دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز و زمستان ۱۴۰۴

استاد: علی شریفی زارچی

گردآورندگان: ارشیا یوسف‌نیا - بنیامین قنبری - زهرا رحمانی - علی باوفا - کیهان هدایی

تمرین اول | یادگیری با نظارت | مهلت ارسال: ۳۰ آبان

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم، برای هر تمرین می‌توانید تا ۵ روز تأخیر مجاز داشته باشید و در مجموع حداکثر ۱۵ روز تأخیر مجاز خواهید داشت. توجه داشته باشید که تأخیر در تمرین‌های عملی و تئوری به صورت جداگانه محاسبه می‌شود و مجموع تأخیر هر دو نباید بیشتر از ۱۵ روز شود. پس از اتمام زمان مجاز، دو روز اضافی برای آپلود غیرمجاز در نظر گرفته شده است که در این بازه، به ازای هر ساعت تأخیر، ۲ درصد از نمره تمرین کسر خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۱۰۰ نمره)

۱. (۱۰ نمره)

(آ) اگر داده‌ها خطی قابل تفکیک نباشند، هیچ راه‌حلی برای SVM به صورت soft-margin نداریم.

(ب) برای یک مجموعه داده که نویز زیادی دارد، یعنی تعداد زیادی از دادگان برچسب نادرست دارند، استفاده از جنگل تصادفی به طور کلی بهتر از boosted decision trees است.

(پ) در روش bagging در استفاده از درخت تصمیم‌گیری، معمولاً n درخت با واریانس زیاد و بایاس کمی داریم. انتخاب نهایی بر اساس اجماع آن‌ها با یکدیگر است که حاصل آن، نتیجه‌ای با بایاس کم و واریانس کم می‌شود.

(ت) در تخمین واریانس-بایاس، استفاده از یک مدل بسیار پیچیده بر روی داده محدود، معمولاً در ابتدا به بایاس کم و واریانس بالا منجر می‌شود.

(ث) الگوریتم perceptron تضمین می‌کند که برای هر مجموعه داده‌ای که به صورت خطی قابل تفکیک باشد، پس از تعداد محدودی گام، به یک راه‌حل همگرا می‌شود.

حل.

(آ) نادرست. اگر داده‌ها به صورت خطی جداپذیر نباشند، می‌توان با استفاده از حقه kernel و بردن داده‌ها به ابعاد بالاتر، آن‌ها را به صورت کامل از هم تفکیک کرد.

(ب) درست. جنگل تصادفی درختان زیادی را بر اساس نمونه‌های bootstrap آموزش می‌دهد و میانگین پیش‌بینی‌های آن‌ها را محاسبه می‌کند. این عمل باعث کاهش اثر خطاهای برچسب‌ها در بین درختان می‌شود، بنابراین این مجموعه در برابر نویز مقاوم است. در مقابل، روش‌های boosting وزن بیشتری به نمونه‌هایی که اشتباه پیش‌بینی شده‌اند اختصاص می‌دهند. برچسب‌های نویزی اغلب اشتباه پیش‌بینی می‌شوند و چون این مجموعه داده حاوی تعداد زیادی از این نوع برچسب‌ها است، باعث می‌شود مدل روی آن‌ها تمرکز کند و آن‌ها را بیش‌برازش کند.

(پ) درست. در bagging معمولا هر درخت واریانس زیادی دارد، اما چون نتیجه نهایی با اجماع به دست می‌آید، واریانس خروجی کاهش می‌یابد و تخمین پایدارتر می‌شود.

(ت) درست. استفاده از یک مدل با پیچیدگی بالا برای یک مجموعه داده محدود، منجر به کد شدن نویز به جای الگوی اصلی داده می‌شود که از آن با عنوان واریانس بالا نیز یاد می‌شود.

(ث) درست. در صورتی که داده‌ها به صورت خطی قابل تفکیک باشند، الگوریتم perceptron قادر است در تعداد محدودی گام، آن خط جداساز را پیدا کند.

