



## تمرین کامپیوتری سوم

سیستم‌های عامل - بهار 1400

طراحی چندریسه‌ای

دانشکده مهندسی برق و کامپیوتر

مسئولان تمرین:

مهلت تحویل:

استاد:

میینا شاه‌بنده، محمدصابر ابراهیم نژاد

تا ساعت 23:55 روز یکشنبه 9 خرداد

دکتر مهدی کارگهی

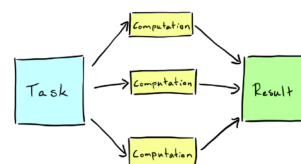
### مقدمه

هدف از این تمرین آشنایی شما با مفاهیم اولیه طراحی چندریسه‌ای<sup>1</sup> یک مسئله است. در این

تمرین شما به تحلیل داده‌هایی که از مشخصات و قیمت فروش خانه‌ها جمع‌آوری شده‌است

می‌پردازید. پیشنهاد می‌شود در ابتدا به مطالعه پرونده<sup>2</sup> مربوط به پیش‌زمینه که در کنار این پرونده بارگذاری شده‌است پرداخته و سپس

شرح تمرین را مطالعه فرمایید.



### شرح تمرین

در این تمرین شما به تحلیل داده‌هایی که از مشخصات و قیمت فروش خانه‌ها جمع‌آوری شده‌است

می‌پردازید. در ابتدا برنامه شما اقدام به خواندن و تجزیه مجموعه داده<sup>3</sup>ی ارائه شده کرده و آنها را در

حافظه خود ذخیره می‌کند. پس از استخراج داده‌ها و ویژگی‌های آنها، برنامه اقدام به برچسب گذاری

ستون قیمت داده‌ها، بدست آوردن میانگین و انحراف معیار داده‌ها و در نهایت اقدام به تعیین رده قیمتی خانه‌ها می‌کند. در این



<sup>1</sup> Muti-Threaded Design

<sup>2</sup> File

<sup>3</sup> Dataset

تمرین شما به دو روش این مسئله را پیاده‌سازی می‌کنید. همچنین در قالب گزارش کاری که در کنار این پرونده بارگذاری شده‌است، به مقایسه این روش‌ها و بررسی پیاده‌سازی‌های انجام‌شده می‌پردازید.

## استخراج ویژگی‌ها

در مجموعه داده تهیه دیده‌شده ستون SalePrice ستون هدف<sup>4</sup> و سایر ستون‌ها، معرف ویژگی‌های هر نمونه هستند. در ابتدا به استخراج ویژگی‌های مربوط به هر نمونه کرده و آن را در حافظه برنامه ذخیره کنید.

## برچسب گذاری

ستون قیمت در واحد دلار بیان شده است. در این تمرین می‌خواهیم رده قیمتی خانه‌ها را بدست آوریم که شامل دو رده ی ارزان و گران است (این دو رده نسبی هستند و معیار واقعی نیستند). به این منظور ستون SalePrice خانه‌ها را به این شکل تغییر دهید که اگر از آستانه قیمت کمتر بود (این آستانه قیمت به عنوان ورودی به برنامه شما داده می‌شود) تبدیل به عدد 0 شود (دسته ی 0 که نماینده «ارزان» است) و اگر بیشتر یا مساوی آستانه قیمت تعیین شده بود تبدیل به عدد 1 شود (دسته ی 1 که نماینده «گران» است). نتیجه را در حافظه برنامه در همین ستون SalePrice ذخیره کنید.

## بدست آوردن میانگین و انحراف معیار

در این مرحله می‌خواهیم میانگین و انحراف معیار ویژگی‌هایی از خانه‌های ارزان و گران را بدست آوریم؛ زیرا از این پارامترها در تعیین رده قیمتی خانه‌ها استفاده خواهیم کرد. پس برای هر ستون ویژگی، میانگین داده‌های آن و سپس انحراف معیار داده‌های آن را هم برای خانه‌های گران و هم برای خانه‌های ارزان بدست آورده و در حافظه برنامه ذخیره کنید. در واقع پس از محاسبات برای هر ستون ویژگی، باید مقدار میانگین و انحراف معیار آن را برای رده قیمتی 0 و رده قیمتی 1 در حافظه برنامه داشته باشید.

