بسمه تعالی

یادگیری ماشین

آذر ۱۴۰۰

سلام بر تمام دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی 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.

۳. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید. شکل ها به طور واضح و در فرمت درست گزارش شوند.

۴. عکس‌ها را به صورت واضح و همراه با زیرنویس در گزارش خود بیاورید.

۵. از بین سوالات شبیه سازی پاسخ دادن به حداقل یکی از سوالات ۸ یا ۹ الزامی است و از بین سوالات تئوری پاسخ دادن به حداقل یکی از سوالات ۳ یا ۵ الزامی است. حداکثر تا نمره ۱۱۰ (۱۰ نمره امتیازی) لحاظ خواهد شد.

۶. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل نمره تمرین صفر می‌شود.

۷. برای هر کد که در فایل نهایی ضمیمه می‌کنید، گزارش بنویسید. کدهای ضمیمه شده بدون گزارش مربوطه نمره‌ای نخواهند داشت. (این گزارش‌ها تنها معیار تفکیک کد شما و کدهای موجود در منابع مختلف مانند اینترنت خواهند بود.)

۷. در صورت داشتن سوال، از طریق ایمیل mesbahamirhossein@gmail.com، سوال خود را مطرح کنید.

- سوال اول (۱۵ نمره)

به سوالات زیر پاسخ دهید.

- مفهوم bias-variance trade off را با توجه به اندازه h_n در روش پارزن k_n در روش knn توضیح دهید.

- تفاوت روش های پارامتریک و نان پارامتریک را توضیح دهید.

- مشکلات روش های kernel based چیست.

- تفاوت مفهوم حجم در روش پارزن و knn را بررسی کنید.

۲- سوال دوم (۱۰ نمره)

دیاگرام voronoi حاصل از الگوریتم نزدیک ترین همسایه را در نظر بگیرید. ثابت کنید cell های حاصل از این دیاگرام محدب هستند.

۳- سوال سوم (۲۰ نمره)

توزیع نرمال $p(x)\sim N(\mu.\sigma^2)$ و تابع پنجره پارزن $\varphi(x)\sim N(0.1)$ را در نظر بگیرید. نشان دهید که تخمین پنجره پارزن

$$P(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

برای h_nهای کوچیک دارای ویژگی‌های زیر است:

- $\tilde{p}_n(x)\sim N(\mu.h_n{}^2 + \sigma^2)$
- $p_n(x) - \tilde{p}_n(x) \cong \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2\left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right]p(x)$
- $var[p_n(x)] \cong \frac{1}{2nh_n\sqrt{\pi}}p(x)$

۴ - سوال چهارم (۱۰ نمره)

متریک فاصله اقلیدسی را در d بعد در نظر بگیرید:

$$D(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{k=1}^{d} (a_k - b_k)^2}.$$

فرض کنید عناصر هر بعد را در یک مقدار حقیقی غیر صفر ضرب میکنیم. یعنی k = 1, 2, ..., d داریم:

$$x_k = \acute{}\alpha_k x_k$$

ثابت کنید پس از ضرب نیز این متریک فاصله همچنان یک استاندارد است. در مورد تاثیر این امر بر طبقه بند knn بحث کنید.

۵ - سوال پنجم (۲۰ نمره)

یک مسئله طبقه بندی با روش knn را در نظر بگیرید. مجموعه داده دو کلاسه D را نیز به صورت $D= \{x^q, \omega_i{}^q\}$, q=1,..., Q داریم. این داده ها نتایج یک نظر سنجی بوده و دیتاپوینت ها به صورت پرچسب خورده، مستقل از هم هستند و فرض می کنیم تعداد داده های دو کلاس یکسان است. برای هر سمپل تست نزدیک ترین k دیتاپوینت را به صورت $\{x_i\}$ i=1, ..., k نمایش می دهیم. هر دوی p(x|1) و p(x|2) توزیع یکنواخت بر روی یک کره به شعاع واحد دارند و مرکز دو ابر کره نیز از هم ۱۰ واحد فاصله دارند.

الف) نشان دهید اگر k فرد باشد متوسط احتمال خطا از رابطه زیر به دست می آید:

$$p_Q(e) = \frac{1}{2^Q}\sum_{j=0}^{\frac{k-1}{2}}\binom{Q}{j}$$

ب) با توجه به بخش قبل نشان دهید که در این حالت خطای طبقه بند نزدیک ترین همسایه کمتر از حالت $k \geq 2$ است و دلیل مشاهده این موضوع را توضیح دهید.

ج) نشان دهید: $\lim_{Q\to\infty} p_Q(e) = 0$

۶- سوال ششم (۱۵ نمره) – پیاده سازی

موارد خواسته شده در قسمت های مختلف را انجام داده و نتایج به دست آمده را تحلیل کنید.

