به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

آمار و احتمال مهندسی

پروژه‌ی نهائی (نسخه‌ی دوم)

طراح:	محمد ربیعی قهفرخی
تاریخ ویرایش این نسخه	۱۸ دی ۱۳۹۹
مهلت تحویل گزارش	۲۴ دی ۱۳۹۹

فهرست مطالب

۱ ضریب اویلر-ماسکرونی

عدد صحیح بزرگی نظیر n در نظر بگیرید. اگر n را بر عدد صحیح r که کوچکتر و مساوی آن تقسیم کنید و حاصل تقسیم را به نزدیک‌ترین عدد صحیح بزرگ‌تر گرد کنیم، در طی عمل گرد کردن به ازای هر r، مقدار ϵ_r مرتکب خطا شده‌ایم. در نگاه اول به نظر می‌رسد به‌طور متوسط، امید ریاضی ϵ_r برابر نیم می‌باشد اما چارلز والی پوسین در سال ۱۸۹۸ ثابت کرد که مقدار حدی این خطا به ثابت اویلر-ماسکرونی [1] هم‌گرا می‌شود. این ثابت به‌عنوان فاصله‌ی حدی nامین عدد هارمونیک و لگاریتم n تعریف می‌شود:

$$\gamma = \lim_{n\to\infty}\left(\sum_{r=1}^{n}\frac{1}{r}-\log(n)\right) \simeq 0/5772 \qquad (1)$$

با توجه به این مقدمه، به سوالات زیر پاسخ دهید:

آ) قصد داریم امید ریاضی خطا را محاسبه کنیم. مشخص کنید متغیر تصادفی ما مقدار r است یا n؟

ب) در یک اسکریپت پایتون، همگرا شدن امید ریاضی خطا را نشان دهید. نموداری که نشان‌دهنده‌ی هم‌گرا شدن این مقادیر به ثابت اویلر باشد. هم‌چنین مقادیر δ_n را برای nهای مختلف محاسبه کنید و در یک نمودار دیگر رسم کنید. رویکرد خود در حل مساله را به‌همراه نمودارها در گزارش خود درج کنید.

$$\delta_n = \frac{\hat{\epsilon_r}(n+1)-\gamma}{\hat{\epsilon_r}(n)-\gamma} \qquad (2)$$

شکل ۱: Charles Vallée Poussin (1866 – 1962)

[1] Euler-Mascheroni Constant

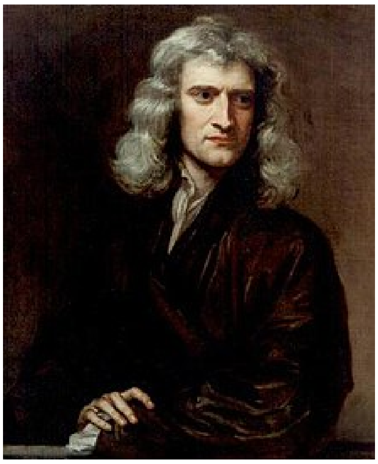
۲ مساله‌ی نیوتن-پیپس

ساموئل پیپس، وقایع‌نویس انگلیسی، در سال ۱۶۹۳ در طی نامه‌ای از آیزاک نیوتن درخواست کرد که یه مساله‌ی احتمالاتی درباره‌ی یک شرط‌بندی را ارزیابی کند. پرسش ساموئل این بود که کدام یک از موارد زیر محتمل‌تر است:

- در پرتاب شش تاس، حداقل یکی از تاس‌ها ۶ باشد.
- حداقل دو تاس از ۱۲ تاس پرتاب شده ۶ باشد.
- حداقل سه‌تاس از ۱۸ تاس پرتاب شده ۶ باشد.

آ) با مراجعه به سایت ویکی‌پدیا پاسخ نیوتن را ارزیابی کنید. نیوتن یک تحلیل مفهومی برای نتیجه‌گیری خود ارائه کرده است که به‌نظر ناصحیح می‌رسد. آیا می‌توانید تحلیل بهتری برای نتیجه‌ی مشاهده شده ارائه کنید؟

ب) حال مساله‌ی پیپس را تعمیم می‌دهیم. فرض کنید P_n احتمال مشاهده‌ی nتا شش در پرتاب $6n$ تاس باشد. مشابه سوال قبل، نشان دهید که مقادیر P_n به مقدار ۰/۵ هم‌گرا می‌شوند.



