دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین
بهار ۱۴۰۴
استاد:

آزمون پایانترم

سوال ۱ حل.

۱.۱) **ساختار دو درخت تصمیم یکسان خواهد بود.** زیرا که درخت تصمیم بر اساس مقادیر ویژگی‌ها و ترتیب نسبی آن‌ها تقسیم‌بندی انجام می‌دهد، نه مقیاس عددی آن‌ها. به عبارت دیگر، الگوریتم درخت تصمیم به دنبال یافتن آستانه‌هایی است که داده‌ها را به‌خوبی تفکیک کند، و این آستانه‌ها از روی مقایسه‌ی مقادیر استخراج می‌شوند نه مقدار مطلق ویژگی‌ها. از طرفی استانداردسازی ترتیب داده‌ها را حفظ می‌کند. در فرآیند استانداردسازی، فقط میانگین و انحراف معیار تغییر داده می‌شود، اما ترتیب نسبی مقادیر هر ویژگی ثابت می‌ماند. بنابراین، درخت تصمیم همان تقسیم‌بندی را انجام خواهد داد. از طرفی دیگر معیارهای ارزیابی تقسیم (مانند Gini و آنتروپی) به مقیاس عددی وابسته نیستند. در نتیجه این معیارها براساس توزیع نمونه‌ها در گره‌ها محاسبه می‌شوند و نه مقادیر دقیق ویژگی‌ها و بنابراین تغییر در مقیاس داده‌ها اثری بر انتخاب تقسیم‌ها نخواهد داشت.

۳.۱) پیشنهاد او از لحاظ مفهومی دچار مشکل است. یکی از فرضیاتی که در رابطه با دیتای تست داریم این است که ما این داده‌ها و اطلاعات مربوط آن‌ها را در زمان آموزش در اختیار نداریم و تنها در هنگام تست می‌توانیم این داده‌ها را دسترسی داشته باشیم. بنابراین اگر الگوریتم PCA را استفاده کنیم، از اطلاعات موجود در داده‌ی تست برای نرمال‌سازی و یافتن بردارهای اصلی استفاده کرده‌ایم و نشت اطلاعات داریم.

۴.۱) خیر، تغییری ایجاد نمی‌شود. علت این امر این است که ویژگی‌ای که مقدار ثابتی (مانند یک) داشته باشد، دارای واریانس صفر می‌باشد و در الگوریتم PCA ویژگی‌های با بیشترین میزان واریانس انتخاب می‌شوند و در نتیجه این ویژگی تاثیری نخواهد داشت.

۵.۱) درست است چون PCA در واقع L بزرگترین مقدار ویژه را در نظر گرفته و از بردار ویژه های نظیر آن ها استفاده می کند؛ پس با دوبار انجام دادنش، همچنان به همان بردارها با بزرگ‌ترین مقدار ویژه می رسیم.

۶.۱) احتمال این‌که نمونه‌ی iام در بین نمونه‌های انتخاب شده نباشد برابر با $(1-\frac{1}{N})^{Np}$ می‌باشد که برای N های بزرگ این مقدار به e^{-p} میل می‌کند. حال اگر متغیر تصادفی X را برابر با تعداد نمونه‌هایی که در نمونه‌گیری قرار ندارند بنامیم، خواهیم داشت:

$$\mathbf{E}[X] = \mathbf{E}[X_1 + X_2 + \cdots + X_N] = N\mathbf{E}[X_1] = N.e^{-p}$$

سوال ۲ حل.