الف) ۵ دیتاپوینت رندوم ۵ بعدی را تولید کرده و فاصله این ۵ دیتا پوینت را از هم حساب کرده ونمودار هیستوگرام فاصله ها را رسم نمایید. این کار را برای ابعاد ۱۰، ۱۰۰، ۵۰۰، ۱۰۰۰، ۱۰۰۰۰ و ۱۰۰۰۰۰۰ تکرار کنید.

ب) برای ۵ دیتا پوینت رندوم ۱۰۰۰۰ بعدی فاصله دیتاپوینت ها را از هم حساب کرده و نمودار هیستوگرام فاصله ها را رسم کنید. این کار را برای دیتاپوینت ها به تعداد ۱۰، ۱۰۰، ۵۰۰۰ و ۱۰۰۰۰ را تکرار کنید.

ج) نتایج تحلیل خود را بیان کنید. با توجه به این نتایج عملکرد الگوریتم knn را چگونه ارزیابی میکنید. همچنین ارتباط این نتایج با curse of dimensionality را بیان کنید.

۷ – سوال هفتم (۲۰ نمره) – پیاده سازی

دیتاست اعداد دست نویس هدی را از این لینک دریافت کنید.

الف) ابتدا از هر کلاس ۳ داده را نمایش دهید.

ب) الگوریتم knn برای طبقه بندی را بدون استفاده از هیچ گونه پکیج آماده پیاده سازی کنید.

ج) دیتا را نرمالایز کرده و مجدد الگوریتم knn را اعمال کنید.نتایج این مرحله را با مرحله ب مقایسه کنید. همچنین روش نرمالیزشن مورد استفاده خود را توضیح دهید.

د) مراحل ب و ج را با استفاده از پکیج آماده scikit-learn تکرار کرده و نتایج خود را مقایسه کنید.

توجه کنید که معیار فاصله را میتوانید به صورت دلخواه انتخاب کنید. و این معیار باید برای همه مراحل یکسان باشد همچنین در نتایج خود confusion matrix و F1 score را گزارش کنید.

۸ – سوال هشتم (۲۰ نمره) – پیاده سازی

در این سوال میخواهیم به پیاده سازی روش تخمین نان پارامتری پارزن بپردازیم. لازم به ذکر است که الگوریتم خواسته شده در این سوال را باید بدون استفاده از کتابخانه های آماده موجود پیاده سازی کنید.

برای شروع ابتدا دیتاست ted talks را دانلود کنید.

الف) ستون duration این دیتاست را استخراج کرده و توزیع دیتای این ستون را با استفاده از روش پنجره پارزن با کرنل گوسی به دست آورده و نتیجه را نمایش دهید. اندازه پنجره را برابر با ۱۰ در نظر بگیرید.

ب) تاثیر اندازه پنجره را با ۳ مقدار مختلف بررسی کنید.

ج) با استفاده از کتابخانه seaborn توزیع ستون duration را رسم کنید. با افزایش مقدار n روند تغییر و همگرا شدن به توزیع اصلی را روی یک نمودار نشان دهید.

د) نتیجه قسمت الف را با نتیجه توابع کتابخانه های آماده مقایسه کنید.

۹ – سوال نهم (۲۰ نمره) – پیاده سازی

در این سوال میخواهیم به پیاده سازی طبقه بند بیز به کمک روش های تخمین توزیع non-parametric بپردازیم.

توجه شود که در این سوال مجاز به استفاده از توابع پکیج های آماده نیستید. برای نتایج ماتریس آشفتگی و f1-score و accuracy را گزارش کنید.

از دیتاست اعداد دست نویس هدی برای این سوال استفاده کنید.

الف) یک طبقه بند naive bayes با استفاده از روش تخمین پنجره پارزن برای تخمین pdf طراحی کرده و نتایج خواسته شده را گزارش کنید. اندازه پنجره را برابر با ۱۰ در نظر بگیرید.

ب) برای روش تخمین پنجره از کرنل های cosine و linear استفاده کرده و نتیجه را با قسمت قبل مقایسه کنید.

ج) برای بهترین کرنلی که از قسمت های قبل به دست آورده اید نتایج را برای ۳ اندازه پنجره مختلف مقایسه کنید.

د) یک طبقه بند naive bayes با استفاده از روش knn برای تخمین pdf طراحی کرده و نتایج خواسته شده را برای ۳ مقدار مختلف k گزارش کنید.

ه) با افزایش مقدار n برای روش پارزن با کرنل گوسی و افزایش مقدار k برای روش knn همگرایی نتایج دو روش به یکدیگر را بررسی کنید.