Sir Isaac Newton (1642 – 1726) (آ)

Samuel Pepys (1633 – 1703) (ب)

۳ تخمین عدد نپرین با روش مونته کارلو

احتمالا با ثابت اویلر آشنا هستید (e). این عدد گاهی با نام عدد نپر نیز شناخته می‌شود اگرچه هر دو نام‌گذاری به عنوان ادای احترام به این ریاضی‌دانان صورت گرفته و کاشف این عدد شخص دیگریست. اولین حضور عدد نپر در مقالات در سال ۱۶۱۸ توسط جان نپر، مخترع لگاریتم، انجام گرفت. اگرچه وی به خود ثابت دست نیافته بود، بلکه لیستی از لگاریتم‌ها که با کمک این ثابت به‌دست آمده بودند را منتشر کرده بود. البته احتمالا لیست ذکر شده توسط ویلیام اوترد تهیه شده بوده است. کشف واقعی این ثابت ۶۵ سال بعد و توسط ژاکوب برنولی در سال ۱۶۸۳ در طی بررسی حد عبارت زیر انجام گرفت:

$$\lim_{n\to\infty}\left(۱+\frac{۱}{n}\right)^{n} \tag{۳}$$

با توجه به این که حداقل دانشجوی ترم سوم مهندسی هستید، احتمالا با کاربرد وسیع این عدد در اکثر زمینه‌های ریاضی آشنا هستید. در این تمرین قصد داریم به روش مونته‌کارلو مقدار این ثابت را تخمین بزنیم. روش‌های مونته‌کارلو دسته‌ی وسیعی از الگوریتم‌های محاسباتی هستند که برپایه‌ی نمونه‌گیری تصادفی از محیط بنا شده‌اند. در حقیقت در این روش‌های سعی می‌کنیم از تصادفی بودن نمونه‌ها استفاده کنیم تا مقداری را حساب کنیم که ذاتا غیرتصادفیست. نخستین پژوهش‌ها روی این متد با حل مساله‌ی سوزن بوفون آغاز شد و پس از آن (تقریبا از سال ۱۹۳۰) از این متد درشبیه‌سازی‌های حوزه‌های بیولوژی، فیزیک، مهندسی و ...استفاده شد. در طی شرح مساله بهتر با این مفهوم آشنا خواهید شد.

William Oughtred
(1574 - 1660)

Jacob Bernoulli
(1655 - 1705)

John Napier
(1550 - 1617)

Leonhard Euler
(1707 - 1803)

قصد داریم ثابت اویلر را تخمین بزنیم. منحنی $f(x) = e^{(-x)}$ را در نظر بگیرید. (شکل ۴آ) توجه خود را به دامنه‌ی $(\cdot, 1)$ معطوف می‌کنیم: (قسمت داخل مستطیل مشخص شده)

آ) با فرض دانستن ثابت اویلر، مساحت بخشی از مستطیل جدا شده از تصویر ۴آ را محاسبه کنید که زیر منحنی نمایی قرار گرفته است. (به صورت پارامتری و بر حسب e)

ب) حال فرض کنید مساحت قسمت مشخص شده در قسمت (آ) را از پیش می‌دانیم (S). با توجه به رابطه‌ی به‌دست آمده از بخش قبل، مقدار e را بر حسب S به دست آورید.

ج) حال از فضای دو بعدی محصور به مستطیل مشخص شده، n نمونه از یک توزیع یکنواخت دو بعدی که روی این فضا تعریف شده است استخراج می‌کنیم. با توجه به قاعده‌ی قانون اعداد بزرگ، اگر تعداد نمونه‌ها به سمت بی‌نهایت میل کند، نسبت تعداد نمونه‌هایی که پایین منحنی قرار گرفته‌اند به تعداد کل نمونه‌ها، برابر با نسبت مساحت سطح زیر نمودار به مساحت کل مستطیل است. یعنی اگر به تصویر شکل ۴ب توجه کنید، داریم:

$$\frac{S}{S_{\text{rect}}} = \lim_{n\to\infty}\left(\frac{n_{\text{red}}}{n}\right) \qquad (۴)$$

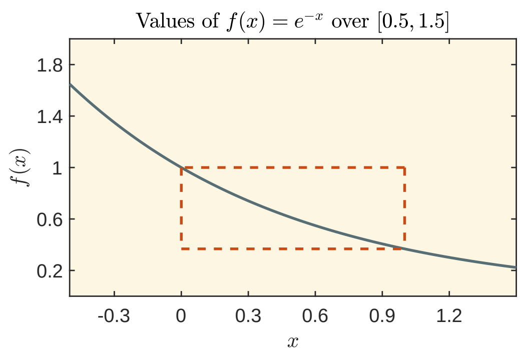
به این ترتیب می‌توان ثابت اویلر را به کمک روش مونته‌کارلو تخمین زد. با فرض $n = ۱۰۰$ تخمین $\hat{e}$ را ارائه دهید.

