بسمه تعالی

یادگیری ماشین

فروردین ۱۴۰۱

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.

۴. شکل ها به طور واضح و در فرمت درست گزارش شوند. عکس‌ها را به صورت واضح و همراه با زیرنویس در گزارش خود بیاورید.

۵. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل نمره تمرین صفر می‌شود.

۶. 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.

۱- سوال اول (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

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

برای h_nهای کوچک دارای ویژگی‌های زیر است:

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

۲- سوال دوم (۱۰ نمره)

متریک فاصله اقلیدسی را در d بعد در نظر بگیرید:

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

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 فاصله استاندارد است یعنی ویژگی‌های گفته شده برای یک فاصله استاندارد را دارا می‌باشد. در مورد تاثیر این امر بر طبقه بند knn بحث کنید.

۳- سوال سوم (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 می‌آید:

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

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

۴- سوال چهارم (۱۵ نمره)

بخش اول – ۵ 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.

بخش دوم (پیاده سازی) – ۱۰ نمره:

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

الف) ۱۰۰۰ دیتاپوینت رندوم ۵ بعدی را تولید کرده و فاصله این ۱۰۰۰ دیتا پوینت را از هم حساب کرده و نمودار هیستوگرام فاصله‌ها را رسم نمایید. این کار را برای ابعاد ۲۰۰۰، ۵۰۰۰، ۱۰۰۰۰ و ۱۰۰۰۰۰۰ تکرار کنید.

ب) برای ۱۰۰۰ دیتا پوینت رندوم ۱۰۰۰۰ بعدی فاصله دیتاپوینت ها را از هم حساب کرده و نمودار هیستوگرام فاصله‌ها را رسم کنید. این کار را برای دیتاپوینت ها به تعداد ۲۰۰۰، ۵۰۰۰، ۱۰۰۰۰ و ۱۰۰۰۰۰۰ 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.

۵ – سوال پنجم (۱۵ 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 ۲۰، ۵۰ و ۱۰۰ مختلف بررسی کنید.

ج) با استفاده از کتابخانه‌های آماده توزیع ستون duration را رسم کنید. با افزایش مقدار n روند تغییر و همگرا شدن به توزیع اصلی را روی یک نمودار نشان دهید. مقدار n را در بازه ۲۵۰ نمونه تا کل دیتا با step برابر با ۲۵۰ برسی کرده و همگرایی برای nهای مختلف را حتما روی یک نمودار نشان دهید.

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

۶– سوال ششم (۲۰ نمره) – پیاده سازی

در این تمرین از دیتاست fashion MNIST استفاده خواهیم کرد. توصیه میشود با توجه به پردازش سنگین مورد نیاز این بخش از google colab برای این بخش استفاده کنید. ابتدا داده‌ها را به ۳ بخش train، validation و test جدا کنید. سپس یک مدل mlp برای طبقه‌بندی داده‌ها طراحی کنید.

الف) نمودار دقت و خطا (loss) را برای داده های train و validation و همچنین این مقادیر را برای داده‌های تست گزارش کنید.

ب) تاثیر تعداد لایه های مخفی و نرخ یادگیری و solver های مختلف را روی عملکرد مدل بررسی کنید. برای پرامترهای مختلف میتوانید گزینه‌های زیر را در نظر بگیرید.

- تعداد لایه‌های مخفی ۱، ۲، ۳ و ۴

- اندازه‌ی لایه‌های مخفی ۵۰، ۱۰۰، ۱۵۰ نورون

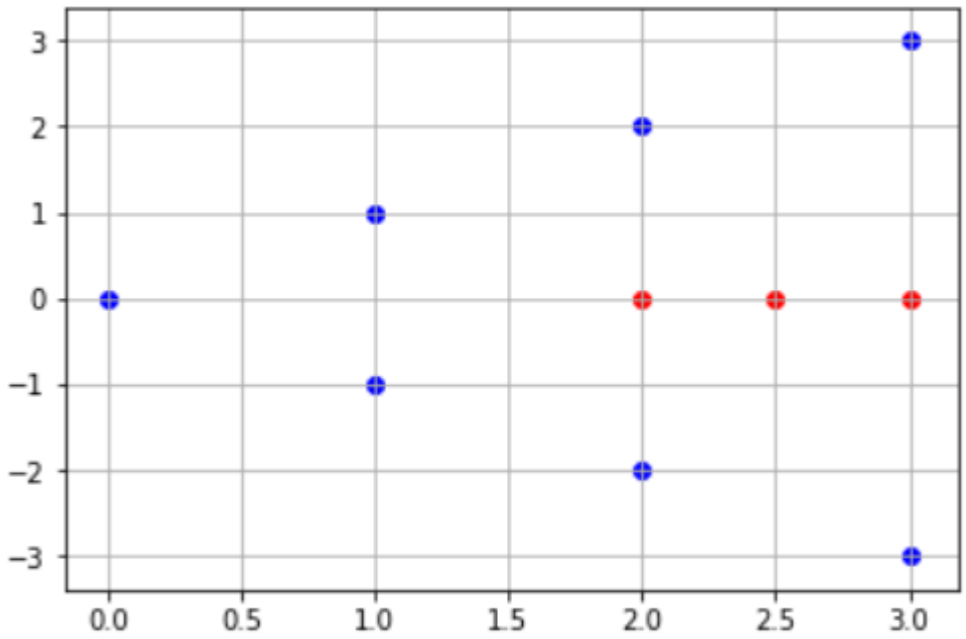
- solverهای sgd، ADAM و rmsprop

- نرخ یادگیری ۰,۰۰۱، ۰,۰۱، ۰,۱، ۰,۵ و ۰,۹

ج) بهترین پارامترهایی که برای این مدل به دست آورده‌اید را در یک جدول گزارش کنید.

