به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش تمرین اول درس یادگیری ماشین

پاسخ مسئله‌های تشریحی

استاد درس: دکتر احسان ناظرفرد

دانشجو: فاطمه غلام‌زاده

۹۹۱۳۱۰۰۳

نیم سال دوم ۱۳۹۹-۱۴۰۰

سوال ۱)

الف) یادگیری بانظارت: یک روش عمومی در یادگیری ماشین است که در آن مجموعه‌ای از جفت‌های ورودی-خروجی به یک سیستم ارائه شده و سیستم تلاش می‌کند تابعی از ورودی به خروجی را فرا گیرد. در این روش از یادگیری، داده‌های آموزشی دارای برچسب هستند و در واقع آموزش پارامترها توسط این برچسب‌ها انجام می‌شود. در این آموزش برچسب‌ها نقش معلم یا ناظر را دارند و به همین دلیل معروف به یادگیری بانظارت است.

ب) یادگیری نیمه نظارتی: در این روش تعداد کمی از داده‌ها دارای برچسب‌اند و تعداد زیادی از آنها برچسب ندارند و الگوریتم‌های یادگیری نیمه نظارتی به گونه‌ای هستند که یادگیریشان را با هردو دسته از داده‌های برچسب‌دار و بدون برچسب انجام می دهند. هنگامی که داده‌های بدون برچسب همراه تعداد کمی داده‌های برچسب‌دار استفاده می‌شوند موجب افزایش چشم‌گیری در دقت یادگیری می‌شود. در اغلب موارد به دست آوردن داده‌ی برچسب دار نیاز به یک عامل انسانی با مهارت دارد و در این صورت هزینه‌ی فرآیند برچسب‌زنی بسیار زیاد است و برای دیتاست‌های بزرگ عملا غیرقابل اجراست. در چنین مواردی یادگیری نیم نظارتی می‌تواند ارزش عملیاتی بالایی داشته باشد.

ج) یادگیری بدون نظارت: در این روش هیچ برچسبی نداریم و مدل باید خودش داده‌ها را براساس ویژگی‌هایشان جدا کند. این روش بیشتر در مسائل خوشه‌بندی مشاهده می‌شود و چون برچسبی برای نظارت درست بودن عملکرد مدل وجود ندارد به یادگیری بدون نظارت معروف است.

برخی از دلایل استفاده از یادگیری بدون نظارت:

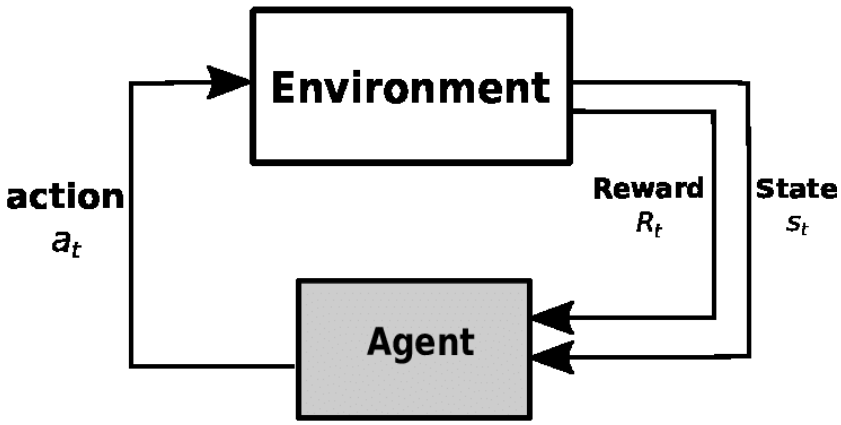
۱) یادگیری بدون نظارت تمام الگوهای ناشناخته موجود در داده‌ها را کشف می‌کند.
۲) هزینه کمتری دارد زیرا برچسب‌زدن داده ها عمل پرهزینه‌ای است اما این روش از داده‌های بدون برچسب استفاده می‌کند.
۳) روش‌های بدون نظارت کمک می‌کنند فیچرهایی را بیابیم که می‌توانند برای دسته‌بندی مفید باشند.