د) تخمین ارائه‌شده در قسمت (ج)، تابعی از نمونه‌های گرفته‌شده از محیط است. در نتیجه این تخمین، خود یک متغیر تصادفی می‌باشد به عبارت دیگر اگر اسکریپت قسمت قبل را چندبار اجرا کنید، احتمالا مقادیر متفاوتی دریافت خواهید کرد. فرض کنید این متغیر تصادفی، توزیعی گاوسی داشته باشد. برای تخمین مقادیر امید ریاضی و واریانس آن می‌بایست چند نمونه از آن گرفت و سپس با استفاده از تخمین بیشینه درست‌نمایی [۲] مقادیر میانگین و واریانس را برای یک n ثابت و بزرگ محاسبه کرد. توزیع تخمینی از متغیر تخمین ارائه‌شده در قسمت (ج) را به‌دست آورید.

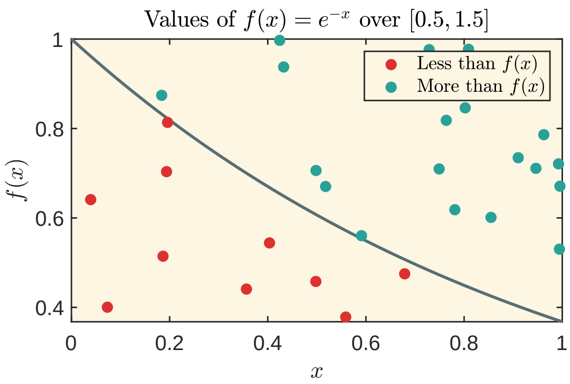
ه) در نهایت می‌توان امید ریاضی توزیع تخمین زده شده را به عنوان تخمین بهتر ارائه کرد و واریانس تخمین گاوسی فوق معیاری از عدم قطعیت ما نسبت به تخمینیست که ارائه کرده‌ایم. بدیهیست

[2] Maximum Likelihood Estimation (MLE)

که عدم قطعیت ما هیچ‌گاه صفر نخواهد شد، به عنوان مثال یکی از اولین عوامل ایجاد کننده‌ی عدم قطعیت تقریبیست که روی تعداد نمونه‌ها اعمال کردیم. بنابر رابطه‌ی ۴ تعداد نمونه‌ها باید به بی‌نهایت میل کند ولی ما در بخش (ج) و (د) تعداد نمونه‌ها را برابر با ۱۰۰ فرض کردیم. پس می‌توان انتظار داشت هرچه n را بزرگتر کنیم، عدم قطعیت از تخمین کمتر شود و به صفر میل کند. برای ارزیابی این فرضیه، تعداد نمونه‌ها را بین ۱ تا ۱۰۰۰ تغییر دهید و به ازای هر مقدار n، امید ریاضی و واریانس توزیع تصادفی تخمین را محاسبه کنید. سپس در یک نمودار، مقادیر امید ریاضی را برحسب تعداد نمونه‌ها رسم کنید سپس به ازای هر داده روی نمودار، با یک خط عمودی بازه‌ی اطمینان ۹۵ درصدی داده‌ها را مشخص کنید. [۳] آیا فرضیه‌ای که مطرح کردیم تایید شد؟ اگر پاسخ منفیست، چرائی این موضوع را بررسی کنید.



(آ) منحنی تابع $\exp(-x)$



(ب) پخش شدن نمونه‌ها روی فضا

[۳] نمونه ای از این نمودارها را با عنوان error bar در اینترنت جست و جو کنید.

