دانشکده مهندسی برق و کامپیوتر

درس

یادگیری ماشین

تمرین تئوری دوم

نیمسال دوم سال تحصیلی ۱۴۰۰ - ۱۴۰۱

سوالات

۱- مثال های آموزشی نشان داده شده در جدول ۱را برای یک طبقه بندی باینری در نظر بگیرید. (۳۵ نمره)

الف) شاخص جینی (Gini index) را برای مجموعه کلی نمونه های آموزشی محاسبه کنید.

ب) میانگین وزنی شاخص جینی تقسیم (weighted average Gini Index of splitting) را در حالت multiway برای ویژگی customer id بدست آورید.

پ) شاخص جینی پراکندگی (gini index of diversity) برای ویژگی gender محاسبه شود.

ت) شاخص جینی پراکندگی (gini index of diversity) در حالت چند راهه برای ویژگی Car type را بدست آوردید.

ث) شاخص جینی پراکندگی (gini index of diversity) در حالت چند راهه برای ویژگی Shirt size محاسبه شود.

ج) کدام ویژگی بهتر است، جنسیت، نوع ماشین یا اندازه پیراهن؟

چ) توضیح دهید که چرا شناسه مشتری نباید به عنوان شرط تست ویژگی استفاده شود حتی اگر کمترین جینی را داشته باشد.

جدول ۱

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

۲- ثابت کنید تابع softmax نسبت به اضافه شدن مقدار ثابت به ورودی، حساس نیست. به عبارت دیگر تساوی زیر برقرار است: (x + c به معنای افزودن مقدار ثابت c به تمام ابعاد x می‌باشد)

$$softmax(x) = softmax(x + c)$$

برای تابع softmax داریم: (۱۵ نمره)

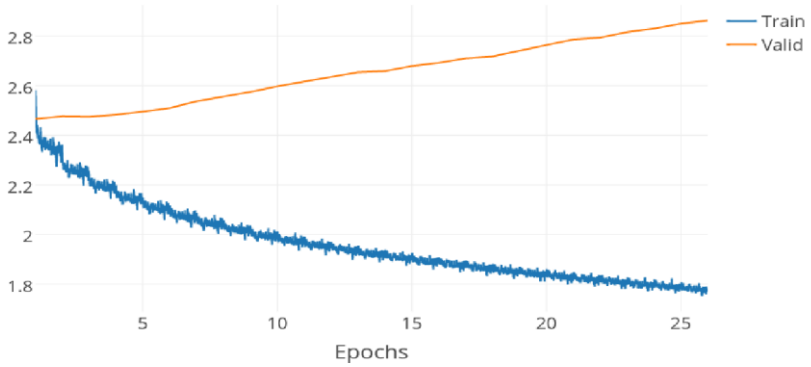
$$softmax(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

۳- فرض کنید یک مدل رگرسیون لاجستیک برای تشخیص بیماری سرطان طراحی کرده‌اید و پس از آموزش شبکه، منحنی‌های آموزش زیر مشاهده شده است. ابتدا توضیح دهید مدل از چه مشکلی رنج می‌برد و سپس بیان کنید کدام یک از موارد زیر می‌تواند به بهبود مدل کمک کند. در هر مورد توضیح کوتاهی ارائه دهید و مشخص کنید مقدار bias و variance بعد از اجرای هر کدام از موارد چه تغییری خواهد کرد. (۱۵ نمره)

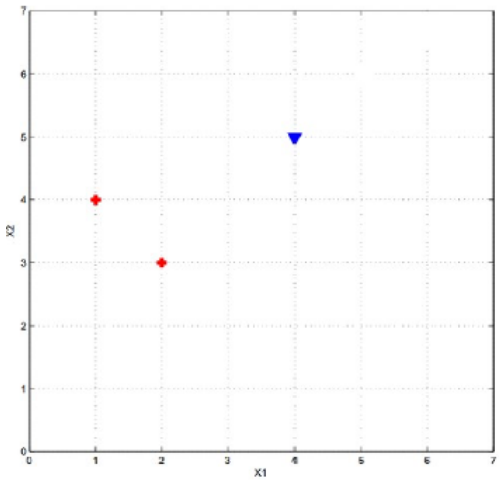
الف) اضافه کردن ویژگی های جدید.

ب) بزرگتر کردن مجموعه آموزشی.

ج) بزرگتر کردن پارامتر منتظم‌سازی.