د) یادگیری تقویتی: یکی از گرایش‌های یادگیری ماشین است که از روانشناسی رفتارگرایی الهام می‌گیرد. این روش بر رفتارهایی تمرکز دارد که ماشین باید برای بیشینه کردن پاداشش انجام دهد. یادگیری تقویتی یک عامل (agent) را قادر به یادگیری در محیطی تعاملی با استفاده از آزمون و خطاها و استفاده از بازخوردهای اعمال و تجربیات خود می‌سازد.

گرچه هم یادگیری نظارت شده و هم یادگیری تقویتی از نگاشت بین ورودی و خروجی استفاده می‌کنند، اما در یادگیری تقویتی که در آن بازخوردهای فراهم شده برای عامل مجموعه صحیحی از اعمال جهت انجام دادن یک وظیفه هستند، بر خلاف یادگیری نظارت شده از پاداش‌ها و تنبیه‌ها به عنوان سیگنال‌هایی برای رفتار مثبت و

منفی بهره برده می‌شود. یادگیری تقویتی در مقایسه با یادگیری نظارت نشده دارای اهداف متفاوتی است. در حالیکه هدف در یادگیری نظارت نشده پیدا کردن مشابهت‌ها و تفاوت‌های بین نقاط داده محسوب می‌شود، در یادگیری تقویتی هدف پیدا کردن مدل داده مناسبی است که پاداش انباره‌ای کل[1] را برای عامل بیشینه می‌کند.

در این یادگیری فقط یک سیگنال به عنوان معلم داریم که تنها بیان می کند خروجی تابع مطلوب هست یا خیر، جواب اصلی را نمی گوییم. این روش در عاملها و رباتیک خیلی موضوعیت دارد. درواقع این روش از یادگیری نظارتی ضعیف‌تر است و زمانی سراغ این روش می‌آییم که نتوانیم از نظارتی استفاده کنیم. یعنی دادهها برچسب نداشته باشند. تصویر زیر ایده اساسی و عناصر درگیر در یک مدل یادگیری تقویتی را نشان می‌دهد.



ه) یادگیری انتقالی: به معنای استفاده از یک مدل از پیش آموزش دیده در یک کاربرد جدید است. این تکنیک امروزه در یادگیری عمیق بسیار موردتوجه است زیرا امکان آموزش شبکه‌های عصبی عمیق را با داده‌های نسبتا کمی فراهم می‌کند. برای مثال، دانش به دست آمده هنگام آموزش یک مدل برای تشخیص اتومبیل‌ها می‌تواند زمان ساخت یک مدل برای تشخیص کامیون‌ها مورد استفاده قرار بگیرد.

و) دسته‌بندی: در مسایل دسته‌بندی داده‌های ما داری برچسب هستند که مشخص می کند هر داده مربوط به کدام دسته است و تعداد کل دسته‌ها چندتا است. در این مسائل مدل به گونهای آموزش می‌بیند که با دریافت داده‌ی جدید بتواند حدس بزند آن داده به کدام دسته تعلق دارد؛ یعنی به داده‌های کدام دسته بیشترین شباهت را دارد. دسته‌بندی‌ها از الگوریتم‌های بانظارت هستند. یکی از معروف‌ترین مثال‌های مسئله‌ی دسته‌بندی، دسته-بندی ایمیل‌ها به دو گروه spam و not spam است.

ز) رگرسیون: در فرهنگ لغت واژه رگرسیون (Regression) از لحاظ لغوی به معنی پسروی، برگشت و بازگشت است. اما از دید آمار و ریاضیات به مفهوم بازگشت به یک مقدار متوسط یا میانگین به‌کارمی‌رود. بدین معنی که