۲. (۱۵ نمره) روش تخمین چگالی در knn برآوردی از چگالی یک نقطه از فضای ویژگی‌ها را نسبت به چگالی k همسایهٔ نزدیک‌تر می‌دهد. نشان دهید که مدل چگالی knn را می‌توان یک توزیع نامناسب دید. این موضوع چه اهمیتی دارد و چه نتیجه‌ای می‌دهد؟

توزیع نامناسب (improper distribution) توزیعی است که انتگرال‌گیری روی آن روی کل بازه واگرا خواهد شد.[۱]

راهنمایی: برای محاسبهٔ $\mathbb{P}(\mathbf{x}_i)$، کره‌ای کوچک با مرکزیت خود آن نقطه در نظر بگیرید و شعاع را تا دربرگرفتن k نقطه افزایش دهید. افزون بر این، کره را چنان کوچک فرض کنید که $\mathbb{P}(\mathbf{x}_i)$ در آن به تقریب ثابت باشد.

حل. هرگاه در kNN بخواهیم چگالی احتمالی را در نقطه‌ای مانند x_i محاسبه کنیم، کره‌ای کوچک در نظر می‌گیریم که نقطه مورد نظر در مرکز آن قرار داشته باشد و شعاع آن را تا جایی افزایش می‌دهیم که این کره k نقطه را در بر گیرد. در این صورت خواهیم داشت:

$$p(x_i) = \frac{k}{NV_i}$$

که N تعداد تمام مشاهدات و V_i حجم کره‌ای به مرکزیت x_i است. فرض می‌کنیم V_i آن‌قدر کوچک است که $p(x_i)$ در آن ثابت باقی بماند. اکنون خواهیم داشت:

$$\int p(x)\,dx \approx \sum_{i=1}^{N} p(x_i)\,V_i = \sum_{i=1}^{N} \frac{k}{NV_i}\,V_i = k \neq 1$$

معادله بالا برقرار خواهد بود اگر و تنها اگر حجم تمامی کره‌ها به اندازه کافی کوچک بوده و N به اندازه کافی بزرگ باشد. هر دو شرط گفته شده نیز در kNN برقرارند. از آن‌جایی که kNN محدودیتی بر روی توزیع داده اعمال نمی‌کند و تنها بر روی یک بخش محلی متمرکز می‌شود، همان‌گونه که مشاهده کردید می‌بینیم که تخمین چگالی از قوانین توزیع‌های احتمالاتی پیروی نمی‌کند. با توجه به این که انتگرال چگالی تخمین زده شده بر روی کل بازه برابر با یک نمی‌شود، این توزیع نامناسب خواهد بود. به همین علت تفسیر و به‌کارگیری kNN برای مواردی که در آن‌ها به توزیع‌های احتمالاتی نیاز داریم، دشوار خواهد بود.

۳. (۱۵ نمره) فرض کنید یک تابع رگرسیون $h(\mathbf{x})$ داریم که برای هر بردار ورودی $\mathbf{x}$ برچسب $y = h(\mathbf{x})$ را به آن نسبت می‌دهد. از دادگان آموزش استفاده کرده‌ایم و m مدل پیش‌بینی‌کننده آموزش داده‌ایم که آن‌ها را با $\hat{h}_1(\mathbf{x}), \hat{h}_2(\mathbf{x}), \ldots, \hat{h}_m(\mathbf{x})$ نشان می‌دهیم. اکنون قرار دهید:

$$\hat{H}_m(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} \hat{h}_i(\mathbf{x})$$

[۱] برای مطالعهٔ بیشتر، مراجعه کنید به: https://arxiv.org/abs/1711.02064

ثابت کنید که:

$$\mathbb{E}_{\mathbf{x}}\left[(\hat{H}_m(\mathbf{x}) - h(\mathbf{x}))^2\right] \leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathbf{x}}\left[(\hat{h}_i(\mathbf{x}) - h(\mathbf{x}))^2\right]$$

مفهوم این حکم چیست؟ حل. مقادیر E_{com} و E_{avg} را به صورت زیر تعریف می‌کنیم:

$$E_{com} = \mathbb{E}_x\big[(H_M(x) - h(x))^2\big], \qquad E_{avg} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_x\big[(y_i(x) - h(x))^2\big].$$

