به نام زیباترین

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳

استاد: علی شریفی زارچی

گردآورندگان: ماهان بیهقی - پیام تائبی - امیررضا آذری

۲۴ آبان ۱۴۰۳ | آزمون میان‌ترم اول | زمان: ۳ ساعت

نام و نام خانوادگی: | شماره دانشجویی:

لطفا پاسخ هر سوال را در یک برگه‌ی جداگانه بنویسید. بالای هر برگه نام و شماره دانشجویی خود را حتما قید کنید. در پایان آزمون، برگه‌ی سوال‌های مختلف را از هم جدا کنید و هر برگه را در دسته‌ی مربوط به خود قرار دهید. هر مساله ۲۰ نمره دارد. نمره‌ی کامل آزمون ۱۰۰ است و ۲۰ نمره اضافه دارد.

الاکلنگ بایاس و واریانس

۱. تفاوت کاربرد داده‌های Validation و Test چیست؟ فقط به مهم‌ترین موضوع اشاره کنید.

۲. فرض کنید قیمت یک رمزارز از تابع زیر تبعیت می‌کند: $y = (۱ + sin(\frac{\pi x}{۱۰})) + \epsilon$ که در آن x شماره‌ی روز از ماه است. سه مدل زیر را در نظر بگیرید:

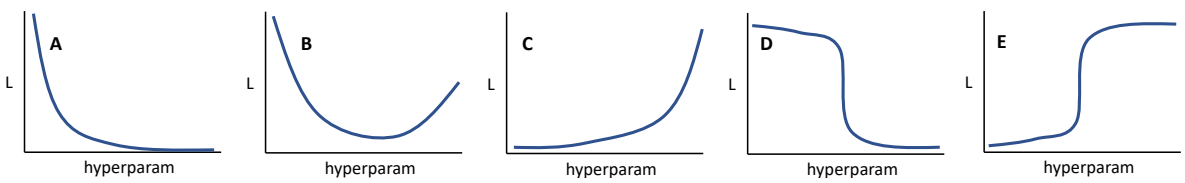
- $\hat{y_1} = \theta_1 x + \theta_0$
- $\hat{y_3} = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$
- $\hat{y_9} = \theta_9 x^9 + \cdots + \theta_2 x^2 + \theta_1 x + \theta_0$

کم یا زیاد بودن مقدار Bias و Variance هریک از این سه مدل را در حالت‌های زیر با دلیل مشخص کنید. نیازی به محاسبات نیست.

- وقتی ۱۰۰ داده از قیمت این رمزارز در طول یک ماه داریم.

- وقتی ۵ داده از قیمت این رمزارز در طول یک ماه داریم.

۳. در شکل زیر محور x تغییرات یک Hyperparameter و محور y مقدار تابع زیان Loss را نشان می‌دهد.



برای هریک از قسمت‌های زیر فقط یکی از شکل‌های A تا E را انتخاب کنید که محتمل‌ترین تغییر رفتار تابع زیان بر اساس تغییرات مقدار Hyperparameter است. هم‌چنین، دلیل انتخاب خود را توضیح دهید.

(الف) k: تعداد همسایگان در الگوریتم k نزدیک‌ترین همسایه (kNN)

- میزان خطا با تغییر k

○A ○B ○C ○D ○E

(ب) d: عمق یک درخت تصمیم

- زیان آموزش Training Loss

○A ○B ○C ○D ○E

- زیان تست Test Loss

○A ○B ○C ○D ○E

(پ) α نرخ یادگیری Learning Rate در رگرسیون لاجستیک Logistic Regression

- زیان آموزش Training Loss

○A ○B ○C ○D ○E

- زیان تست Test Loss

○A ○B ○C ○D ○E

(ت) تعداد درختان در جنگل تصادفی Random Forest

- زیان آموزش Training Loss

○A ○B ○C ○D ○E

- زیان تست Test Loss

○A ○B ○C ○D ○E

نام و نام خانوادگی: شماره دانشجویی:

خوشه خوشه

در این مسئله می‌خواهیم به بررسی الگوریتم خوشه بندی k-means بپردازیم. فرض کنید $X = x_1, x_2, ..., x_n$ داده‌های ما باشد و γ یک ماتریس Indicator باشد به این صورت که $\gamma_{ij} = 1$ اگر x_i متعلق به خوشه j ام باشد و در غیر این صورت برابر ۰ است. فرض کنید $\mu_1, ..., \mu_k$ میانگین خوشه ها باشند. اعوجاج J برای داده‌ها به صورت زیر محاسبه می‌شود :

$$J(\gamma, \mu_1, ..., \mu_k) = n \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2$$

همچنین $C = 1, ..., k$ را به عنوان مجموعه خوشه ها در نظر بگیرید.