## تعیین رده قیمتی خانه‌ها

برای هر سطر از داده (هر خانه)، می‌خواهیم با استفاده از داده‌های مرحله قبل (میانگین و انحراف معیار) رده قیمتی آن را حدس بزنیم. به این منظور از ستون ویژگی GrLivArea استفاده می‌کنیم (علت انتخاب این ویژگی، هم بستگی بالای آن با ستون هدف است که این مباحث خارج از این پروژه است). نحوه حدس به این صورت است که اگر مقدار این ویژگی خانه ی فعلی (هر سطر از

---

<sup>4</sup> ستونی که قصد طبقه‌بندی آن وجود دارد: Target

داده) در بازه ی حدس رده قیمتی 1 قرار گرفت، رده قیمتی آن را 1 تعیین کرده و در غیر این صورت 0. این حدس ها را در حافظه ی برنامه ذخیره کنید. بازه حدس رده قیمتی 1 برای هر ویژگی به شکل زیر بدست می آید:

$$(\mu_1 - \sigma_1, \mu_1 + \sigma_1)$$

که در آن  $\mu_1$  میانگین ویژگی مورد نظر برای خانه هایی است که ستون SalePrice آنها 1 (گران) و  $\sigma_1$  نیز انحراف معیار آنهاست.

## محاسبه دقت

برای هر نمونه، رده قیمتی تعیین شده در قسمت قبل را با مقدار واقعی آن که در ستون SalePrice قرار دارد مقایسه می کنید. پس از اتمام مقایسه برای تمام نمونه ها، دقت مدل از رابطه زیر به دست آورید:

$$Accuracy = \frac{\text{number of samples classified correctly}}{\text{number of all samples}}$$

در انتها باید مقدار این دقت را تا دو رقم اعشار گزارش کنید (در صورتی که مقدار دقت بدست آمده به 0 منتهی می شد، باید 0 را هم در پاسخ بیاورید؛ به عنوان مثال 79.8 باید به صورت 79.80 گزارش شود). شکل دقیق این گزارش در بخش ورودی و خروجی برنامه شرح داده شده است.

## پیاده سازی سری<sup>5</sup>

در این بخش از تمرین شما به پیاده سازی سری برنامه خواسته شده می پردازید. سعی کنید در این بخش از تمرین بهترین پیاده سازی که می توانید را از لحاظ زمان اجرا انجام دهید. پس از انجام این بخش از تمرین به کامل کردن بخش مربوطه در گزارش کار اقدام کنید.

## پیاده سازی چندریسه ای

در این بخش از تمرین به موازی سازی اعمال صورت گرفته در توابعی که در بخش قبل به عنوان Hotspot<sup>6</sup> از آن ها یاد کردید می پردازید. تجزیه کردن پرونده ها و ذخیره سازی آنها در حافظه از اعمال زمان گیر در بسیاری از برنامه ها است که احتمالاً از توابع مربوط به آنها (درکنار سایر توابع) به عنوان Hotspot های برنامه یاد کرده اید. برای موازی سازی این بخش می توانید مجموعه داده ای را که در اختیارتان قرار گرفته است به پرونده هایی کوچکتر تقسیم کرده و اعمال مربوطه را توسط چندین ریسه انجام دهید.

<sup>5</sup> Serial

<sup>6</sup> توابعی که در برنامه تان بیشترین زمان اجرا را به خود اختصاص می دهند.

- **دقت کنید** که مجاز به تغییر ساختار مجموعه داده، از قبیل تغییر ستون‌های مربوطه نیستید و تنها می‌توانید نمونه‌های موجود در هر پرونده را بین پرونده‌های کوچکتر تقسیم کنید.

- **دقت شود** که در صورتی که تعداد ریشه‌هایی که برای این منظور در نظر می‌گیرید برابر با  $n$  باشد، اسامی مجموعه‌های داده کوچک شده باید به صورت (dataset\_[0, n - 1].csv) باشد و استخراج هر پرونده را توسط یک ریشه انجام دهید.

○ برای مثال ۲ ریشه، این مجموعه‌ها بصورت dataset\_0.csv و dataset\_1.csv هستند.

○ همچنین **دقت شود** که در برنامه خود تصور کنید که در هر پرونده شکسته شده، سطر اول نام مربوط به ستون‌ها آورده شده است.

- **توجه شود** که این بخش از تمرین باید به صورت چندریشه‌ای پیاده‌سازی گردد و سایر پیاده‌سازی‌ها قابل قبول نیست. پس از انجام این بخش از تمرین به کامل کردن بخش مربوطه در گزارش کار اقدام کنید.