۴ سری فیبوناچی تصادفی

با دنباله‌ی معروف فیبوناچی هستید. این دنباله با مقادیر اولیه‌ی $F_۱ = F_۲ = ۱$ به صورت رابطه‌ی ۵ تعریف می‌شود.

$$F_n = F_{n-۱} + F_{n-۲} \tag{۵}$$

n امین عضو دنیاله‌ی فیبوناچی به صورت مجانبی به ϕ^n هم‌گرا می‌شود که ϕ نسبت طلایی نامیده می‌شود. یعنی:

$$\lim_{n\to\infty} (F_n)^{۱/n} = \phi \tag{۶}$$

دنباله‌ی فیبوناچی تصادفی به شکل رابطه‌ی ۷ تعریف می‌شود که مقدار β_n یک متغیر تصادفی گسسته است که مقادیر ± ۱ را با احتمال برابر در بر می‌گیرد.

$$f_n = f_{n-۱} + \beta_n f_{n-۲} \tag{۷}$$

مشخصا عبارت f_n یک متغیر تصادفی پیچیده و وابسته به α_nهای قبلی خود می‌باشد. در نتیجه این دنباله مقدار مجانبی ندارد. اگرچه دنباله‌ی فیبوناچی تصادفی همواره کوچکتر از دنباله‌ی فیبوناچی معمولیست، اما مقدار مجانبی ندارد، چون با میل کردن n به سمت بی‌نهایت، مقدار f_n بین مقادیر مثبت و منفی نوسان می‌کند. اما در سال ۱۹۶۰، فورستنبرگ و کستن [4] ثابت کردند که قدرمطلق دنباله به صورت مجانبی به α^n هم‌گرا می‌شود.

آ) به کمک روش‌هایی که در این پروژه آموختید، یک بازه‌ی اطمینان ۹۵درصدی برای α تعیین کنید.

ب) دنباله‌های nتایی از این دنباله را در نظر بگیرید. بزرگترین عبارت ظاهر شده را با $f_{\max}(n)$ نمایش می‌دهیم که خود یک متغیر تصادفی با ساپورت مثبت است. می‌خواهیم توزیع این متغیر تصادفی را به صورت یک توزیع نمایی تخمین بزنیم. ابتدا رابطه‌ی تخمین بیشینه‌ی بزرگ‌نمایی پارامتر توزیع نمایی (λ) را اثبات کنید، سپس با نمونه‌گیری از $f_{\max}(۲۵)$، یک توزیع بر آن برازش کنید.

[4] Furstenberg and Kestenm, Products of random matrices, Ann. Math Stat. 31, 456-469

ج) یک روش ناپارامتری تخمین تابع چگالی براساس هیستوگرام توزیع می‌باشد. به کمک نمونه‌هایی که در قسمت قبل گرفتید، تابع چگالی تجربی را به داده‌ها برازش کنید. سپس این تابع چگالی را به همراه تابع تخمین زده شده در قسمت قبل در یک نمودار رسم کنید.

د) همان‌طور که پیشتر به آن پرداخته شد، تخمین پارامتر توزیع در قسمت (ب) خود یک متغیر تصادفی است. اثر افزایش n را بر امید ریاضی λ با رسم نمودار بررسی کنید. نتیجه را تحلیل کنید.

۵ قاعده‌ی دایره‌ای برای مقادیر ویژه

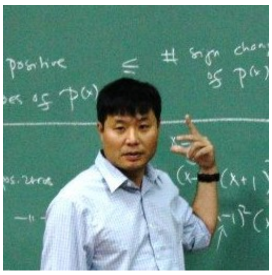
با مقادیر ویژه در ریاضی۲ آشنا شده‌اید. λ مقدار ویژه‌ی ماتریس M نامیده می‌شود اگر بردار غیر بدیهی مانند v وجود داشته باشد که $Mv = \lambda v$. هم‌چنین برای به‌دست آوردن مقادیر ویژه، چنین عمل می‌کردیم:

$$
\begin{aligned}
& Mv = \lambda v && (۸)\\
\Rightarrow \quad & 0 = (\lambda I - M)\, v && (۹)\\
\Rightarrow \quad & 0 = \det(\lambda I - M) && (۱۰)
\end{aligned}
$$

می‌دانید که لزوما مقادیر ویژه‌ی ماتریس M حقیقی نیستند و می‌توانند به فضای اعداد مختلط تعلق داشته باشند. قانون دایره [5] در احتمالات بیان می‌کند که مقادیر ویژه‌ی یک ماتریس $n \times n$ که درایه‌های آن نمونه‌هایی .i.i.d از یک توزیع نرمال با میانگین صفر و واریانس $\frac{1}{n}$ هستند به داخل دایره‌ی واحد هم‌گرا می‌شوند، اگر $n \rightarrow \infty$. این قاعده اولین بار توسط ژان ژینیبر در سال ۱۹۶۰ برای توزیع گاوسی مطرح شد. ویچسلاو گیرکو در سال ۱۹۸۰ این قاعده را به‌روز رسانی کرد تا توزیع‌های بیشتری را در بر بگیرد. ترنس تائو و وان ه. و. در سال ۲۰۱۰ این قاعده را در حالت کلی تعمیم دادند تا همه‌ی توزیع‌هایی که در دو شرط زیر صدق می‌کنند را در بر گیرد.