۱. آیا k-means نسبت به انتخاب نقاط اولیه حساس است، یعنی پاسخ آن بر اساس مجموعه‌ی نقاط اولیه تغییر می‌کند؟ اگر بله یک مثال ارائه کنید و اگر خیر، اثبات کنید.

۲. نشان دهید که الگوریتم k-means در زمان متناهی قدم به پایان می‌رسد. (راهنمایی: نشان دهید J تعداد محدودی حالت دارد.)

۳. اگر ابعاد داده نسبت به تعداد نمونه‌ها خیلی زیاد باشد و عملا نمونه‌ها در یک فضای بزرگ پراکنده باشند، برای بهبود خوشه‌بندی از چه روشی استفاده می‌کنید؟

۴. نشان دهید که کمینه J یک تابع غیرافزایشی بر حسب k یا همان تعداد خوشه هاست. در این صورت آیا انتخاب مقدار هایپرپارامتر k بر اساس کمینه‌کردن مقداز J ایده‌ی خوبی است؟ اگرنه، چه ایده‌ی بهتری دارید؟

۵. فرض کنید $\hat{x}$ میانگین داده‌های نمونه باشد. مقادیر زیر را در نظر بگیرید.

$$T(X) = \frac{\sum_{i=1}^{n} \|x_i - \hat{x}\|^2}{n}$$

$$W_j(X) = \frac{\sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$B(X) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \|\mu_j - \hat{x}\|^2$$

در اینجا $T(X)$ نشان دهنده انحراف کلی، $W_j(X)$ انحراف درون خوشه‌ای و $B(X)$ انحراف بین خوشه‌ای است. رابطه‌ی بین این ۳ مقدار به چه صورت است؟ نشان دهید که k-means میتواند به عنوان کمینه کننده میانگین وزن دار مقادیر درون خوشه‌ای و به طور تقریبی بیشینه کردن انحراف بین خوشه‌ای دیده شود.

خلاف شیب

در بسیاری از سناریوهای دنیای واقعی، داده‌های ما دارای میلیون‌ها بُعد هستند، اما یک نمونه خاص فقط دارای صدها ویژگی غیرصفر است. به عنوان مثال، در تحلیل اسناد با تعداد کلمات به عنوان ویژگی‌ها، ممکن است فرهنگ لغت ما میلیون‌ها کلمه داشته باشد، اما یک سند خاص فقط دارای صدها کلمه منحصر به فرد است. در این سؤال، می‌خواهیم نرم ℓ_2-Regularized Stochastic Gradient Descent (SGD) برای زمانی که داده‌های ورودی ما sparse است را کارا کنیم. به خاطر داشته باشید که در Logistic Regression ℓ_2 ، می‌خواهیم تابع هدف زیر را کمینه کنیم (در این مسئله برای سادگی w_0 حذف شده است):

$$F(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N} l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) + \frac{\lambda}{2}\sum_{i=1}^{d} w_i^2$$

که در آن $l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w})$ تابع هدف است:

$$l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) = \ln\left(1 + \exp\left(\sum_{i=1}^{d} w_i x_i^{(j)}\right)\right) - y^{(j)}\left(\sum_{i=1}^{d} w_i x_i^{(j)}\right)$$

و باقی مانده‌ی جمع، میزان Regularization Penalty خواهد بود.

وقتی روی نقطه $(\mathbf{x}^{(j)}, y^{(j)})$ SGD انجام می‌دهیم، تابع هدف را به صورت زیر تقریب می‌زنیم:

$$F(\mathbf{w}) \approx l(\mathbf{x}^{(j)}, y^{(j)}, \mathbf{w}) + \frac{\lambda}{2}\sum_{i=1}^{d} w_i^2$$

۱. ابتدا حالت $\lambda = ۰$ را در نظر بگیرید. قانون به‌روزرسانی SGD برای w_i را هنگامی که $\lambda = ۰$ است، با استفاده از اندازه گام η و با توجه به نمونه $(\mathbf{x}^{(j)}, y^{(j)})$ بنویسید.