در این صورت، با تعریف $\epsilon_i(x) = y_i(x) - h(x)$، برای E_{com} خواهیم داشت:

$$\begin{aligned}
E_{com} &= \mathbb{E}_x\big[(H_M(x) - h(x))^2\big] \\
&= \mathbb{E}_x\big[\big(\tfrac{1}{M}\sum_{i=1}^{M} y_i(x) - h(x)\big)^2\big] \\
&= \mathbb{E}_x\big[\big(\tfrac{1}{M}\sum_{i=1}^{M} y_i(x) - \tfrac{M}{M}h(x)\big)^2\big] \\
&= \mathbb{E}_x\big[\big(\tfrac{1}{M}\sum_{i=1}^{M} (y_i(x) - h(x))\big)^2\big] \\
&= \mathbb{E}_x\big[\big(\tfrac{1}{M}\sum_{i=1}^{M} \epsilon_i(x)\big)^2\big] \\
&= \frac{1}{M^2}\,\mathbb{E}_x\Big[\Big(\sum_{i=1}^{M} \epsilon_i(x)\Big)^2\Big].
\end{aligned}$$

همچنین، با همان تعریف $\epsilon_i(x) = y_i(x) - h(x)$، برای E_{avg} داریم:

$$E_{avg} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_x\big[(y_i(x) - h(x))^2\big] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_x\big[(\epsilon_i(x))^2\big].$$

حال با استفاده از نامساوی کوشی–شوارتز به ازای هر x داریم:

$$\Big(\sum_{i=1}^{M}\epsilon_i(x)\Big)^2 \leq M\sum_{i=1}^{M}(\epsilon_i(x))^2.$$

طرفین را در $\frac{1}{M^2}$ ضرب می‌کنیم:

$$\frac{1}{M^2}\Big(\sum_{i=1}^{M}\epsilon_i(x)\Big)^2 \leq \frac{1}{M}\sum_{i=1}^{M}(\epsilon_i(x))^2.$$

اکنون از امیدریاضی نسبت به x می‌گیریم:

$$\mathbb{E}_x\Big[\frac{1}{M^2}\Big(\sum_{i=1}^{M}\epsilon_i(x)\Big)^2\Big] \leq \mathbb{E}_x\Big[\frac{1}{M}\sum_{i=1}^{M}(\epsilon_i(x))^2\Big].$$

بنابراین،

$$\frac{1}{M^2}\mathbb{E}_x\Big[\Big(\sum_{i=1}^{M}\epsilon_i(x)\Big)^2\Big] \le \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_x\big[(\epsilon_i(x))^2\big],$$

و در نتیجه

$$E_{com} \le E_{avg}.$$

۴. (۱۰ نمره) به سوالاتی که در ادامه آمده‌اند با بیان توضیحات کامل پاسخ دهید.

(آ) یک مدل Random Forest روی تعداد زیادی داده برای یک مسئلهٔ دسته‌بندی آموزش داده‌ایم. میزان خطای مدل روی دادگان آموزشی پایین است ولی روی دادگان آزمون خطای زیادی گزارش شده است. دو دلیل برای این مشکل بیان کنید و برای هر کدام راه‌حلی بگویید.

(ب) فرض کنید دادگان $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ با میانگین و واریانس μ, σ^2 داده شده‌اند. بگیرید:

$$\mathbf{z}_m = \frac{1}{m}\Big(\sum_{i=1}^{m}\mathbf{a}_i\Big)$$

که در آن $\mathbf{a}_i$ ها تصادفی و متمایز از $\mathbf{X}$ گزینش شده‌اند. میانگین و واریانس $\mathbf{z}_m$ را بیابید. اکنون بیان کنید مجموعه دادگان $\{\mathbf{z}_m\}$ چه خوبی و بدی‌هایی نسبت به $\mathbf{X}$ دارد و در چه شرایطی استفاده از آن برای آموزش مناسب است.

حل.

(آ) مورد اول: پیچیدگی بالای مدل جنگل تصادفی. این یعنی مدل ما به داده‌ها overfit می‌شود و خاصیت تعمیم‌پذیری ندارد.
راه‌حل: محدود کردن عمق و تعداد درخت‌ها تا واریانس کاهش یابد و افزایش حداقل تعداد نمونه لازم برای جدا کردن گره‌ها. این کارها باعث می‌شود مدل محدودیت بیشتری پیدا کند و از بیش‌برازش جلوگیری شود.