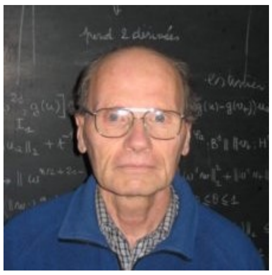
- $\mathcal{E}(X) = 0$
- $\mathcal{E}(X^2) = \dfrac{1}{n}$

Terence Tao
(1975 - -)



Van H. Vu
(1970 -)

Vyacheslav Girko
(1550 - 1617)



Jean Ginibre
(1946(?) -)

[5] Girko Circular Law

آ) درستی این قاعده را توسط رسم نمودار با nهای مختلف روی توزیع‌های گاوسی و یکنواخت بررسی کنید.

ب) حال برای مقادیر ۵، ۱۵ و ۶۰، ۱۰۰ ماتریس تصادفی از توزیع نرمال نمونه‌گیری کنید و مقادیر ویژه‌ی هر ۱۰۰ ماتریس را در یک نمودار به ازای هر n رسم کنید. آیا توزیع مقادیر ویژه برای nهای مختلف به صورت یکنواخت انجام گرفته است؟

۶ داده بازی

در این بخش سعی می‌کنیم به بررسی یک مجموعه داده‌ی واقعی بپردازیم. مجموعه‌ی داده‌ی ماشین‌ها بر روی سامانه قرار گرفته است. با مراجعه به این نیز می‌توانید به داده‌ها دسترسی پیدا کنید. در این مجموعه داده هر سطر متناظر با یک خودرو می‌باشد که بعضی ویژگی‌های آن در ستون‌های مختلف مجموعه داده جمع‌آوری شده است.

آ) سوالات رسم نمودار:

(آ) تنها با رسم نمودار پراکندگی و نگاشت سایر ویژگی‌ها به خصیخه‌های نمودار، فرض‌های زیر را به‌صورت شهودی بررسی کنید.

i. هرچه ماشینی بلندتر باشد، عریض‌تر هم هست را واکاوی کنیم.

ii. در توزیع نقاط مربوط به طول و عرض ماشین‌ها، نقاط گوشه‌ی راست‌تر (با طول بیشتر) خودروهای چهار در هستند.

iii. در توزیع نقاط مربوط به طول و عرض ماشین‌ها، تمرکز خودروهای هاچبک به سمت گوشه‌ی پایین‌چپ صفحه است (با طول و عرض کمتر)

iv. خودروهای عریض مقاومت بیشتری در برابر هوا دارند. هم‌چنین احتمالا سنگین‌تر نیز هستند در نتیجه ظرفیت موتور بیشتری دارند.

(ب) مشابه موارد فوق، حداقل سه فرض مطرح کنید و نمودار پراکندگی مناسبی برای دید شهودی به فرض خود رسم کنید

(ج) مفهوم لغزش [۶] در نمودار پراکندگی را توضیح دهید. سپس یک نمودار پراکندگی از موارد رسم شده در بالا را با استفاده از این مفهوم بهبود دهید.

(د) در دل داده‌ها بکاوید. حداقل ۵ نمودار از انواع دیگر (میله‌ای، دایره‌ای و ...) رسم کنید و سعی کنید با نگاشت سایر ستون‌ها به خصیصه‌های هر نمودار، نکات نهفته در دل داده‌ها را بیرون آورید.

[6] jitter

ب) دو آزمون فرض دل‌خواه طراحی کنید و برروی داده‌ها پیاده‌سازی کنید. آیا این نتیجه با انتظارات شما همسوست؟

ج) با مفهوم رگرسیون آشنا هستید. با فرض انتخاب دو ویژگی x و y از یک پدیده، وجود رابطه‌ی خطی ادعا می‌کند که:

$$y = \theta^T x + \epsilon \tag{11}$$

که x یک کمیت عددی غیر تصادفی و ϵ یک متغیر تصادفی نرمال با میانگین صفر می‌باشد. در نتیجه بنا به رابطه‌ی ۱۱ متغیر y نیز یک متغیر تصادفی نرمال با میانگین x و واریانس σ_ϵ می‌باشد. با توجه به دیدگاه آماری رگرسیون خطی، رابطه‌ای برای تخمین پارامترهای θ بر اساس روش بیشینه درست‌نمائی [7] بیابید و سپس رابطه‌ی خطی بین دو متغیر دل‌خواه از داده‌ها را بدون استفاده از کتابخانه‌های آماده‌ی رگرسیون به دست آورید.