۱۴ شهریور ۱۴۰۴

تعریف ۱ (مسئله اولیه SVM با حاشیه سخت) فرض کنید یک مجموعه داده خطی‌پذیر $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ داشته باشیم که در آن $\mathbf{x}_i \in \mathbb{R}^2$ و $y_i \in \{+1, -1\}$ باشد و حداقل یک نقطه از هر کلاس وجود داشته باشد. پارامترهای ابرصفحه بهینه $\mathbf{w} \in \mathbb{R}^2$ و $b \in \mathbb{R}$ با حل مسئله زیر به دست می‌آیند:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{to subject} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \text{برای} i = 1, 2, \ldots, N. \end{aligned}$$

یک **بردار پشتیبان** نقطه‌ای است که قید مربوط به آن فعال باشد، یعنی $y_i(\mathbf{w}^*\cdot\mathbf{x}_i + b^*) = 1$.

قضیه ۱ برای $N \geq 3$ نقطه در $\mathbb{R}^2$، پاسخ SVM با حاشیه سخت حداقل ۳ بردار پشتیبان دارد.

اثبات در سه بخش ارائه می‌شود: (۱) اثبات یکتایی بردار وزن بهینه $\mathbf{w}^*$، (۲) اثبات یکتایی بایاس بهینه b^*، و (۳) نشان دادن اینکه این یکتایی مستلزم وجود حداقل سه بردار پشتیبان است.

بخش ۱ (یکتایی $\mathbf{w}^*$) تابع هدف $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ **به شدت محدب** است زیرا هسین آن $\nabla^2 f(\mathbf{w}) = \mathbf{I}$ (ماتریس همانی) مثبت معین است. قیود نیز همگی خطی هستند و بنابراین یک ناحیه شدنی محدب را تعریف می‌کنند. یک تابع به شدت محدب که روی یک مجموعه محدب کمینه شود، یک **کمینه‌سراسری یکتا** دارد. در نتیجه، $\mathbf{w}^*$ یکتا است.

بخش ۲ (یکتایی b^*) فرض کنید برای تناقض، دو جواب بهینه متفاوت وجود داشته باشند: $(\mathbf{w}^*, b_1)$ و $(\mathbf{w}^*, b_2)$ به طوری که $b_1 \neq b_2$. هر دو باید تمام قیود را ارضا کنند:

$$\begin{aligned} y_i(\mathbf{w}^*\cdot\mathbf{x}_i + b_1) \geq 1 \quad \forall i \in \{1, \ldots, N\}, \\ y_i(\mathbf{w}^*\cdot\mathbf{x}_i + b_2) \geq 1 \quad \forall i \in \{1, \ldots, N\}. \end{aligned}$$

با تفاضل قید اول از دوم برای هر نقطه دلخواه i داریم:

$$\begin{aligned} y_i(\mathbf{w}^*\cdot\mathbf{x}_i + b_1) - y_i(\mathbf{w}^*\cdot\mathbf{x}_i + b_2) \geq 0, \\ y_i(b_1 - b_2) \geq 0. \end{aligned}$$

این نامساوی باید برای تمام نقاط برقرار باشد. با تحلیل بر اساس برچسب کلاس:

- برای هر i با $y_i = +1$: $\cdot\, 1\cdot(b_1 - b_2) \geq 0 \implies b_1 \geq b_2$.
- برای هر i با $y_i = -1$: $\cdot\, -1\cdot(b_1 - b_2) \geq 0 \implies b_1 \leq b_2$.

تنها مقداری که بتواند همزمان $b_1 \geq b_2$ و $b_1 \leq b_2$ را برای تمام نقاط (با فرض وجود حداقل یک نقطه از هر کلاس) برآورده کند، $b_1 = b_2$ است. این با فرض اولیه در تناقض است. بنابراین، b^* نیز باید یکتا باشد.

بخش ۳ (لزوم وجود حداقل سه بردار پشتیبان) شرایط KKT برای جواب بهینه $(\mathbf{w}^*, b^*)$ بیان می‌کند که:

(آ) $\mathbf{w}^*$ را می‌توان به صورت ترکیب خطی بردارهای پشتیبان نوشت: $\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$، که در آن $\alpha_i > 0$ تنها اگر $\mathbf{x}_i$ یک بردار پشتیبان باشد.