مورد دوم: وجود تعداد زیادی ویژگی نامناسب.
راه‌حل: استفاده از روش‌هایی مانند PCA و سایر روش‌های کاهش بُعد برای استخراج تعدادی ویژگی مفید و مهم و ادامهٔ کار تنها بر روی آن‌ها.

مورد سوم: تعداد داده‌ها کم است یا کل دامنهٔ رفتار را پوشش نمی‌دهند.
راه‌حل: اضافه کردن داده‌ها از سراسر دامنهٔ ممکن (گردآوری دادهٔ بیشتر و متنوع‌تر).

(ب) فرض کنیم داده‌ها از مجموعه‌ی $X = \{x_i\}_{i=1}^{n}$ با میانگین μ و واریانس σ^2 انتخاب شده‌اند. آماره‌ی z_m به صورت زیر تعریف می‌شود:

$$z_m = \frac{1}{m}\sum_{i=1}^{m}a_i,$$

که در آن a_i نمونه‌هایی تصادفی و متمایز (نمونه‌گیری بدون جایگذاری) از X هستند.
میانگین آماره‌ی z_m برابر با میانگین جامعه μ است، زیرا z_m یک برآوردگر ناریب برای μ است:

$$\mathbb{E}[z_m] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}a_i\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[a_i] = \frac{1}{m}\sum_{i=1}^{m}\mu = \mu.$$

واریانس آماره‌ی z_m در نمونه‌گیری تصادفی ساده بدون جایگذاری از یک جامعه‌ی متناهی با اندازه‌ی n از رابطهٔ زیر به دست می‌آید:

$$\mathrm{Var}(z_m) = \frac{\sigma^2}{m}\left(\frac{n-m}{n-1}\right),$$

که در آن عامل $\frac{n-m}{n-1}$ به عنوان «عامل تصحیح جامعه‌ی متناهی» (FPC) شناخته می‌شود.

مزایا و معایب مجموعه‌ی داده‌های $\{z_m\}$ نسبت به X:

- **واریانس کاهش‌یافته (دقت بالاتر):** واریانس $\mathrm{Var}(z_m)$ به دلیل تقسیم بر اندازه‌ی نمونه m و اعمال FPC بسیار کمتر از واریانس σ^2 داده‌های اصلی X است؛ بنابراین z_m برآوردگر دقیق‌تری برای μ محسوب می‌شود.
- **ناریبی:** چون $\mathbb{E}[z_m] = \mu$، به طور متوسط این آماره دقیقاً برابر با مقدار واقعی میانگین جامعه است.
- **از دست دادن جزئیات:** z_m تنها میانگین نمونه را نشان می‌دهد و اطلاعات مربوط به پراکندگی و ویژگی‌های منحصربه‌فرد (مانند داده‌های پرت) تک‌تک نقاط در X را نادیده می‌گیرد.

در مجموع، آماره‌های $\{z_m\}$ که میانگین نمونه‌های mـتایی هستند، به دلیل ناریب بودن و واریانس پایین (که با افزایش m کاهش می‌یابد)، برآوردگرهای بسیار خوبی برای میانگین جامعه μ به شمار می‌آیند. این ویژگی آن‌ها را برای استفاده در فرایندهای آموزشی مانند Mini-Batch SGD مناسب می‌سازد، زیرا استفاده از آن‌ها نویز محاسباتی (واریانس گرادیان‌ها) را کاهش داده و به همگرایی سریع‌تر و پایدارتر بهینه‌ساز کمک می‌کند؛ مشروط بر این‌که اندازه‌ی m به‌قدر کافی بزرگ انتخاب شود تا نمونه نماینده‌ی خوبی از جامعه باشد.

۵. (۱۰ نمره) به سوالاتی که در ادامه آمده‌اند با بیان توضیحات کامل پاسخ دهید.

(آ) با توجه به تابع سیگموید به سوالات زیر پاسخ دهید.