[1] total cumulative reward

برخی پدیده‌ها به مرور زمان از نظر کمی به طرف یک مقدار متوسط میل می‌کنند. رگرسیون یک فرایند آماری برای تخمین روابط بین متغیرها می‌باشد. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرهای خاص و منحصر بفرد، با تمرکز بر رابطه بین متغیر وابسته و یک یا چند متغیر مستقل، می‌باشد. تحلیل رگرسیون خصوصاً کمک می‌کند در فهم اینکه چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیرهای مستقل و با ثابت بودن دیگر متغیرهای مستقل تغییر می‌کند. تحلیل رگرسیون به صورت گسترده برای پیش‌بینی استفاده شده‌است. تحلیل رگرسیون همچنین برای شناخت ارتباط میان متغیر مستقل و وابسته و شکل این روابط استفاده شده‌است. در رگرسیون سعی براین است که یک عدد را به عنوان خروجی تخمین بزنیم. در رگرسیون، داده ها دارای یک برچسب عددی هستند و مدل باید به گونهای آموزش ببیند که در نهایت با دریافت یک دادهی جدید که قبال آن را ندیده است یک مقدار مناسب تخمین بزند. رگرسیون بیشتر شبیه یک تابع است ($y = f(x)$)

ح) یادگیری برخط: در این روش داده ها به صورت متوالی و پشت سرهم در دسترس قرار می گیرند. این داده ها برای به روزرسانی بهترین مدل برای داده های آینده، در هر مرحله از یادگیری استفاده میشود. یادگیری برخط یک روش رایج است و در مواردی از یادگیری ماشین مورد استفاده قرار می گیرد که از نظر محاسباتی آموزش کل مجموعه داده غیرممکن است و نیاز به الگوریتم های خارج از هسته است. همچنین در شرایطی که لازم است الگوریتم به صورت پویا با الگوهای جدید داده ها سازگار شود، یا وقتی داده ها خود به عنوان تابعی از زمان تولید می شوند، مورد استفاده قرار می گیرد.

ط) بیش برازش: در مرحله ی یادگیری مدل باید پارامترهای مدل را با استفاده از داده های آموزشی محاسبه کرد؛ بنابراین لازم است که مدل رفتار داده ها را یاد بگیرد. اگر زیاد از حد مدل یادگیری انجام دهد و سعی کند که داده ها را حفظ کند و مدل پیچیدهای در نهایت شود، به این حالت می گوییم مدل بیش برازش شده است. به اصطلاح گفته می‌شود که مدل داده‌ها را یاد نگرفته بلکه به خاطر سپرده است (حفظ کرده است). در حالت بیش برازش واریانس مدل بالا و بایاس پایین است.

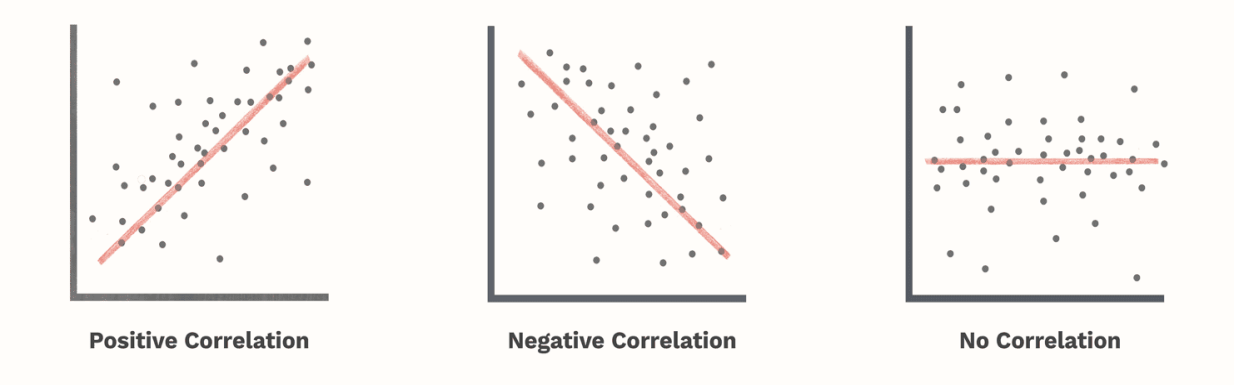
ی) یادگیری فعال: در مواردی که داده‌های بسیار زیادی داریم که تعداد زیادی از آنها برچسب ندارند از این یادگیری استفاده می‌کنیم. در واقع در این روش، از داده‌های برچسب‌خورده استفاده می‌شود تا مدل یادگیری را انجام دهد سپس از همین مدل استفاده می‌شود تا تعدادی از داده‌های بدون برچسب، برچسب گذاری شوند. بعد از آن دوباره یادگیری روی داده‌های جدید برچسب خورده و داده‌های قبلی انجام می‌شود. این فرایند انقدر تکرار می‌شود تا تمام داده‌ها برچسب بخورند.

