


دانشکده‌ی علوم ریاضی

تحلیل رگرسیون

تمرین : سری ۷

مدرّس: دکتر محمد صفدری

مهلت تحویل دوم خردادماه

- پاسخ‌های خود را در قالب یک فایل فشرده در درس‌افزار شریف ارسال کنید.
- در پوشه‌ی پاسخ لازم است به ازای هر سوال یک پوشه ایجاد کنید و کدهای مربوط به آن سوال را در آن قرار دهید.
- در پوشه‌ی مربوط به هر سوال، پوشه‌ای حاوی نمودارهای و توضیحات تکمیلی (در صورت لزوم) بگنجانید.
- پیش از اجرای هر بخش seed را برابر ۱۰۰ قرار دهید.
- سوالات خود پیرامون تمرینات را با دستیاران درس مطرح کنید.

تمرینات نظری

از فصل ۵ کتاب درس، تمرینات ۱ و ۴ را حل کنید.

شبه بوت‌استرپ

هر آماره یک تابع قطعی از بردار داده است. فرض کنید بردار $x \in \mathbb{R}$ داده است و آماره‌ای با مقدار حقیقی از این داده (مثلا میانگین) مورد توجه ماست. تابع f تعمیمی از آماره‌ی ماست و به این صورت تعریف شده که برداری n مولفه‌ای که هر مولفه نشانگر تکرر هرکدام از داده‌هاست را ورودی می‌گیرد و آماره را روی مجموعه داده‌ای متناظر با این تکررها محاسبه می‌کند. فرض می‌کنیم جمع این تکررها (نرم $\mathcal{L}_1$ بردار تکرر) برابر n هست و آماره صرفا برای مجموعه داده‌های nتایی مورد توجه است. بنابراین داریم:

$$f : S \to \mathbb{R}$$

که

$$S = \{p = (p_1, p_2, ..., p_n) \in \mathbb{R}^n : p_i \geq 0, \|p\|_1 = n, p_i \in \mathbb{Z} \quad \forall i\}.$$

برای مثال به ازای $p = (1, 1, ..., 1)$ مقدار $f(p)$ همان مقدار آماره برای مجموعه داده‌ای است که در دست داریم.

الف) متغیر تصادفی P در S مقدار می‌گیرد و توزیع یکنواخت دارد. روشی برای نمونه‌گیری از P معرفی کنید.

ب) فرض کنید K نمونه از P گرفته‌ایم و $f(P)$ را محاسبه کرده‌ایم. نشان دهید واریانس نمونه‌ای این مقادیر برابر با تخمین بوت‌استرپ با K تکرار از واریانس آماره‌ی اولیه است.

بردار تکرر را به بردار وزن تعمیم می‌دهیم به این صورت که

$$\tilde{S} = \{p = (p_1, p_2, ..., p_n) \in \mathbb{R}^n : p_i \geq 0, \|p\|_1 = n, p_i \in \mathbb{R} \quad \forall i\}.$$

و فرض می‌کنیم تابع تعمیم یافته‌ی $\tilde{f} : \tilde{S} \to \mathbb{R}$ سازگار با f تعریف شده است، به عبارت دیگر در مجموعه‌ی $S \subset \tilde{S}$ مقادیر یکسان با f اتخاذ می‌کند. برای برخی آماره‌ها این تعمیم می‌تواند معنی دار هم باشد، مثلا برای میانگین می‌توان $\tilde{f}(p) = p.x$ را معرفی کرد که برای حالت تکرر کماکان معنی میانگین داده‌ای با تکرر p را دارد و برای نقاطی که معنای تکرر نمی‌دهند یک میانگین وزندار است. کماکان میانگین کل داده مقدار f در نقطه‌ی $p = (1, 1, ..., 1)$ خواهد بود.

ج) متغیر تصادفی $\tilde{P}$ در $\tilde{S}$ مقدار می‌گیرد و تابع چگالی یکنواخت دارد. روشی برای نمونه‌گیری از $\tilde{P}$ معرفی کنید. (نابدیهی، لغزنده)

ج) نشان دهید چنانچه $\tilde{f}$ تابعی خطی باشد، آنگاه

$$\mathbb{E}[f(p)] = \mathbb{E}[\tilde{f}(p)]$$

تمرینات عملی

در این بخش قصد داریم در یک مثال ساده هر دو روش بازنمونه‌گیری را پیاده‌سازی کنیم.

- مجموعه داده‌ی ۱۰۰ تایی با نمونه‌گیری از توزیع نرمال استاندارد بسازید و آن را A بنامید.
- مجموعه داده‌ی ۱۰۰ تایی با نمونه‌گیری از توزیع کوشی استاندارد بسازید و آن را B بنامید.

اطلاعات مرتبط با توزیع کوشی (cauchy) را می‌توانید با جستجو در اینترنت پیدا کنید. با توجه به توابع چگالی توزیع‌های کوشی و نرمال استاندارد، مشاهده می‌کنیم که مرکز (محور تقارن، میانه) برای هردوی آنها در صفر واقع است. قصد داریم عملکرد آماره‌ی میانگین و میانه‌ی نمونه‌ای را برای تخمین مرکز دو توزیع کوشی و نرمال استاندارد (که هر دو صفر هستند) ارزیابی کنیم. به طور مشخص واریانس این تخمینگرها مورد توجه است.

کراس ولیدیشن

الف) تابعی بنویسید که یک مجموعه داده و یک تخمین‌گر (تابعی که یک زیرمجموعه از داده را ورودی می‌گیرد و عددی حقیقی برمی‌گرداند) به عنوان ورودی بگیرد و با روش k-fold کراس ولیدیشن واریانس تخمین‌گر را روی این مجموعه داده تخمین بزند.

ب) عملکرد تخمین‌گرهای میانگین و میانه را روی دو مجموعه‌داده‌ی A و B ارزیابی کنید. پاسخ خود را قدری شرح دهید، با عنایت به اینکه برای توزیع کوشی میانگین تعریف نمی‌شود!

ج) به ازای مقادیر k از ۵ تا ۱۰۰ واریانس تخمین‌گر میانگین روی مجموعه‌داده‌ی A را با استفاده از کراس ولیدیشن بدست آورید. مقدار واقعی واریانس تخمین‌گر میانگین را به طور تحلیلی بدست آورید. به ازای کدام یک از مقادیر k فاصله‌ی مقدار تخمین زده شده به مقدار واقعی آن نزدیک‌تر است؟ آیا شما بهترین k برای k-fold کراس ولیدیشن را پیدا کرده‌اید؟ توضیح دهید.

بوت استرپ

الف) تابعی بنویسید که یک مجموعه داده و یک تخمین‌گر (تابعی که یک زیرمجموعه از داده را ورودی می‌گیرد و عددی حقیقی برمی‌گرداند) به عنوان ورودی بگیرد و با روش بوت‌استرپ و تکرار ۲۰۰ بار، واریانس تخمین‌گر را روی این مجموعه داده تخمین بزند.

ب) عملکرد تخمین‌گرهای میانگین و میانه را روی دو مجموعه‌داده‌ی A و B ارزیابی کنید. پاسخ خود را قدری شرح دهید.