۷ – سوال هفتم (۱۵ نمره)

تصویر زیر را در نظر بگیرید.



شکل ۱

برای جداسازی داده‌های دو کلاس یک شبکه پرسپترون طراحی کرده و پیاده سازی کنید. توجه داشته باشید که در این سوال مجاز به استفاده از کتابخانه و توابع آماده نمی‌باشید.

۸ – سوال هشتم (۲۰ نمره)

یک رستوران بر این است که بررسی نماید با توجه به عوامل موثر، افرادی که به رستوران مراجعه می‌کنند در صورتی که تمام میزها پر باشد، برای خالی شدن میز صبر می‌کنند یا نه؟

داده‌های ثبت شده از ۱۲ مراجعه کننده، جنبه‌های مختلف و اینکه صبر می‌کنند یانه را در جدول ۱ مشاهده می‌فرمایید.

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

جدول ۱ – داده‌های ثبت شده از ۱۲ مراجعه کننده

توضیح فیچرهای مختلف نیز به شرح زیر است:

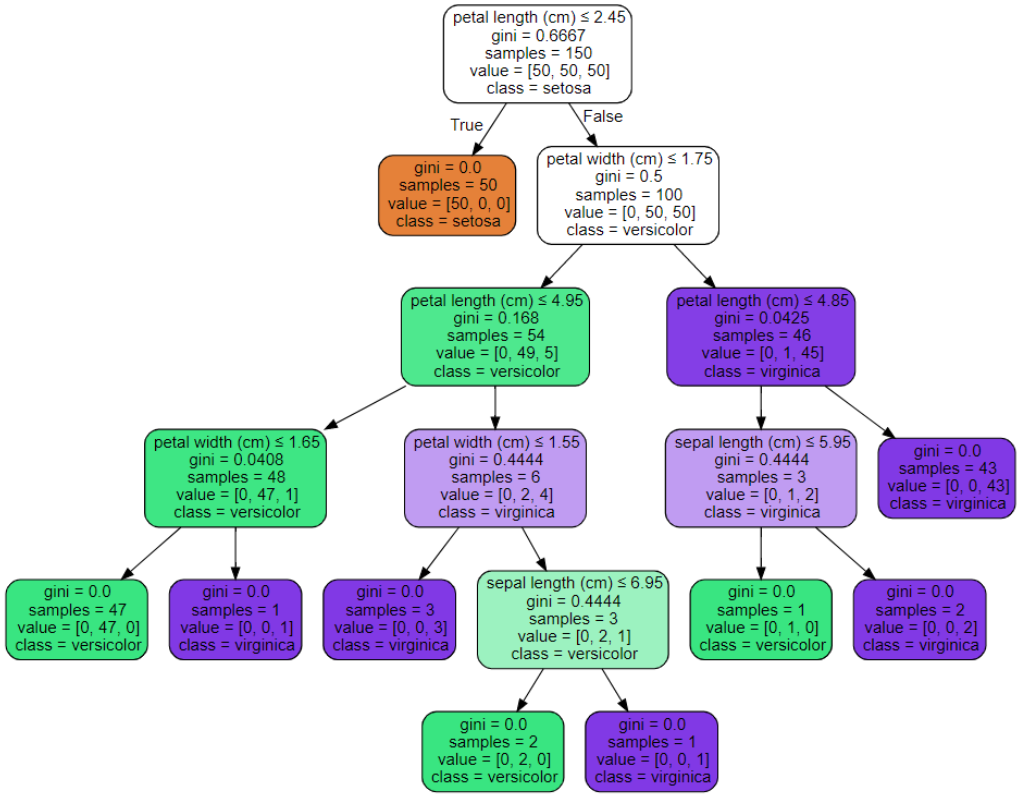
1.	Alternate: whether there is a suitable alternative restaurant nearby.
2.	Bar: whether the restaurant has a comfortable bar area to wait in.
3.	Fri/Sat: true on Fridays and Saturdays.
4.	Hungry: whether we are hungry.
5.	Patrons: how many people are in the restaurant (values are None, Some, and Full).
6.	Price: the restaurant's price range ($, $$, $$$).
7.	Raining: whether it is raining outside.
8.	Reservation: whether we made a reservation.
9.	Type: the kind of restaurant (French, Italian, Thai or Burger).
10.	WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

- مرحله (یا لایه) اول درخت تصمیم را با استفاده از معیار آنتروپی به صورت دستی حل کنید.

- طبقه بند درخت تصمیم را بدون استفاده از پکیج آماده و با در نظر گرفتن معیار آنتروپی کرده پیاده سازی کرده و نتایج پیاده‌سازی را گزارش کنید.

- طبقه بند درخت تصمیم را با استفاده از پکیج‌های آماده و با در نظر گرفتن معیار آنتروپی پیاده سازی کرده و آن را رسم کنید. شکل ۲ نشان‌دهنده نمونه از تصویر رسم شده برای درخت تصمیم می‌باشد.

- تفسیر خود از این درخت تصمیم را شرح دهید.



شکل ۲ – درخت تصمیم رسم شده توسط پکیج scikit learn