- مشتق تابع سیگموید را بدست آورید (تمامی مراحل به طور کامل نوشته شود).
- تابع سیگموید و مشتق آن را رسم کنید. از نمودار بدست آمده کمک بگیرید و به طور شهودی توضیح دهید که مشکل تابع سیگموید در بروز رسانی وزن‌ها با کاهش گرادیان چیست.

(ب) چرا رگرسیون لجستیک که با الگوریتم کاهش گرادیان آموزش داده می‌شود، بعضی وقت‌ها که دادگان تقریباً به طور خطی قابل تفکیک هستند، کند همگرا می‌شود؟

(ج) در یک مسئلهٔ طبقه‌بندی چندکلاسه با K کلاس، به‌جای استفاده از چندین مدل دودویی، می‌توان از یک مدل چندکلاسه استفاده کرد. نشان دهید که اگر به همهٔ بردارهای پارامتر $\theta_1, \theta_2, \ldots, \theta_K$ یک بردار ثابت $\mathbf{c}$ اضافه کنیم، مقادیر احتمالات تغییری نمی‌کنند.

حل.

(آ) **(بخش اول)**

$$y = \sigma(x) = \frac{1}{1+e^{-x}}$$

$$u = 1 + e^{-x}, \qquad y = \frac{1}{u}$$

$$\frac{du}{dx} = -e^{-x}, \qquad \frac{dy}{du} = -\frac{1}{u^2}$$

با استفاده از قاعدهٔ زنجیره‌ای داریم:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = \left(-\frac{۱}{u^۲}\right)\left(-e^{-x}\right) = \frac{e^{-x}}{u^۲} = \frac{e^{-x}}{(۱+e^{-x})^۲}.$$

از طرفی:

$$\sigma(x) = \frac{۱}{۱+e^{-x}}, \qquad ۱ - \sigma(x) = \frac{e^{-x}}{۱+e^{-x}},$$

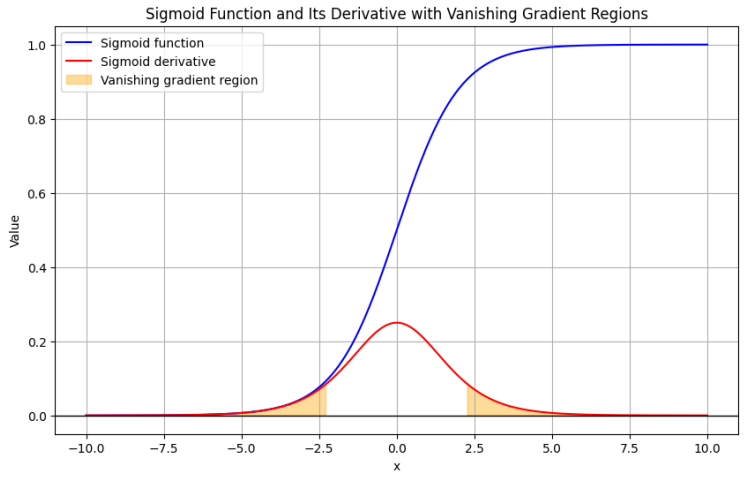
پس می‌توان نوشت:

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(۱+e^{-x})^۲} = \sigma(x)\big(۱ - \sigma(x)\big).$$

در نتیجه:

$$\sigma'(x) = \sigma(x)\big(۱ - \sigma(x)\big).$$

(بخش دوم)



ناحیهٔ نارنجی‌رنگ نشان‌دهندهٔ بخش‌هایی است که مشتق $\sigma'(x)$ بسیار کوچک (نزدیک به صفر) است. در این نواحی، گرادیان‌هایی که برای به‌روزرسانی وزن‌ها و بایاس‌ها استفاده می‌شوند بسیار کوچک می‌شوند، در نتیجه مدل بسیار آهسته یاد می‌گیرد یا حتی یادگیری تقریباً متوقف می‌شود.

(ب) وقتی داده‌ها تقریباً به صورت خطی قابل‌تفکیک باشند، لجستیک رگرسیون سعی می‌کند پارامترهایی پیدا کند که دو کلاس را به طور کامل از هم جدا کنند. برای رسیدن به این هدف، مقدار وزن‌ها به‌طور مداوم بزرگ‌تر می‌شود تا احتمال‌های پیش‌بینی‌شده برای نمونه‌های هر کلاس تا حد ممکن به ۰ یا ۱ نزدیک شوند.