۲. اگر از یک ساختار داده متراکم استفاده کنیم، میانگین پیچیدگی زمانی برای به‌روزرسانی w_i هنگامی که $\lambda = ۰$ است، چقدر است؟ اگر از یک ساختار داده تنک استفاده کنیم، چطور؟ پاسخ خود را در یک یا دو جمله توضیح دهید.

۳. اکنون حالت کلی را که $\lambda > ۰$ در نظر بگیرید. قانون به‌روزرسانی SGD برای w_i را هنگامی که $\lambda > ۰$ است، با استفاده از اندازه گام η و با توجه به نمونه $(\mathbf{x}^{(j)}, y^{(j)})$ بنویسید.

۴. اگر از یک ساختار داده متراکم استفاده کنیم، میانگین پیچیدگی زمانی برای به‌روزرسانی w_i هنگامی که $\lambda > ۰$ است، چقدر است؟

۵. فرض کنید $\mathbf{w}_i^{(t)}$ بردار وزن بعد از به‌روزرسانی tام باشد. اکنون فرض کنید که k به‌روزرسانی SGD روی $\mathbf{w}$ با استفاده از نمونه‌های $(\mathbf{x}^{(t+۱)}, y^{(t+۱)}), \cdots, (\mathbf{x}^{(t+k)}, y^{(t+k)})$ انجام می‌دهیم، که در آن $x_i^{(j)} = ۰$ برای هر نمونه در دنباله باشد (یعنی ویژگی iام برای تمام نمونه‌ها در دنباله صفر است). وزن جدید $\mathbf{w}_i^{(t+k)}$ را بر حسب $\mathbf{w}_i^{(t)}$، k، η، و λ حساب کنید.

۶. با استفاده از پاسخ خود در قسمت قبل، یک الگوریتم کارا برای Regularized SGD ارائه دهید زمانی که از ساختار داده sparse استفاده می‌کنیم. میانگین پیچیدگی زمانی به‌ازای هر نمونه چقدر است؟
راهنمایی: چه زمانی نیاز به به‌روزرسانی w_i دارید؟

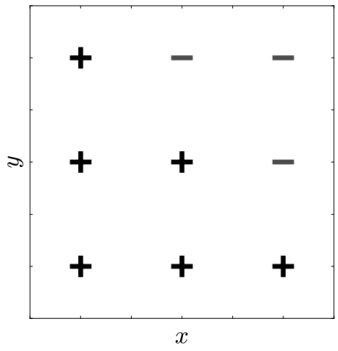
خرد جمعی

می‌دانیم که Adaboost یک دسته‌بند H را با استفاده از جمع وزن‌دار یادگیرنده‌های ضعیف h_t به صورت زیر یاد می‌گیرد:

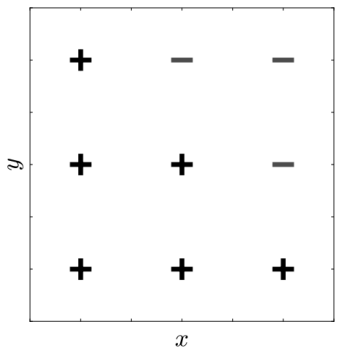
$$H(x) = \text{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

در این سوال، ما از درخت‌های تصمیم به عنوان یادگیرنده‌های ضعیف خود استفاده می‌کنیم، که یک نقطه را به عنوان $\{۱, -۱\}$ بر اساس دنباله‌ای از thresholdها روی ویژگی‌های آن (اینجا x, y) طبقه‌بندی می‌کنند.
در سوالات زیر فرض کنید که در صورت برابری امتیاز برای کلاس مثبت و منفی، خروجی دسته‌بندها به طور دلخواه تعیین می‌شود.