ک) همبستگی و استقلال ویژگی‌ها: همبستگی چگونگی ارتباط یک یا چند متغیر با یکدیگر را توضیح می دهد. این متغیرها می توانند ویژگی های داده ورودی باشند که برای پیش بینی متغیر هدف ما استفاده شده است. همبستگی ، روش آماری تعیین کننده نحوه حرکت / تغییر یک متغیر در رابطه با متغیر دیگر است. اگر دو متغیر

با یکدیگر همبستگی داشته باشند آن‌گاه می‌توانیم یکی را با استفاده از دیگری پیش‌بینی کنیم. همبستگی میان ویژگی‌ها می‌تواند مثبت یا منفی باشد. اگر با افزایش یک ویژگی، ویژگی دیگر هم افزایش یابد گفته می‌شود که بین این دو ویژگی همبستگی مثبت داریم اما اگر با افزایش یکی دیگری کاهش یابد میان این دو ویژگی همبستگی منفی داریم.

مفهوم استقلال دو ویژگی در مقابل مفهوم همبستگی قرار دارد و بدین معناست که دو یا چند متغیر ارتباطی با یکدیگر ندارند و نمی‌توان یکی را با استفاده از دیگری پیش‌بینی کرد. در واقع عدم وجود همبستگی، استقلال دو ویژگی (متغیر) را نتیجه می‌دهد.

شکل‌های زیر همبستگی مثبت، منفی و عدم وجود همبستگی(استقلال) میان دو ویژگی را نشان می‌دهد.



سوال ۲:

الف) با افزایش تعداد داده‌های آموزش، واریانس مدل کاهش پیدا می‌کند اما بایاس می‌تواند افزایش یا کاهش پیدا کند[1] . در مقاله ذکر شده بر روی ۴ دیتاست این آزمایش صورت گرفته و در تمام ۴ دیتاست با افزایش تعداد داده‌های آموزش واریانس کاهش پیدا کرده است اما بایاس در برخی موارد کاهش و در برخی موارد افزایش داشته است.

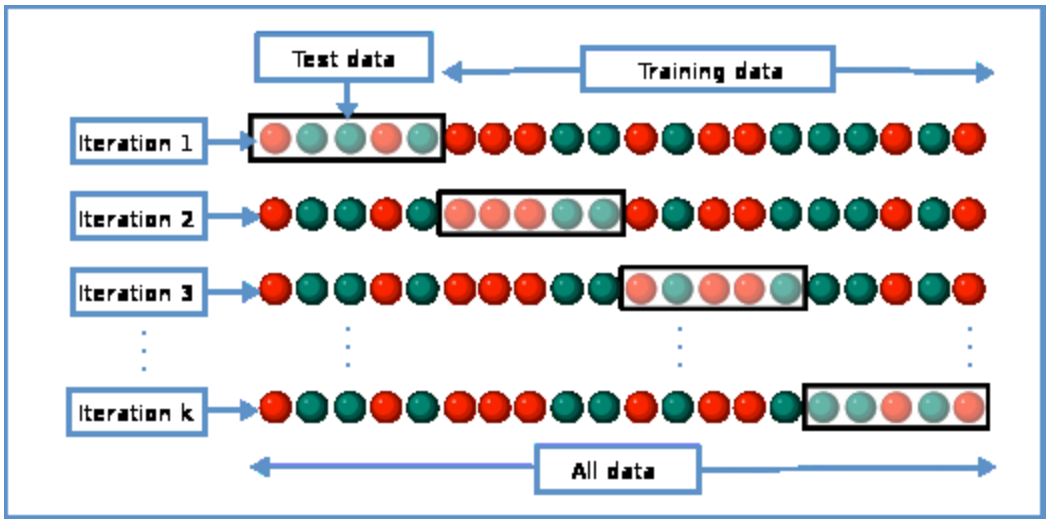
ب)

۱- رگولاریزاسیون(regularization)، در یادگیری ماشین، راهی برای پیشگیری از بیش برازش به حساب می آید. رگولاریزاسیون، بیش برازش را از طریق افزودن جریمه (Penalty) به تابع زیان (Loss Function) کاهش می دهد.

