانکودر در بازسازی داده است؛ بدین معنا که می‌تواند یک ورودی را فشرده کرده و سپس بازسازی کند. یک اتوانکودر از دو بخش اصلی "Encoder" و "Decoder" تشکیل شده است:

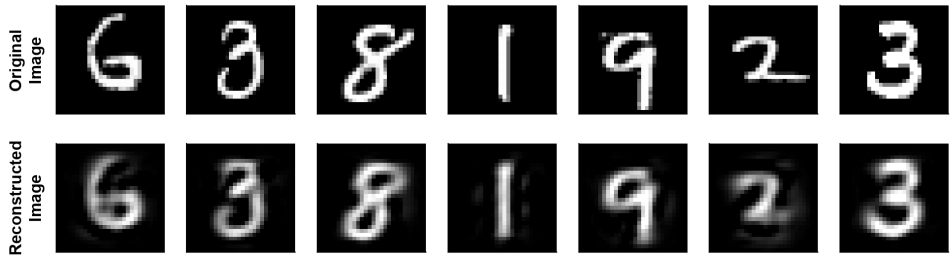
- **انکودر (Encoder):** این قسمت وظیفه تبدیل داده ورودی به فضای نهان (Latent) را دارد. انکودر، ویژگی‌های مهم و معنادار، که معمولا بُعد کمتری از ورودی دارند، را از داده‌های ورودی استخراج می‌کند.
- **دیکودر (Decoder):** وظیفه‌ی این بخش، بازسازی داده از فضای نهان است. دیکودر تلاش می‌کند با استفاده از ویژگی‌های کم‌بعد (Low Dimensional Features) تولیدشده توسط انکودر، داده را به شکل اصلی ورودی بازسازی کند.

در واقع، اتوانکودرها می‌توانند با ایجاد یک نمایش کم‌بعد از داده‌های ورودی، اطلاعات مهمی از داده را بازیابی و استخراج کنند.



۱ـ دو مورد از کاربردهایی که اتوانکودرها می‌توانند داشته باشند را به اختصار بنویسید.

۲ـ داده‌های بازسازی‌شده به کمک اتوانکودر دارای خطا هستند. در تصویر زیر نمونه‌هایی از تصاویر دیتاست mnist، که شامل تصاویر دست‌نویس از اعداد 0 تا 9 است، به همراه تصاویر بازسازی‌شده آن‌ها توسط اتوانکودر نمایش داده شده است. مشاهده می‌کنید که تصاویر بازسازی‌شده نسبت به تصاویر اصلی تار (Blur) هستند.



در مورد دلیل وجود این خطا و ارتباط آن با اندازه فضای پنهان (Latent) تحقیق کنید و نتیجه را به طور خلاصه بیان کنید.

۳ـ در این بخش می‌خواهیم تعدادی از تصاویر دیتاست mnist را توسط یک اتوانکودر از پیش آموزش داده شده (Pre-trained) بازسازی کنیم و میزان خطای بین تصاویر بازسازی‌شده و تصاویر اصلی متناظر را بدست آوریم. در فایل mnist_AE.h5 که در ضمیمه این تمرین قرار داده شده است، یک مدل اتوانکودر از پیش آموزش داده شده، ذخیره شده است.

آ. به کمک کد زیر دیتای تست دیتاست mnist را لود و پیش پردازش کنید.

```
from keras.datasets import mnist
(_, _,) (test_images, _) = mnist.load_data()
test_images = test_images.reshape(test_images.shape[0,] -1)
test_images = test_images.astype('float32') / 255.0
```

نهایتا متغیر test_images ، آرایه‌ای از 10000 تصویر که Flatten شده‌اند و مقادیر هر پیکسل آن به بازه 0 تا 1 اسکیل شده است، می‌باشد. منظور از Flatten کردن یک عکس، کنار هم قرار دادن سطرهای آن به منظور تبدیل آن به آرایه یک بعدی است. از آنجایی که اندازه هر تصویر دیتاست mnist برابر 28*28 پیکسل می‌باشد، ابعاد متغیر test_images برابر (10000 , 28*28=784) می‌باشد.