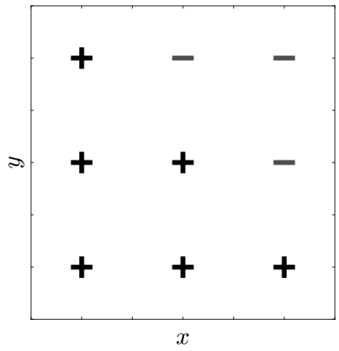
۱. فرض کنید یادگیرنده‌های ضعیف ما درخت‌های تصمیم با عمق ۱ هستند (Decision Stumps)، که خطای آموزشی وزن‌دار را کمینه می‌کنند. با استفاده از مجموعه داده زیر، مرز تصمیمی که توسط $h_۱$ یاد گرفته شده است را ترسیم کنید.



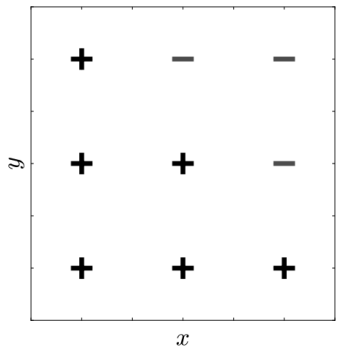
۲. در مجموعه داده‌ی زیر، نقطه(های) با بیشترین وزن در iteration دوم را مشخص و مرز تصمیمی که توسط $h_۲$ یاد گرفته شده است را ترسیم کنید.



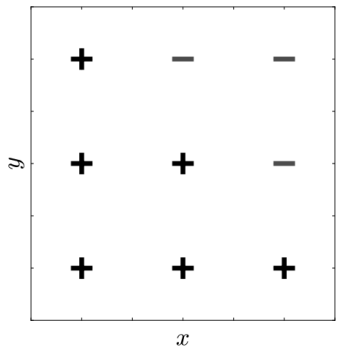
۳. در مجموعه داده‌ی زیر، مرز تصمیم $H = \text{sgn}(\alpha_۱ h_۱ + \alpha_۲ h_۲)$ را ترسیم کنید. (راهنمایی: نیازی به محاسبه صریح α ها نیست).



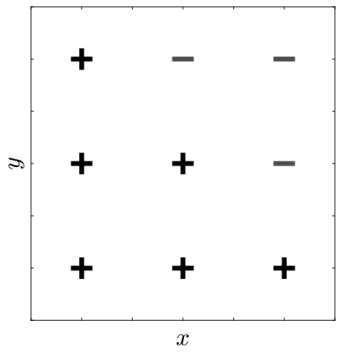
۴. اکنون فرض کنید که یادگیرنده‌های ضعیف ما درخت‌های تصمیم با عمق حداکثر ۲ هستند، که خطای آموزشی وزن‌دار را کمینه می‌کنند. با استفاده از مجموعه داده‌ی زیر، مرز تصمیمی که توسط h_1 یاد گرفته شده است را ترسیم کنید.



۵. در مجموعه داده‌ی زیر، نقطه(ها) با بیشترین وزن در iteration دوم را دایره بکشید و مرز تصمیمی که توسط h_2 یاد گرفته شده است را ترسیم کنید.



۶. در مجموعه داده‌ی زیر، مرز تصمیم $H = \text{sgn}(\alpha_1 h_1 + \alpha_2 h_2)$ را ترسیم کنید. (راهنمایی: نیازی به محاسبه صریح α ها نیست).



Tikhonov

تابع هزینه مسئله رگرسیون خطی به صورت زیر تعریف می‌شود:

$$\mathbb{L}_۱(w) = \|y - Xw\|_۲^۲$$

همانطور که در درس دیدید، می‌توانیم چند عنصر دیگر به عنوان Regularization Term به این تابع هزینه اضافه کنیم. در این صورت خواهیم داشت:

$$\mathbb{L}_۲(w) = \|y - Xw\|_۲^۲ + \|\Gamma w\|_۲^۲$$

که به Γ، Tikhonov matrix می‌گویند. این حالت کلی مسئله ridge regression است که در اینجا به جای λ، از یک ماتریس استفاده می‌کنیم. برای آموزش مدل رگرسیون خطی خود، می‌خواهیم از تکنیکی به نام $dropout$ استفاده کنیم. این تکنیک برای ورودی d بعدی، هر ویژگی را با احتمال p نگه داشته و در غیر این صورت برابر صفر خواهد شد. با استفاده از این تکنیک، تابع هزینه به شکل زیر تغییر خواهد کرد :

$$\mathbb{L}_۳(w) = \mathbb{E}_{D \sim Bernouli(p)}[\|y - (D \odot X)\hat{w}\|_۲^۲]]$$

توجه کنید در اینجا $\hat{w}$ پارامتر های پیدا شده توسط مدلی است که با $dropout$ آموزش داده شده است. همچنین ضرب $\odot$، ضرب element wise می‌باشد.