اما با بزرگ شدن وزن‌ها، (Sigmoid) سیگموید تابع در نواحی اشباع خود قرار می‌گیرد؛ یعنی خروجی آن حتی با تغییرات بزرگ در وزن‌ها، تغییر اندکی می‌کند. در این وضعیت، گرادیان‌ها بسیار کوچک می‌شوند و هر گام الگوریتم (Gradient Descent) گرادیان کاهش فقط یک به‌روزرسانی بسیار کوچک در پارامترها ایجاد می‌کند.

به همین دلیل، فرآیند آموزش ظاهراً «گیر می‌کند» و بسیار کند پیش می‌رود، در حالی که در واقع مدل به‌آرامی در جهت وزن‌های بسیار بزرگ حرکت می‌کند. افزودن (Regularization) رگولاریزیشن با محدود کردن رشد وزن‌ها، از این رفتار جلوگیری کرده و به مدل کمک می‌کند سریع‌تر و پایدارتر همگرا شود.

(ج) اگر برای همهٔ کلاس‌ها $\theta'_j = \theta_j + c$ (برای یک بردار ثابت c) داشته باشیم، آنگاه:

$$e^{\theta_j'^T x} = e^{(\theta_j + c)^T x} = e^{c^T x} e^{\theta_j^T x}.$$

در نتیجه در مدل Softmax داریم:

$$P'(y=k) = \frac{e^{\theta_k'^T x}}{\sum_j e^{\theta_j'^T x}} = \frac{e^{c^T x} e^{\theta_k^T x}}{e^{c^T x} \sum_j e^{\theta_j^T x}} = \frac{e^{\theta_k^T x}}{\sum_j e^{\theta_j^T x}} = P(y=k).$$

بنابراین اضافه کردن یک بردار ثابت یکسان به همهٔ θ_jها، احتمال‌های خروجی را تغییر نمی‌دهد و تنها یک افزونگی در پارامترها ایجاد می‌کند؛ یعنی مدل از نظر آماری معادل باقی می‌ماند.

۶. (۱۵ نمره) به سوالاتی که در ادامه آمده‌اند با بیان توضیحات کامل پاسخ دهید.

(آ) برای یک مجموعه binary classification به صورت $(\mathbf{x}_i, y_i)$ که $y_i \in \{+۱, -۱\}$ مسئله hard-margin[۲] را در نظر بگیرید که به صورت مسئله بهینه‌سازی

$$\min \left\{ \frac{۱}{۲} \|\mathbf{w}\|^۲ \right\} \quad s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq ۱$$

تعریف می‌شود.

- توضیح دهید چرا Geometric Margin مربوط به این classifier برابر با

$$\gamma = \min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

است.

- چرا با کاهش دادن $\frac{۱}{۲}\|\mathbf{w}\|^۲$ با توجه به شروط داده شده، این margin افزایش می‌یابد؟

(ب) فرض کنید مدل Linear Regression را داریم:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(۰, \sigma \mathbf{I})$$

- فرم بسته‌ای برای تخمین‌گر least squares برای پارامتر $\boldsymbol{\beta}$ به دست آورید.
- نشان دهید که این تخمین‌گر unbiased است و ماتریس covariance آن را نیز به دست آورید.

حل.

(آ) بخش اول

می‌دانیم Geometric Margin برای نقطه‌ی i برابر با فاصله‌ی آن نقطه از ابرصفحه‌ی classifier است. با توجه به این‌که مقدار $w^T x + b$ می‌تواند مثبت یا منفی باشد و $y_i \in \{-۱, +۱\}$، به صورت زیر عمل می‌کنیم:

- نقطه‌ی $x_.$ را روی ابرصفحه در نظر می‌گیریم:

$$w^T x_. + b = ۰.$$

- بردار از $x_.$ به x_i برابر است با:

$$x_i - x_.$$

[۲] https://en.wikipedia.org/wiki/Support_vector_machine

- برای به‌دست‌آوردن فاصله از ابرصفحه، تصویر این بردار را روی بردار واحد $\frac{w}{\|w\|}$ محاسبه می‌کنیم:

$$\gamma_i = \left| \left\langle \frac{w}{\|w\|}, x_i - x_\cdot \right\rangle \right| = \left| \frac{w^T(x_i - x_\cdot)}{\|w\|} \right|.$$