ب. به کمک قطعه کد زیر، مدل اتوانکودر Pre-trained را لود کنید و تصاویر test_images را به کمک آن بازسازی کنید.

```
import tensorflow as tf
autoencoder = tf.keras.models.load_model('mnist_AE.h5')
reconstructed_images = autoencoder.predict(test_images)
```

ج. 4 نمونه از تصاویر test_images را به همراه تصویر بازسازی‌شده متناظر با آن رسم کنید. (توجه داشته باشید که برای نمایش تصاویر Flatten شده در test_images، باید آن‌ها را به آرایه‌ای دو بعدی با ابعاد 28*28 تغییر شکل دهید)

د. تابعی برای محاسبه «میانگین مربع خطاها» (Mean Squared Error) بنویسید. سپس میزان MSE برای تمامی این 10000 تصویر بازسازی شده را بدست آورید. در نهایت هیستوگرام MSE ها را رسم کنید. (توجه داشته باشید که استفاده از کتابخانه یا تابع آماده برای محاسبه MSE مجاز نیست)

ه. در ادامه به کمک «آزمون کولموگروف‌ـاسمیرنوف» میخواهیم بررسی کنیم آیا 10000 MSE تصویر بازسازی شده دارای توزیع نرمال هستند یا خیر. آزمون کولموگروف‌ـاسمیرنوف نوعی آزمون نیکوئی برازش (Goodness of Fit) برای مقایسه یک توزیع نظری با توزیع مشاهده شده است. ابتدا با محاسبه میانگین و انحراف معیار نمونه‌ای MSE ها، μ و σ توزیع نرمال فرضی را بدست آورید. سپس به کمک دستور زیر آزمون نیکویی برازش کولموگروف‌ـاسمیرنوف را روی داده‌های MSE انجام دهید.

```
from scipy import stats
ks_statistic, p_value = stats.kstest(data, cdf='norm', args=(mean, std))
```

بر اساس p_value بدست آمده، تعیین کنید که آیا میتوان پذیرفت که داده‌های MSE از توزیع نرمال با μ و σ برآورد شده پیروی می‌کنند یا خیر؟

۲. Regression & Least Squares (۳۵) نمره

هشت نقطه اصلی و سه نقطه دیگر را که در جدول‌های زیر داده شده‌اند در نظر بگیرید. این سه نقطه به ترتیب از چپ به راست، نقطه «پرت» [1] ، نقطه «اهرمی» (نافذ) [2] و نقطه‌ای با هر دو ویژگی «دور افتادگی» و «اهرمی» هستند.

x	-2.3	-1.1	0.5	3.2	4.0	6.7	10.3	11.5
y	-9.6	-4.9	-4.1	2.7	5.9	10.8	18.9	20.5

x	5.8	20.4 (L)	20.4 (L)
y	31.3 (O)	14.1	31.3 (O)

۱ ـ در مورد نقاط پرت و نقاط اهرمی و نقاطی با هر دو ویژگی تحقیق کنید و تاثیر منفی این نقاط را بر معادله رگرسیونی توضیح دهید.

۲ ـ «ضریب تعیین» [3] (R^2) یکی از شاخص‌هایی است که میزان ارتباط خطی بین دو متغیر را اندازه‌گیری می‌کند. این ضریب می‌تواند به عنوان شاخصی برای بررسی نیکوئی برازش رگرسیون خطی استفاده شود. در مورد این ضریب تحقیق و به صورت خلاصه آن را توضیح دهید.

۳ـ تاثیر نقاط غیر عادی را با اجرای 4 رگرسیون خطی جداگانه به شرح زیر بررسی کنید.

- رگرسیون بر پایه هشت داده اصلی
- رگرسیون بر پایه هشت داده اصلی به اضافه نقطه دور افتاده
- رگرسیون بر پایه هشت داده اصلی به اضافه نقطه اهرمی
- رگرسیون بر پایه هشت داده اصلی به اضافه نقطه دور افتاده‌ـاهرمی

از روش رگرسیون خطی مبتنی بر روش کمترین مربعات خطا (Least Squares) استفاده کنید. در هر یک از چهار رگرسیون فوق، نمودار داده‌ها همراه با خط رگرسیون را در یک صفحه رسم کنید و ضریب تعیین (R^2)هر یک را بیان کنید. (توجه داشته باشید که استفاده از کتابخانه یا تابع آماده برای پیاده سازی رگرسیون خطی مجاز نیست)

۴ـ راهکارهایی برای یافتن مدل رگرسیونی بهتر (نسبت به مدل مبتنی بر کمترین مربعات خطا) در حضور نقطه دور افتاده و یا اهرمی پیشنهاد کنید.

[1] Outlier
[2] High Leverage Point
[3] Coefficient of Determination

۳. Central Limit Theorem & Sampling (۴۰ نمره)

تصویر یک «**تابلوی گالتون**» (**Galton Board**) ـ در این تخته، تعداد زیادی توپ از بالا به پایین سرازیر میشوند. در طی مسیر چندین لایه از موانع وجود دارند که هر توپ در هر مرحله با برخورد به این موانع، به یکی از دو سمت راست یا چپ منحرف می‌شوند. این توپ‌ها نهایتا توزیعی شبیه توزیع نرمال ایجاد می‌کنند.