۲- اعتبار سنجی متقابل (Cross-validation) : در این حالت داده‌های نمونه به دو یا چند بخش تفکیک شده و در هر مرحله یکی از بخش‌ها برای برآورد پارامترهای مدل به کار می‌رود. این بخش از نمونه را مجموعه داده‌های آموزشی (Training Set) می‌نامند. بخش‌های دیگر نمونه که به آن مجموعه داده‌های آزمایشی(Test Set) می‌گویند برای سنجش میزان خطای پیش‌بینی مدل به کار می‌روند. روند اعتبار سنجی متقاطع به صورت زیر است:

۱. داده‌ها را به دو بخش آموزشی و آزمایشی تفکیک می‌کنیم.
۲. برای داده‌های آموزشی پارامترهای مدل مناسب را براساس کمینه سازی تابع خطا، برآورد می‌کنیم.
۳. خطای برازش مدل ایجاد شده را روی داده‌های آزمایشی اندازه‌گیری می‌کنیم.
۴. نسبت خطای بدست آمده از مدل، برای داده‌های آزمایشی و آموزشی نباید خیلی بزرگ باشد.
5. مراحل ۱ تا ۴ را با توجه به همگرا شدن نسبت حاصل از مرحله ۴ ادامه می‌دهیم در غیر این صورت به تعداد تکرار مشخص، عملیات پایان می‌یابد (هر کدام زودتر به وقوع بپیوندد). با توجه به میزان نسبت خطاهای ذکر شده بهترین مدل در این مرحله حاصل می‌شود.

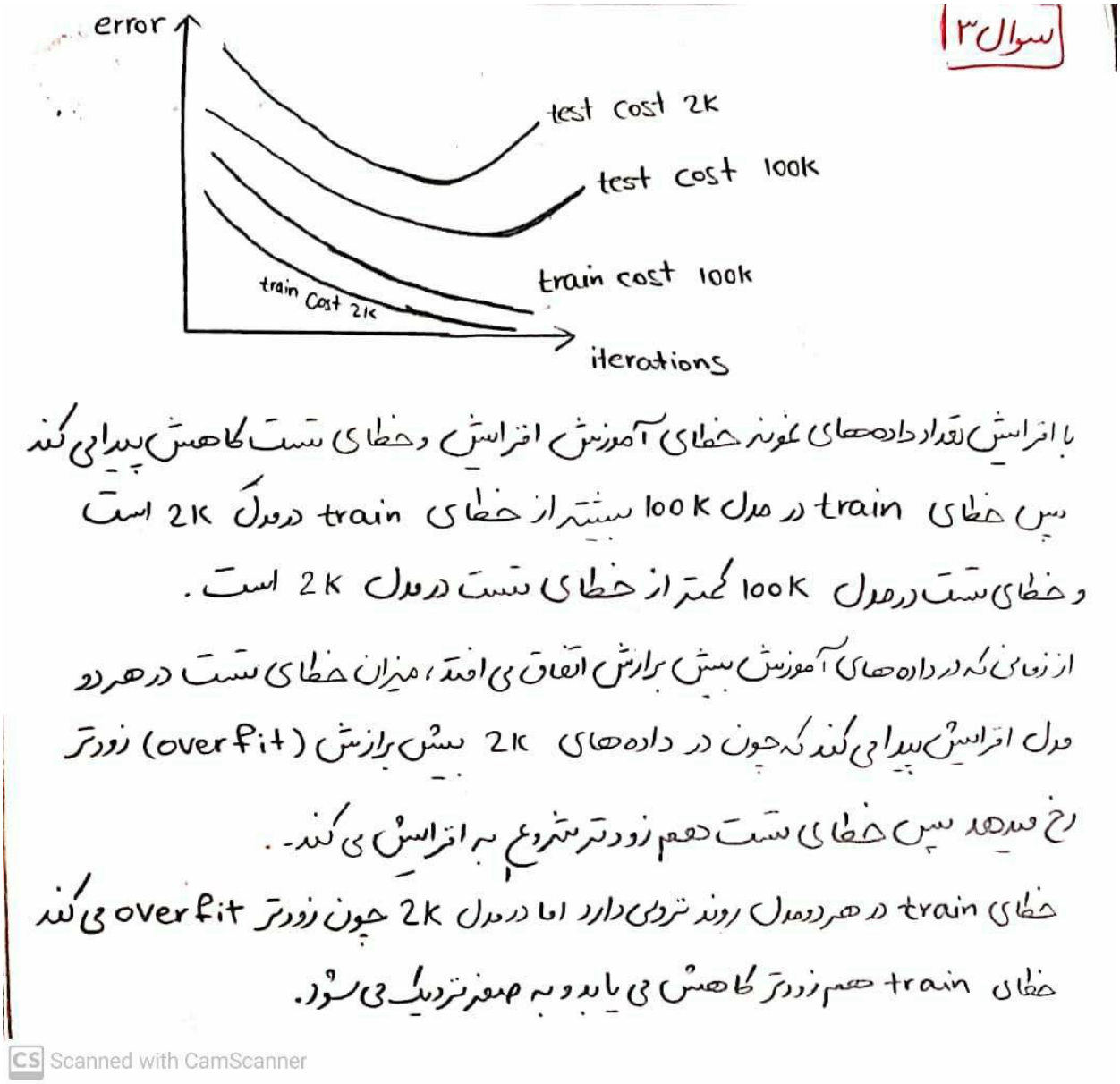
شکل زیر نحوه اعتبارسنجی متقابل (Cross-validation) را نشان می‌دهد.