۱. ابتدا معادله نرمال برای حل مساله‌ی minimization بدون توجه به $dropout$ به دست آورید. در اینجا شما مانند دیگر مسئله های رگرسیون، باید وزن های بهینه را با استفاده از مشتق و ... با استفاده از تابع هزینه به دست آورید.

۲. یک شرط ساده، کافی و لازم برای ماتریس Γ بیان کنید که تضمین کند تابع هزینه یک جواب منحصر به فرد و بهینه برای $\hat{w}$ دارد.

۳. حال اثبات کنید هنگام استفاده از تکنیک $dropout$، می‌توان تابع هزینه را به شکل زیر بازنویسی کرد :

$$\mathbb{L}(w) = \|y - pX\hat{w}\|_۲^۲ + p(۱-p)\|\hat{\Gamma}\hat{w}\|_۲^۲$$

به طوری که $\hat{\Gamma}$ یک ماتریس قطری بوده که عنصر j ام قطری این ماتریس، برابر نرم ستون j ام ماتریس دادگان X می باشد.

۴. فرض کنید Γ معکوس پذیر باشد. با یک تغییر متغیر سعی کنید تا تابع هزینه گفته شده در حالت بدون $dropout$را به صورت تابع هزینه مسئله ridge regression بازنویسی کنید :

$$\mathbb{L}(\hat{w}) = \|y - \hat{X}\hat{w}\|_۲^۲ + \lambda\|\hat{w}\|_۲^۲$$

نام و نام خانوادگی: شماره دانشجویی:

اصلی

در یک شرکت بزرگ کاریابی کار می‌کنید. هر فرد یک پروفایل دارد که بعضی ویژگی‌های آن (نظیر سن، آخرین حقوق) عدد پیوسته و بعضی ویژگی‌های دیگر (نظیر رشته‌ی تحصیلی) Categorical است. همچنین بعضی از ویژگی‌ها (نظیر آشنایی با هریک از زبان‌های برنامه‌نویسی) به صورت صفر و یک درج شده است.

۱. می‌خواهید برای هر کاربر جدید که پروفایل خود را تکمیل کرده‌است، یک مبلغ حقوق تخمین بزنید. از چه الگوریتمی استفاده می‌کنید؟ چه تغییری روی ویژگی‌ها می‌دهید؟ برای هریک از ویژگی‌ها از چه پیش‌پردازشی استفاده می‌کنید؟ جزئیات را توضیح دهید.

۲. حال می‌خواهید داده‌های این شرکت را به صورت یک نمودار نمایش دهید. برای این کار تصمیم دارید از تحلیل مولفه‌های اصلی (PCA) استفاده کنید. آیا به نظر شما تغییر مقیاس ویژگی‌ها لازم است؟ اگر بله، چرا؟ و چطور این کار را انجام می‌دهید؟ اگر خیر، دلیل شما چیست؟

۳. فرض کنید ماتریس کوواریانس زیر را داشته باشید. چطور از روی آن مولفه‌های اصلی را مشخص می‌کنید؟ محاسبات خود را بنویسید.

$$C = \begin{bmatrix} 3 & -4 \\ -4 & 3 \end{bmatrix}$$

۴. فرض کنید می‌خواهید با PCA ابعاد داده‌هاهای شرکت را با بردن آن‌ها فضای جدیدی کاهش دهید که عمده‌ی اطلاعات حفظ شود. چطور تعداد بعدهای فضای جدید را مشخص می‌کنید؟

۵. ثابت کنید تصویرکردن داده‌ها توسط برداررویژه‌ی با بزرگ‌ترین مقدار ویژه‌ی ماتریس کوواریانس، واریانس داده‌های تصویرشده را بیشینه می‌کند.

۶. به دل‌خواه یک مساله‌ی زیبای یادگیری ماشین بنویسید و آن را حل کنید. نمره‌ی این بخش به زیبایی و منحصر به فرد بودن مساله و درستی راه‌حل شما اختصاص می‌یابد.