(ب) برای هر بردار پشتیبان $\mathbf{x}_j$، تساوی زیر برقرار است: $y_j(\mathbf{w}^* \cdot \mathbf{x}_j + b^*) = 1$.

معادله مربوط به یک بردار پشتیبان را می‌توان به صورت زیر بازنویسی کرد:

$$y_j b^* + w_1^*(x_j)_1 + w_2^*(x_j)_2 = 1, \qquad (1)$$

این یک **معادله خطی** با سه مجهول است: b^*، w_1^*، و w_2^*.
جواب یکتا $(\mathbf{w}^*, b^*)$ کاملاً توسط مجموعه معادلاتی از نوع (۱) که توسط بردارهای پشتیبان ارائه می‌شوند، تعیین می‌گردد. برای تعیین یکتای سه مجهول، یک دستگاه معادلات نیاز به حداقل سه معادله مستقل خطی دارد. از آنجایی که داده‌ها عمومی هستند و نقاطی از هر دو کلاس وجود دارند، حداقل تعداد بردارهای پشتیبان مورد نیاز برای تشکیل یک دستگاه با رتبه کامل، **سه** عدد است. داشتن کمتر از سه بردار پشتیبان منجر به یک دستگاه معادلات تحت‌تعیین می‌شود که با یکتایی $(\mathbf{w}^*, b^*)$ که در بخش‌های ۱ و ۲ اثبات شد، در تناقض است.

بنابراین، جواب SVM با حاشیه سخت باید حداقل سه بردار پشتیبان داشته باشد.

سوال ۳ حل. رابطه ی بین این ۳ مقدار به صورت زیر است:

$$\begin{aligned} nT(X) &= \sum_{i=1}^{n} \|x_i - \hat{x}\|^2 \\ &= \sum_{i=1}^{n} \|x_i - \hat{x}\|^2 = \sum_{i=1}^{n} \|x_i - \mu_j + \mu_j - \hat{x}\|^2 \\ &= \sum_{i=1}^{n} \|x_i - \mu_j\|^2 + \sum_{i=1}^{n} \|\mu_j - \hat{x}\|^2 + \sum_{i=1}^{n} (x_i - \mu_j)^T (\mu_j - \hat{x}) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2 + \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|\mu_j - \hat{x}\|^2 + \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)^T (\mu_j - \hat{x}) \end{aligned}$$

حال داریم :

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2 = \sum_{j=1}^{k} n_j W_j(X) \quad \text{where} n_j = \sum_{i=1}^{n} \gamma_{ij}$$

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

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}} \quad \Longrightarrow \quad \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)^T (\mu_j - \hat{x}) = 0$$

پس داریم :

$$nT(X) = \sum_{j=۱}^{k} n_j W_j(X) + nB(X)$$

چون $T(X)$ ثابت است الگوریتم k-means $W(X)$ را کمینه و در نتیجه $B(X)$ را بیشینه می کند.

سوال ۴ حل.

ابتدا محاسبات مربوط به هر حالت را انجام می‌دهیم:

مقدار حالت $S_۲$:

$$V(S_۲) = ۱ + ۰/۹ \cdot V(S_۲) \Rightarrow V(S_۲) = ۱۰$$

مقدار حالت $S_۴$:

$$V(S_۴) = -۵ + ۰/۹ \cdot \max\left(V(S_۲),\ ۰/۲V(S_۴) + ۰/۸V(S_۲)\right)$$

از آن‌جا که $V(S_۲) = ۱۰$:

$$V(S_۴) = -۵ + ۰/۹ \cdot ۱۰ = ۴$$

مقدار حالت $S_۳$:

اکشن اول:

$$V(S_۳) = -۱ + ۰/۹(۰/۲V(S_۴) + ۰/۸V(S_۲)) = -۱ + ۰/۹(۰/۲ \cdot ۴ + ۰/۸ \cdot ۱۰) = -۱ + ۰/۹(۸/۸) = ۶/۹۲$$