[7] maximum likelihood

۷ توضیحات

دانشجویان عزیز حتما به نکات زیر توجه داشته باشند.

- پروژه به‌گونه‌ای طراحی شده که با دانش آماری فراتر از آن‌چه در این درس آموخته‌اید نیاز نداشته باشد وآن‌چه را که آموخته‌اید تثبیت و تفهیم می‌کند. به همین جهت انجام آن برای یادگیری درس اکیدا توصیه می‌شود.
- صرف‌نظر از رویکرد آموزشی این پروژه، آخرین نقطه‌ی جبران نمراتتان در این درس می‌باشد و بنا به سابقه‌ی چندساله، به اسکیل‌شدن نمرات امیدی نیست، در نتیجه از اهمیت این موقعیت غافل نشوید.
- شما می‌بایست علاوه بر کدهای پیاده‌شده، گزارشی تحلیلی از نتایج خود ارائه دهید. توجه داشته باشید که مفهوم گزارش پروژه با مفهوم توضیح کد متفاوت است در نتیجه در فایل گزارش، از درج کد جدا بپرهیزید.
- کدهای پایتون و آر خود را حتما در قالب دفترچه‌ی ژوپیتر بارگذاری کنید. دستیاران آموزشی موظف به اجرای کدهای شما نیستند.
- اسکریپت‌های خود را خوانا و تمیز بنویسید. طبیعتا این درس، درس برنامه‌نویسی نیست اما کد بسیار پیچیده و غیرقابل فهم نمره‌ی کامل را دریافت نمی‌کند. استفاده از توابع و نام‌های متغیرهای بامعنا به خوانایی کد می‌افزاید.
- گزارش‌کار، اولین و مهم‌ترین آیتم نمره‌دهی می‌باشد در نتیجه با صرف زمان مناسب، گزارشی تهیه کنید که بازتاب‌گر زحماتی باشد که برای انجام پروژه کشیده‌اید. استفاده‌ی صحیح از نیم‌فاصله، علائم نگارشی، گویا بودن جملات و پاراگراف‌بندی مناسب از جمله مواردیست که در نگاه اول جلب توجه می‌کند و نکاتی نظیر استفاده از زیرنویس برای تصاویر و بالانویس برای جداول، ارجاع‌دادن به روابط و تصاویر با شماره‌ی مربوط به هر کدام و ... از جمله خصوصیت‌های یک نوشته‌ی آکادمیک است. متن گزارش را با فونت B Nazanin و اندازه‌ی ۱۴ در قالب گزارش قرار داده شده روی سایت تایپ نمائید. از قرار دادن عکس از نوشته‌ی دست‌نویس خود در گزارش به‌شدت پرهیز کنید و روابط ریاضی را نیز تایپ کنید.
- با توجه به مفهوم امتیازی بودن پروژه، به شدت با موارد تقلب چه در کد و چه در گزارش برخورد خواهد شد.
- سعی می‌شود از برخی از دوستان از طریق تماس تصویری سوالاتی در قالب تحویل پروژه پرسیده شود. در نتیجه مشخص است که هر شخص باید به تمامی محتوایی که ارائه می‌دهد مسلط باشد.
- در نهایت یک فایل گزارش پی‌دی‌اف را در کنار دفترچه‌های ژوپیتر زیپ کرده و با نام <sid>-surname.zip در صفحه‌ی درس بارگذاری کنید.
- ابهامات خود در مورد سوالات و یا قالب گزارش در گروه تلگرامی درس مطرح کنید. در انتهای هر پیام بنده (محمد ربیعی) را منشن کنید. سوالات در گروه پرسیده شده و همان‌جا پاسخ داده خواهند شد تا در دسترس همه‌ی دانشجوها قرار بگیرند.
- بارم‌بندی سوالات پروژه به صورت زیر می‌باشد:

نام سوال	درصد نمره
ضریب اویلر-ماسکرونی	۸
مساله‌ی نیوتن-پیپس	۷
تخمین عدد نپرین با روش مونته کارلو	۱۷
سری فیبوناچی تصادفی	۳۵
قاعده‌ی دایره‌ای برای مقادیر ویژه	۸
داده بازی	۲۵