۳- توقف زودهنگام (Early stopping): هنگام استفاده از یک مدل تکرارشونده (iterative) روند آموزش را قبل از آخرین تکرار متوقف می‌کنیم. این کار موجب می‌شود که مدل مجموعه داده‌ها را حفظ نکند (به خاطر نسپارد).

۴- استفاده از داده‌های آموزشی بیشتر و هم چنین انتخاب ویژگی‌ها (feature selection) نیز راهکارهایی برای مقابله با بیش برازش هستند.

سوال ۳)



سوال ۴)

خطای MSE : خطای میانگین مربعات یا MSE در واقع میانگینی از مربعات اختلاف مقدار پیش‌بینی شده برای داده‌ها و مقدار واقعی آن‌هاست. برای بدست آوردن خطای میانگین مربعات از یک مجموعه یا n داده می‌توان از رابطه زیر استفاده کرد:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

که در آن $\left(\frac{1}{n}\sum_{i=1}^{n}\right)$ عمل میانگین‌گیری را انجام می‌دهد و $(Y_i - \hat{Y}_i)^2$ مقدار مربع خطای هر داده را محاسبه می‌کند. پس MSE میانگین مربع خطاها است.

خطای RMSE: اگر از خطای MSE جذر بگیریم این خطا به دست می‌آید و در واقع انحراف استاندارد باقی‌مانده‌ها یا خطاهای پیش‌بینی (residuals) را نشان می‌دهد. باقی‌مانده‌ها معیاری برای اندازه‌گیری میزان فاصله خط رگرسیون و داده‌های واقعی هستند. خطای RMSE معیاری برای اندازه‌گیری گسترش این باقی‌مانده‌هاست. فرمول خطای RMSE :

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

خطای MAE: اختلاف مقدار اندازه‌گیری شده کمیت از مقدار واقعی آن را خطای مطلق می‌گویند. در مدل یادگیری ماشین اگر از اختلاف مقادیر پیش‌بینی شده برای داده‌ها با مقدار واقعی داده‌ها، میانگین بگیریم خطای میانگین مطلق یا MAE به دست می‌آید. فرمول محاسبه این خطا:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

معمولا وقتی مقدار خطاها بالاست از خطای MSE و RMSE استفاده می‌شود و سعی می‌شود از خطای MAE استفاده نشود. همچنین استفاده از خطای RMSE از MSE رایج تر است زیرا واحد آن با واحد متغیر وابسته (Y) یکسان است.

MSE یک تابع مشتق‌پذیر است و همین امر انجام عملیات ریاضی را در مقایسه با یک تابع مشتق‌ناپذیر مثل MAE و RMSE آسان می‌کند بنابراین در بسیاری از مدل‌ها خطای MSE و RMSE استفاده می‌شوند اگر چه تفسیر آن‌ها نسبت به MAE دشوارتر است.