اکشن دوم:

$$V(S_۳) = -۱ + ۰/۹(۰/۲V(S_۳) + ۰/۸V(S_۴)) \Rightarrow V(S_۳) = -۱ + ۰/۱۸V(S_۳) + ۰/۹ \cdot ۴ = -۱ + ۰/۱۸V(S_۳) + ۳/۶$$

$$\Rightarrow V(S_۳) = ۰/۱۸V(S_۳) + ۲/۶ \Rightarrow V(S_۳) < ۳$$

پس بهترین اکشن، اولی است:

$$V(S_۳) = ۶/۹۲$$

مقدار حالت $S_۱$:

اکشن اول:

$$V(S_۱) = ۰ + ۰/۹(۰/۴V(S_۱) + ۰/۱V(S_۲) + ۰/۵V(S_۳)) = ۰/۹(۰/۴V(S_۱) + ۱ + ۳/۴۶) = ۰/۳۶V(S_۱) + ۴/۰۱۴$$

حل این معادله:

$$V(S_۱) = ۰/۳۶V(S_۱) + ۴/۰۱۴ \Rightarrow ۰/۶۴V(S_۱) = ۴/۰۱۴ \Rightarrow V(S_۱) = \frac{۴/۰۱۴}{۰/۶۴} \approx ۶/۲۷$$

اکشن دوم:

$$V(S_۱) = ۰ + ۰/۹(۰/۲V(S_۲) + ۰/۸V(S_۳)) = ۰/۹(۲ + ۵/۵۳۶) = ۶/۷۸$$

پس مقدار بهینه:

$$V(S_۱) = ۶/۷۸$$

در صورتی که فرض کنیم پاداش در هنگام ورود به حالت مقصد (sj) دریافت می‌شود، مقادیر نهایی به شکل زیر خواهد بود:

$$\boxed{V(S_۱) = ۷/۵۳۶, \quad V(S_۲) = ۱۰, \quad V(S_۳) = ۸/۸, \quad V(S_۴) = ۱۰}$$

سوال ۵ حل.

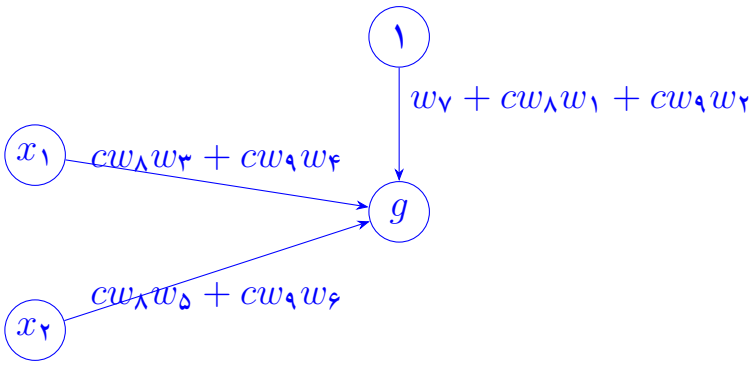
(آ)

$$= \frac{g(w_۷ + u_۸\, h(w_۱ + w_۳x_۱ + w_۵x_۲) + u_۹\, h(w_۲ + w_۴x_۱ + w_۶x_۲))}{۱ + \exp\Big(- \big(w_۷ + cw_۸y_۱ + cw_۹y_۲ + (cw_۸w_۴ + cw_۹w_۴)x_۱ + (cw_۸w_۵ + cw_۹w_۶)x_۲\big)\Big)}.$$

The classification boundary is:

$$w_۷ + cw_۸w_۱ + cw_۹w_۲ + (cw_۸w_۳ + cw_۹w_۴)x_۱ + (cw_۸w_۵ + cw_۹w_۶)x_۲ = ۰.$$

(ب)



(ج) خیر چون شبکه جدید معادل شبکه اولیه است.

