سلام بر تمام دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.
2. نکته 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. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده از کنید. شکل‌ها به طور واضح و در فرمت درست گزارش شوند.
4. از بین سوالات شبیه سازی حتما به هر دو مورد پاسخ داده شود. حداکثر تا ۱۱۰ نمره (۱۰ نمره امتیازی) لحاظ خواهد شد.
5. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر می‌شود.
6. در صورت داشتن سوال، از طریق ایمیل m.zarvandi97@gmail.com ، سوال خود را مطرح کنید.

---

1. دیتاستی متشکل از n نقطه به صورت $(x_i, y_i), x_i \in \mathbb{R}^{d\times 1}$ از مدل خطی زیر گرفته شده است:(۳۰ نمره)

$$y = x^T\beta^\star + \epsilon$$

با در نظر گرفتن مدل رگرسیون خطی L_2 رگولارایز شده زیر با پارامتر رگولاریزیشن $\lambda \geq 0$ به صورت زیر:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

ماتریس $X \in R^{n\times d}$ که در هر سطر آن x_i^T برای هر نقطه قرار دارد را درنظر بگیرید:

1. فرم بسته $\hat{\beta}_\lambda$ را بیابید.

2. عبارت مربوط به بایاس $\mathbf{x}^\top \boldsymbol{\beta}^\star - \mathbb{E}\left[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda\right]$ را به عنوان تابعی از λ و یک x ثابت به دست آورید.
3. عبارت مربوط به واریانس $\mathbb{E}\left[\left(\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda - \mathbb{E}\left[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda\right]\right)^2\right]$ را به عنوان تابعی از λ و یک x ثابت به دست آورید.
4. با استفاده از قسمت های 2 و 3 و نظریه بایاس- واریانس، تاثیر پارامتر λ را در خطای مربعی (squared error) بررسی کنید. مشخص کنید هنگامی که λ کوچک یا بزرگ باشد، کدام عبارت غالب است؟

---

2. الگوریتم گرادیان کاهشی را توضیح دهید، سپس روابط آن را برای تابع هزینه زیر محاسبه نمایید.(۱۰ نمره)

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{q}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$h(x) = \frac{e^{(wx+b)}}{1+e^{(wx+b)}}$$

---

3. متغیرهای تصادفی X, Y را در نظر بگیرید. داریم: (۲۰ نمره)

$$E(X) = \mu_x, E(Y) = \mu_y, \mathrm{Var}(X) = \sigma_x^2, \mathrm{Var}(Y) = \sigma_y^2, \mathrm{Cov}(X, Y) = \sigma_{xy}, r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

در رابطه $\hat{Y} = \alpha + \beta X$، که α, β به منظور کمینه کردن $E(Y - \hat{Y})^2$، امید مربع خطای پیش بینی، انتخاب شده‌اند:

- نشان دهید که مقدار α, β برابر است با:

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$\alpha = \mu_y - \beta\mu_x$$

• نشان دهید که برای این α, β داریم:

$$\frac{\mathrm{Var}(Y) - \mathrm{Var}(Y - \hat{Y})}{\mathrm{Var}(Y)} = r_{xy}^2$$

---

4. با در نظر گرفتن رابطه $y = \beta_0 x + \beta_1 x^2$ برای برازش نقطه‌های $(x_i, y_i)\ for\ i = 1,\ldots,k+1$. (۳۰ نمره)

الف) با استفاده از روابط نویسی به صورت ماتریس، تخمین *least square* را از ضرایب رابطه به دست آورید.

ب) یک رابطه برای ماتریس کواریانس حاصل از تخمین‌ها به دست آورید.

ج) مقدار MSE مربوط به پیش‌بینی را به دست آورید.

---

5. با در نظر گرفتن جدول زیر، که مربوط به یک مسیله رگرسیون خطی ساده می باشد، سوالات مربوطه را بررسی کنید. (۲۰ نمره)

i	x_i	y_i
1	*4*	*31*
2	*9*	*58*
3	*10*	*65*
4	*14*	*73*
5	*4*	*37*
6	*7*	*44*
7	*12*	*60*
8	*22*	*91*
9	*1*	*21*
10	*17*	*84*

الف) مقدار پارامترهای $\beta_0, \beta_1, \sigma^2$ و واریانس مربوط به β_0, β_1 را محاسبه نمایید.

ب)مقدار کورلیشن مربوط به دو پارامتر β_0, β_1 را به دست آورید.

---

6. (شبیه‌سازی) در این سوال، طبقه بندی طراحی کنید که بتوانیم، که ۲ کلاس متفاوت (دو تیم فوتبال منچستریونایتد و چلسی) با استفاده از دیتاست داده شده، را تشخیص دهیم. جهت طبقه‌بندی، میتوانید میانگین رنگ در هر عکس را محاسبه نمایید، سپس بر اساس مقدار به دست آمده، با مقدار رنگ آبی و قرمز مقایسه نمایید.. برای دیتاست داده شده، این طبقه بند را تست کنید. ماتریس *Confusion* را گزارش دهید. مقادیر *accuracy*، *precision* و *recall* را محاسبه کنید، و نتایج هر کدام را توضیح دهید. (۲۰ نمره)

---

7. (شبیه‌سازی) در این سوال می‌خواهیم، *overfitting* و *under-fitting* را برای یک سری داده بررسی کنیم. ابتدا، دیتا ست *Dummy Data HSS.csv* را لود کنید، این دیتاست میزان هزینه تبلیغات کالاها را در تلویزیون، رادیو و شبکه‌های اجتماعی و همچنین میزان تاثیرگذار بودن هرکدام و در آخر میزان فروش محصول را نشان می‌دهد. میزان فروش محصول را به عنوان خروجی فرض کنید. ابتدا داده‌ها را به سه دسته آموزش، تست و ارزیابی تقسیم کنید. (تعیین درصد اختصاص داده به هر دسته را خودتان به شکلی که قابل قبول باشد تعیین کنید.) حال سعی کنید که تابع درجه یک تا هفت برای این داده‌ها برازش کنید. مقادیر *MSE* را برای داده‌های آموزش و ارزیابی در هر درجه تعیین و هر دو را بر روی یک نمودار نمایش دهید و براساس آن تعیین کنید که بهترین درجه برای تخمین میزان فروش براساس این ویژگی ها چقدر است. مقادیر بایاس، واریانس و MSE را برای همه درجات از یک تا هفت برای داده‌های تست نیز به صورت جداگونه بدست بیاورید و روی یک نمودار نمایش دهید. مشاهده خود را از نتایج به دست آمده شرح دهید. .(۲۰ نمره)