خطای MAE نسبت به داده‌هایی که دارای outlier هستند مقاوم‌تر است بنابراین در چنین شرایطی استفاده از MAE نسبت به MSE ارجحیت دارد. همچنین در مواردی که می‌خواهید به وجود outlierها در داده‌های خود پی ببرید یک روش استفاده از خطای MSE و RMSE است.

سوال۵)

اثر تکانه: روشی است که به تسریع بردارهای شیب در جهت های مناسب کمک می کند ، بنابراین منجر به همگرایی سریع تر می شود. یکی از محبوب ترین الگوریتم های بهینه سازی است و بسیاری از مدل های پیشرفته با استفاده از آن آموزش می بینند. یک مثالی که درمورد گرادیان نزولی و اثر تکانه بیان می‌شود به این شکل است: گرادیان نزولی مثل یک فرد است که از یک تپه پایین می رود. او شیب دارترین مسیر را به سمت پایین دنبال می کند. پیشرفت او کند اما ثابت است. تکانه یک توپ سنگین است که از همان تپه می چرخد. اینرسی اضافه شده هم به عنوان نرم کننده و هم تسریع کننده عمل می کند، نوسانات را کاهش می دهد و باعث می شود ما دره های باریک، کوهان های کوچک و مینیمم های محلی را با سرعت رد کنیم و در آن‌ها گرفتار نشویم.

مزایا: همگرایی سریع‌تر، کاهش نوسانات

مشکلات تکانه زیاد و کم : اگر تکانه کم باشد نمی‌تواند مدل را از گیر افتادن در مینیمم‌های محلی نجات بدهد و هم‌چنین روند همگرایی خیلی کند صورت می‌گیرد.

اگر تکانه زیاد باشد موجب می‌شود که همگرایی به سرعت رخ بدهد و در نتیجه به جواب بهینه نرسیم در صورتی که تکانه زیاد باشد بهتر است نرخ یادگیری را کاهش بدهیم تا از سرعت همگرایی کاسته شود.

سوال ۶ الف)

$$J(w)=\sum_{i=1}^{n}(y_i-w^T x_i)^2=(y-Xw)^T(y-Xw)$$

$$=((Xw)^T-y^T)(Xw-y)=(Xw)^TXw-(Xw)^Ty-y^TXw+y^Ty$$

(عدد ثابت است ←)

$$=(Xw)^TXw-2(Xw)^Ty+y^Ty$$

$$\Rightarrow J(w)=(Xw)^TXw-2(Xw)^Ty+y^Ty$$

از تابع هزینه $J(w)$ نسبت به w مشتق می‌گیریم تا مقدار بهینه w به دست آید:

$$\frac{\partial J(w)}{\partial w}=2X^TXw-2X^Ty=0$$

$$\Rightarrow \hat{w}=(X^TX)^{-1}X^Ty$$

ب) ① زمانی که متغیر وابسته داشته باشیم (یعنی همه‌ی ستون‌های X از هم مستقل نباشند) نمی‌توان از رابطه بالا استفاده کرد زیرا حلقه X^TX وارون ندارد.

راه حل: متغیری اضافه کنیم که از سایر متغیرها مستقل باشد و آن متغیر وابسته را حذف کنیم

② زمانی که تعداد ستون‌ها خیلی بیشتر از تعداد سطرها باشد ← یعنی تعداد زیادی فیچر داشته باشیم

در این صورت هم X^TX وارونپذیر نیست و نمی‌توان از این رابطه استفاده کرد

راه حل ← حذف برخی از فیچرهایی که کمتر مفید هستند (تعداد ستون‌ها کاهش می‌یابد)

ج) اگر یکی از ابعاد داده ها ترکیب خطی از سایر ابعاد داده ها باشد بدین معناست که ستون های ماتریس X مستقل از هم نیستند و طبق تعیینی در جبر خطی اگر ستون های ماتریس از هم مستقل نباشند ماتریس وارون پذیر نیست و در نتیجه جمله ی X^TX هم که در رابطه وجود دارد وارون پذیر نیست.