(د) Yes. If linear activation functions are used for all the hidden units, the output from hidden units can be written as a linear combination of the inputs; since these intermediate outputs serve as input for the final layer, there is an equivalent network without hidden layers.

سوال ۶ حل.

(الف)

برای اینکه نشان دهیم فرم مجموع مربعات وزنی با فرم ماتریسی معادل است، ابتدا باید خطای پیش‌بینی برای تمام نمونه‌ها را در یک بردار نمایش دهیم. این بردار خطا $\mathbf{e}$ از تفاضل بین مقادیر پیش‌بینی شده $(\mathbf{X}\theta)$ و مقادیر واقعی $(\mathbf{y})$ به دست می‌آید:

$$\mathbf{e} = \mathbf{X}\theta - \mathbf{y} = \begin{pmatrix} \theta^T x^{(1)} - y^{(1)} \\ \theta^T x^{(2)} - y^{(2)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{pmatrix}$$

همانطور که مشخص است، هر درایه e_i از این بردار، همان خطای نمونه iـام است. در قدم بعدی، برای اینکه وزن‌های $w^{(i)}$ را وارد محاسبات ماتریسی کنیم، یک ماتریس قطری $\mathbf{W}$ می‌سازیم که درایه‌های قطر اصلی آن همان وزن‌ها هستند و بقیه درایه‌ها صفرند.

$$\mathbf{W} = \begin{pmatrix} w^{(1)} & 0 & \dots & 0 \\ 0 & w^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w^{(m)} \end{pmatrix}$$

حالا اگر عبارت ماتریسی $(\mathbf{X}\theta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\theta - \mathbf{y})$ را که معادل $\mathbf{e}^T\mathbf{W}\mathbf{e}$ است بسط دهیم، به نتیجه زیر می‌رسیم:

$$\begin{aligned} \mathbf{e}^T\mathbf{W}\mathbf{e} &= \begin{pmatrix} e_1 & \dots & e_m \end{pmatrix} \begin{pmatrix} w^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w^{(m)} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_m \end{pmatrix} \\ &= \begin{pmatrix} e_1 & \dots & e_m \end{pmatrix} \begin{pmatrix} w^{(1)}e_1 \\ \vdots \\ w^{(m)}e_m \end{pmatrix} \\ &= \sum_{i=1}^{m} w^{(i)} e_i^2 \end{aligned}$$

این عبارت نهایی دقیقاً همان فرم مجموع مربعات خطا است. بنابراین نشان دادیم که با تعریف ماتریس $\mathbf{W}$ به این شکل، دو عبارت با هم معادل هستند.

(ب)

در این قسمت، هدف ما پیدا کردن مقدار بهینه‌ی θ است که تابع هزینه $J(\theta)$ را کمینه کند. برای این کار، باید از تابع هزینه نسبت به θ گرادیان بگیریم و آن را برابر صفر قرار دهیم. تابع هزینه ما از قسمت قبل به این صورت است:

$$J(\theta) = \frac{1}{2}(\mathbf{X}\theta - \mathbf{y})^T\mathbf{W}(\mathbf{X}\theta - \mathbf{y})$$

برای سادگی در مشتق‌گیری، ابتدا این عبارت را بسط می‌دهیم. پس از ضرب کردن جملات در هم، به عبارت زیر می‌رسیم:

$$J(\theta) = \frac{1}{2}(\theta^T\mathbf{X}^T\mathbf{W}\mathbf{X}\theta - 2\theta^T\mathbf{X}^T\mathbf{W}\mathbf{y} + \mathbf{y}^T\mathbf{W}\mathbf{y})$$