دیتاست ضمیمه‌شده FIFA2020.csv شامل اطلاعات مربوط به بهترین بازیکنان تاریخ فوتبال جهان تا سال 2020 می‌باشد که شامل ستون‌هایی مانند: ملیت (nationality)، امتیاز (overall)، وزن (weight)، قد (height) ، توانایی شوت زدن (shooting) ، توانایی دریبل زدن (dribbling) ، سرعت (pace) و... می باشد. در واقع هر یک از ستون‌ها یک متغیر تصادفی می‌باشد.
برای لود کردن دیتاست از دستور زیر استفاده کنید.

```
import pandas as pd
df = pd.read_csv('FIFA2020.csv', encoding = "ISO-8859-1")
```

۱ـ ،در این دیتاست، تعدادی از داده‌های کمّی N/A (Not A Number) هستند و همچنین تعدادی از داده‌های کیفی، Icons هستند که نشان‌دهنده نامعلوم بودن این مقادیر می‌باشد. برای جایگزین کردن داده‌های کمّی نامعلوم چه راهکاری پیشنهاد می‌کنید؟ راهکار خود را برای داده‌های ستون (pace) و ستون (dribbling) پیاده کنید و دیتاست جدید را جایگزین دیتاست قبل کنید.

۲ـ نمودار جعبه‌ای متغیر تصادفی age را رسم کنید و مقادیر (min, Q1, Q2, Q3, max) را بدست آورید. به صورت خلاصه توضیح دهید هر کدام از این مقادیر به چه معنا هستند.

۳ـ متغیر تصادفی weight را در نظر بگیرید و به صورت تصادفی و بدون جایگذاری، $n = 100$ نمونه از این متغیر انتخاب کنید:

آ. میانگین، واریانس و انحراف معیار این نمونه‌ها را بیابید.

ب. یکی از ابزارهایی که برای مقایسه شهودی دو توزیع به کار می‌رود، نمودار Q-Q می‌باشد. نحوه استفاده از این نمودار را در یک یا دو جمله توضیح دهید.

ج. یک نمونه‌ی $n=100$ تایی از توزیع نرمال با μ و σ (میانگین و واریانس نمونه‌ای n نمونه) برآورد شده در قسمت " آ" ایجاد کنید. سپس با استفاده این دو مجموعه n تایی و نمودار Q-Q، توزیع آماری وزن بازیکنان را با توزیع نرمال مقایسه کنید و نتیجه را تحلیل کنید.

د. در ادامه به کمک آزمون Shapiro-Wilk مشخص کنید که آیا توزیع آماری وزن 100 بازیکن انتخاب شده از توزیع نرمال پیروی می‌کند یا نه. آزمون «شاپیرو ویلک» (Shapiro-Wilk Test) از آزمون‌های برازش توزیع نرمال محسوب می‌شود. به کمک این آزمون می‌توان مشخص کرد که آیا داده‌ها از توزیع نرمال پیروی می‌کنند یا خیر. برای پیاده سازی این آزمون از کد زیر استفاده کنید:

```
import scipy.stats as stats
statistic, p_value = stats.shapiro(data)
```

ه. سپس قسمت‌های " آ، ب، ج " را به ازای $n = 500, 2000$ تکرار کنید، چه نتیجه‌ای می‌گیرید؟

۴ـ یکی از توزیع‌های آماری مهم، «توزیع پواسون» (Poisson) است. این توزیع بیانگر رویدادهایی است که در طول زمان اتفاق می‌افتند و فقط میانگین فاصله‌ی بین این رویدادها را از داده‌های گذشته می‌دانیم:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad (x \in \mathbb{Z})$$

آ. به ازای $\lambda = 3$ تعداد $n = 5000$ از این توزیع بدون جایگذاری نمونه‌برداری کنید و هیستوگرام آن را رسم کنید.

ب. به ازای $n = 5, 50, 5000$ و $\lambda = 3$ با استفاده از نمودار Q-Q توزیع این نمونه‌ها را با توزیع نرمال مقایسه کنید. سپس p_value آزمون Shapiro-Wilk را برای هر یک بدست آورید و فرضیه نرمال بودن توزیع هر یک از این نمونه‌ها را آزمون کنید. نهایتا نتایج بدست آمده برای این 3 نمونه را بر اساس قضیه‌ی حد مرکزی (CLT) توجیه کنید.