۴- می‌خواهیم یک طبقه‌بند ماشین بردار پشتیبان را روی داده‌های زیر آموزش دهیم. در این شکل ۲ داده با مقدار -۱ (مثبت‌های قرمز) و ۱ داده با مقدار +۱ (مثلث آبی) نشان داده شده است.(سوال به صورت تحلیلی پاسخ داده شود) (۱۵ نمره)

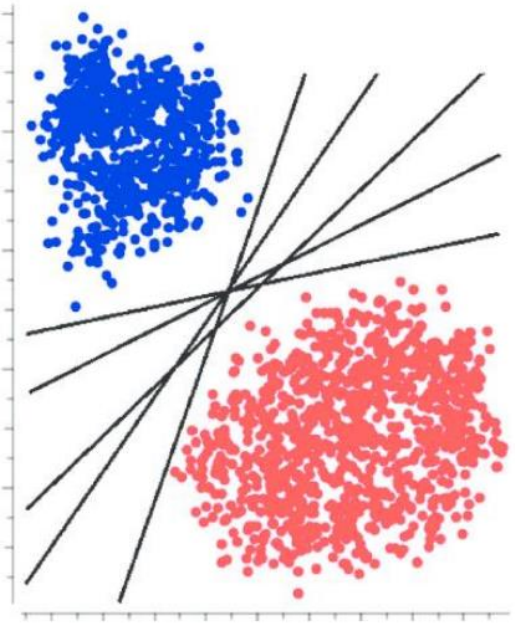


الف) معادله ی خط تصمیم را بدست آورید. (مقادیر w ,b و مارجین یا m را بدست آورید)

ب) نقاط بردار پشتیبان را روی تصویر مشخص کرده و خط تصمیم را رسم کنید.

۵- در مسئله regularized logistic regression زیر، فرض کنید j می‌تواند یکی از سه مقدار ۰،۱،۲ باشد (به عبارت دیگر بردار θ یک بردار با ابعاد یک در سه است) با توجه به داده‌های آموزشی زیر، توضیح دهید بعد از منتظم‌سازی با مقادیر بزرگ λ به ازای هر پارامتر، خطای آموزش چه تغییری می‌کند(به عبارت دیگر با منتظم سازی θ_0 میزان خطا چه تغییری می کند، و سپس به ترتیب منتظم‌سازی θ_1 و θ_2) درباره‌ی تغییرات هر مورد توضیح دهید.(۲۰ نمره)

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^i(\log\left(h_\theta(x^i)\right) - (1-y^i)(\log\left(1-h_\theta(x^i)\right)\right] + \lambda\theta_j^2$$



نکات پیاده سازی و تحویل.

- انجام این تمرین به صورت یک نفره می باشد.
- به همراه فایل ارسالی قالب گزارشی جهت ارسال پاسخ ها قرار داده شده است. در صورت استفاده از قابل فوق و ارسال تمامی پاسخ‌ها به صورت تایپ شده، 10 نمره‌ی امتیازی به نمره‌ی تمرین اضافه خواهد شد. بدیهی است در صورت ارسال پاسخ ها به صورت دست‌نویس نمره‌ی امتیازی در نظر گرفته نخواهد شد.
- برای انجام تمرینها استفاده از زبان برنامه‌نویسی پایتون الزامی می‌باشد.
- در تمرین های برنامه نویسی حتما پیاده سازی خود را در محیط Jupyter Notebook و در یک فایل ipynb انجام دهید.
- نیازی به یک فایل پی دی اف جداگانه برای گزارش بخش پیاده‌سازی نیست. توضیحات خود را در همان فایل ipynb بنویسید . توضیحات به فارسی نوشته شوند در صورت تحویل فایل جداگانه‌ای برای گزارش نمره‌ی این بخش اعمال نخواهد شد.
- در فایل Jupyter Notebook هر سوال از تمرین به همراه پاسخ آن مشخص شده و خروجی‌های مورد نیاز نیز ذخیره شده باشد . همچنین هرگونه نتیجه و یا تحلیلی که در شرح سوال از شما خواسته شده است را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- بعد از تکمیل پاسخ‌ها در فایل Jupyter Notebook مجددا kernel را راه‌اندازی کرده و فایل را اجرا نمایید به صورتیکه شماره‌ی هر سلول در فایل دقیقا مطابق با ترتیب سلول ها باشد.
- تکالیف کامپیوتری تا یک هفته بعد از موعد مقرر قابل تحویل می‌باشند و به ازاء هر روز تأخیر ۷ ٪ از نمرهٔ کل کسر می‌گردد.
- در صورت مشاهده تقلب امتیاز تمامی افراد شرکت‌کننده در آن، ۰ لحاظ می‌شود.
- در صورت وجود سوال و یا ابهام میتوانید از طریق آیدی تلگرام زیر با دستیار آموزشی در ارتباط باشید:

@givkashi

@basir_ebr

@hamidravaee