## ورودی و خروجی برنامه

نمونه اجرای برنامه با فرض اینکه پوشه مربوط به مجموعه‌های داده در کنار پرونده اجرایی شما قرار گرفته است در زیر آمده است که در آن datasets نام پوشه مجموعه داده مذکور و 330000 آستانه قیمت دسته بندی ارزان/گران است.

نمونه اجرا
./HousePricePrediction.out datasets/ 330000

قالب و نمونه خروجی این اجرای برنامه در زیر آمده است:

قالب خروجی
Accuracy: <accuracy_percentage>%
نمونه خروجی
Accuracy: 82.05%

توجه: نام فایل خروجی برنامه و متن قالب خروجی را از اینجا کپی نکنید و آن را تایپ کنید!

## نکات تکمیلی

- تمام خروجی‌های برنامه را در جریان خروجی استاندارد<sup>7</sup> چاپ کنید.
- تضمین می‌شود که ورودی‌هایی که به برنامه شما داده می‌شود صحیح هستند و نیازی به بررسی صحت ورودی توسط برنامه شما نیست.
- طراحی درست، کارایی<sup>8</sup> برنامه و شکستن برنامه به بخش‌های کوچکتر تأثیر زیادی در نمره‌ی تمرین دارد.
- دقت شود برای موازی‌سازی پروژه تنها مجاز به استفاده از کتابخانه **PThread** هستید و استفاده از کتابخانه‌های دیگر (بجز کتابخانه‌های پایه زبان C++) مجاز نیست.

## نحوه‌ی تحویل

- **دقت کنید** که فایل آپلودی شما با نام OS\_CA3\_<SID>.zip حتماً باید شامل دو پوشه<sup>9</sup> مجزا باشد که در یک پوشه پیاده‌سازی سری و در پوشه دیگر پیاده‌سازی موازی آورده شده است. در کنار پوشه‌ها گزارش کار خود را در قالب pdf قرار دهید. دقت کنید که فایل zip شما شامل فولدر بیرونی نباشد و مستقیماً پس از unzip کردن آن، دو پوشه پیاده‌سازی سریال و موازی و گزارش کار شما بدست آید.
- برای مثال، نمونه فایل مورد قبول در زیر آمده است:

OS\_CA3\_810198999.zip

```
|— parallel
|   |— main.cpp
|   |— makefile
|— report.pdf
|— serial
    |— main.cpp
    |— makefile
```

---

<sup>7</sup> Standard Output Stream

<sup>8</sup> Performance

<sup>9</sup> Directory

- **حتماً** در قسمت متن آنلاین محل تحویل، تعداد ریسسه‌هایی که برای انجام عملیات‌های مربوط به خواندن فایل‌ها استفاده کردید را مشخص کنید.
- برنامه شما باید در سیستم عامل لینوکس و با مترجم g++ با استاندارد c++11 ترجمه و در زمان معقول برای ورودی‌های آزمون اجرا شود.
- **دقت کنید** که پروژه شما باید دارای Makefile باشد. همچنین در Makefile خود مشخص کنید که از استاندارد c++11 استفاده می‌کنید.
- نام فایل اجرایی شما که در کنار Makefile خود ساخته می‌شود باید HousePricePrediction.out باشد.
- نیازی به قرار دادن فایل‌های مجموعه داده نیست، زیرا کدهای شما با مجموعه داده‌ی متفاوتی تست می‌شوند که تعداد سطرهای آن می‌تواند متفاوت باشد اما ستون‌های آن مشابه همان مجموعه داده‌ی ارائه شده است.
- **دقت کنید** که برنامه شما توسط آزمون‌های خودکار سنجیده می‌شود. به همین منظور به قالب‌های ذکر شده برای مجموعه‌های داده و فایل آپلودی دقت کافی را داشته باشید.
- نکته‌هایی که در جلسه توجیهی تمرین گفته می‌شود و یا در فروم‌های مربوطه مطرح می‌شود بخشی از تمرین هستند؛ بنابراین به آن‌ها توجه داشته باشید.
- هدف این تمرین یادگیری شماست. لطفاً تمرین را خودتان انجام دهید. در صورت کشف تقلب مطابق قوانین درس با آن برخورد خواهد شد.