توجه کنید که دو جمله میانی در بسط اولیه با هم برابر بودند و در اینجا با هم ترکیب شدند. حالا از این عبارت گرادیان می‌گیریم. با استفاده از قواعد مشتق‌گیری ماتریسی، گرادیان هر جمله محاسبه می‌شود:

$$\begin{aligned}\nabla_\theta J(\theta) &= \frac{۱}{۲}[۲(\mathbf{X}^T\mathbf{W}\mathbf{X})\theta - ۲\mathbf{X}^T\mathbf{W}\mathbf{y} + ۰] \\ &= \mathbf{X}^T\mathbf{W}\mathbf{X}\theta - \mathbf{X}^T\mathbf{W}\mathbf{y}\end{aligned}$$

برای یافتن نقطه کمینه، گرادیان را برابر صفر می‌گذاریم و معادله را برای θ حل می‌کنیم:

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\theta = \mathbf{X}^T\mathbf{W}\mathbf{y}$$

و در نهایت با فرض معکوس‌پذیر بودن ماتریس $\mathbf{X}^T\mathbf{W}\mathbf{X}$، پاسخ نهایی که به معادله نرمال وزنی معروف است، به دست می‌آید:

$$\theta = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-۱}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

این رابطه، فرم بسته برای یافتن پارامترهای بهینه در رگرسیون خطی وزنی است.

(ج)

در این بخش می‌خواهیم نشان دهیم که مسئله کمترین مربعات وزنی، از دیدگاه آماری با روش تخمین بیشینه درست‌نمایی (MLE) معادل است. فرض مسئله این است که هر $y^{(i)}$ از یک توزیع گوسی با میانگین $\theta^T x^{(i)}$ و واریانس $\sigma^{(i)۲}$ می‌آید. برای شروع، تابع درست‌نمایی (Likelihood) کل داده‌ها را می‌نویسیم که حاصل‌ضرب احتمال هر یک از نمونه‌هاست:

$$L(\theta) = \prod_{i=۱}^{m} \frac{۱}{\sqrt{۲\pi\sigma^{(i)۲}}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^۲}{۲\sigma^{(i)۲}}\right)$$

کار کردن با ضرب سخت است، بنابراین از لگاریتم این تابع (Log-Likelihood) استفاده می‌کنیم که ضرب را به جمع تبدیل می‌کند و ماکزیمم کردن آن معادل ماکزیمم کردن خود تابع اصلی است.

$$\begin{aligned}\ell(\theta) = \log L(\theta) &= \sum_{i=۱}^{m} \log\left[\frac{۱}{\sqrt{۲\pi\sigma^{(i)۲}}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^۲}{۲\sigma^{(i)۲}}\right)\right] \\ &= -\frac{۱}{۲}\sum_{i=۱}^{m} \log(۲\pi\sigma^{(i)۲}) - \frac{۱}{۲}\sum_{i=۱}^{m} \frac{۱}{\sigma^{(i)۲}}(y^{(i)} - \theta^T x^{(i)})^۲\end{aligned}$$

برای یافتن θ که این عبارت را ماکزیمم کند، متوجه می‌شویم که جمله اول (شامل لگاریتم) اصلاً به θ وابسته نیست و یک مقدار ثابت است. بنابراین، ماکزیمم کردن کل عبارت معادل کمینه کردن جمله دوم است:

$$\text{minimize}_\theta \left[\frac{۱}{۲}\sum_{i=۱}^{m} \frac{۱}{\sigma^{(i)۲}}(y^{(i)} - \theta^T x^{(i)})^۲\right]$$

این عبارت دقیقاً مشابه تابع هزینه رگرسیون وزنی $J(\theta)$ است. با مقایسه این دو، به راحتی می‌توان دریافت که وزن هر نمونه $w^{(i)}$ با معکوس واریانس آن نمونه برابر است:

$$w^{(i)} = \frac{۱}{\sigma^{(i)۲}}$$

این نتیجه بسیار جالب است و نشان می‌دهد نمونه‌هایی که واریانس کمتری دارند (و در نتیجه قابل اعتمادتر هستند)، وزن بیشتری در مدل‌سازی خواهند داشت.