راه حل: یکی از ستون هایی که به هم وابستگی دارند را حذف کنیم و به جای آن یک ستون دیگه که مستقل از سایر ستون هاست اضافه کنیم.

د) با اضافه کردن جمله منظم ساز $\|w\|^2$، فرم بسته w بهینه به صورت زیر است:

$$\hat{w} = (X^TX + \lambda I)^{-1} X^T y$$

ه)
$$J(w) = \sum_{i=1}^{n} F_i (y_i - w^T x_i)^2 = \frac{1}{2}(Xw-y)^T F (Xw-y)$$

$$\frac{\partial J(w)}{\partial w} = \frac{1}{2} \frac{\partial}{\partial w} (Xw-y)^T F (Xw-y)$$

$$= \frac{1}{2} \frac{\partial}{\partial w} (w^T X^T F X w - w^T X^T F y - y^T F X w + y^T F y)$$

$$= X^T F X w - X^T F y$$

برای یافتن بهینه ی w باید مشتق را مساوی صفر قرار دهیم:

$$\frac{\partial J(w)}{\partial w} = 0 \Rightarrow X^T F X w - X^T F y = 0$$

$$\Rightarrow \hat{w} = (X^T F X)^{-1} X^T F y$$

سوال ۷

الف)

$$y|x_1,x_2 \sim N(w_0+w_1x_1+w_2x_2+w_3x_1^2,\sigma^2)$$

$$\Rightarrow P(y|x_1,x_2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-w_0-w_1x_1-w_2x_2-w_3x_1^2)^2}{2\sigma^2}}$$

ب)

$$\ell(w_0,w_1,w_2,w_3)=\log\prod_{i=1}^{n}P(y^{(i)}|x_1^{(i)},x_2^{(i)})$$

$$=\sum_{i=1}^{n}\log P(y^{(i)}|x_1^{(i)},x_2^{(i)})=\sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y^{(i)}-w_0-w_1(x_1^{(i)})-w_2(x_2^{(i)})-w_3(x_1^{(i)})^2)^2}{2\sigma^2}}$$

$$=\sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi\sigma^2}}+\sum_{i=1}^{n}\log e^{-\frac{(y^{(i)}-w_0-w_1(x_1^{(i)})-w_2(x_2^{(i)})-w_3(x_1^{(i)})^2)^2}{2\sigma^2}}$$

$$=\sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi\sigma^2}}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^{(i)}-w_0-w_1(x_1^{(i)})-w_2(x_2^{(i)})-w_3(x_1^{(i)})^2)^2$$

ج) ماکزیمم کردنِ Conditional log likelihood معادلِ مینیمم کردنِ negative conditional log likelihood

می باشد:

$$f(w_0,w_1,w_2,w_3)=-\ell(w_0,w_1,w_2,w_3)$$

$$=\sum_{i=1}^{n}(y^{(i)}-w_0-w_1(x_1^{(i)})-w_2(x_2^{(i)})-w_3(x_1^{(i)})^2)^2$$

$$\frac{\partial f(\mathbf{w})}{\partial w_0} = -2\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_1(x_1^{(i)}) - w_2(x_2^{(i)}) - w_3(x_1^{(i)})^2\right)$$

$$\frac{\partial f(\mathbf{w})}{\partial w_1} = -2\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_1(x_1^{(i)}) - w_2(x_2^{(i)}) - w_3(x_1^{(i)})^2\right)(x_1^{(i)})$$

$$\frac{\partial f(\mathbf{w})}{\partial w_2} = -2\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_1(x_1^{(i)}) - w_2(x_2^{(i)}) - w_3(x_1^{(i)})^2\right)(x_2^{(i)})$$

$$\frac{\partial f(\mathbf{w})}{\partial w_3} = -2\sum_{i=1}^{n}\left(y^{(i)} - w_0 - w_1(x_1^{(i)}) - w_2(x_2^{(i)}) - w_3(x_1^{(i)})^2\right)(x_1^{(i)})^2$$

مراجع:

[1] G. I. W. Damien Brain, “On the effect of data set size on bias and variance in classification .learning,” *School of Computing and Mathematics, Deakin University*