- با توجه به این‌که $w^T x_\cdot + b = ۰$، داریم:

$$\gamma_i = \left| \frac{w^T x_i + b}{\|w\|} \right|.$$

از آن‌جا که $w^T x_i + b$ هم‌علامت با y_i است، می‌توان نوشت:

$$\gamma_i = \frac{y_i\,(w^T x_i + b)}{\|w\|}.$$

در نهایت، Geometric Margin برای یک مجموعه‌ی داده به صورت کمینه‌ی marginهای نقاط تعریف می‌شود:

$$\gamma = \min_i \left\{ \frac{y_i\,(w^T x_i + b)}{\|w\|} \right\}.$$

شرط مسئله‌ی بهینه‌سازی داده‌شده برابر است با:

$$y_i(w^T x_i + b) \geq ۱.$$

بنابراین:

$$\forall i:\ y_i(w^T x_i + b) \geq ۱ \implies \min_i \{y_i(w^T x_i + b)\} \geq ۱,$$

و در نتیجه:

$$\gamma = \frac{۱}{\|w\|} \min_i \{y_i(w^T x_i + b)\} \geq \frac{۱}{\|w\|}.$$

بخش دوم

هدف مسئله‌ی بهینه‌سازی مینیمم کردن $\frac{۱}{۲}\|w\|^۲$ است. این کار معادل با کوچک‌کردن $\|w\|$ و در نتیجه بزرگ‌کردن $\frac{۱}{\|w\|}$ است. با توجه به این‌که $\gamma \geq \frac{۱}{\|w\|}$، کمینه کردن $\|w\|$ در حضور قیدهای

$$y_i(w^T x_i + b) \geq ۱$$

منجر به حداکثر شدن Geometric Margin روی مجموعه داده می‌شود. پس با حل این مسئله‌ی بهینه‌سازی، ابرصفحه‌ای را می‌یابیم که Geometric Margin آن حداکثر است.

(ب) **بخش اول**

ابتدا تابع loss (تابع هزینه) را به‌صورت زیر در نظر می‌گیریم:

$$J(\beta) = (y - X\beta)^T (y - X\beta).$$

آن را بسط می‌دهیم:

$$J(\beta) = y^T y - ۲\beta^T X^T y + \beta^T X^T X \beta.$$

اکنون نسبت به β مشتق می‌گیریم:

$$\frac{\partial J}{\partial \beta} = -۲X^T y + ۲X^T X\beta.$$

شرط بهینگی $\frac{\partial J}{\partial \beta} = ۰$ را اعمال می‌کنیم:

$$-۲X^T y + ۲X^T X\hat{\beta} = ۰ \implies X^T X\hat{\beta} = X^T y \implies \hat{\beta} = (X^T X)^{-۱} X^T y.$$

بخش دوم

برای نشان‌دادن این‌که $\hat{\beta}$ یک برآوردگر unbiased است، باید نشان دهیم:

$$\mathbb{E}[\hat{\beta}] = \beta.$$

مدل را به صورت $y = X\beta + \epsilon$ می‌نویسیم، که در آن $\mathbb{E}[\epsilon] = ۰$ و $\mathrm{Var}(\epsilon) = \sigma^۲ I$ است. آنگاه:

$$\hat{\beta} = (X^T X)^{-۱} X^T y = (X^T X)^{-۱} X^T (X\beta + \epsilon) = (X^T X)^{-۱} X^T X\beta + (X^T X)^{-۱} X^T \epsilon.$$

از امیدریاضی می‌گیریم:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\big[(X^T X)^{-۱} X^T X\beta\big] + \mathbb{E}\big[(X^T X)^{-۱} X^T \epsilon\big] = \beta + (X^T X)^{-۱} X^T \mathbb{E}[\epsilon] = \beta.$$

پس:

$$\mathbb{E}[\hat{\beta}] = \beta \quad \implies \quad \hat{\beta} \text{ است unbiased.}$$

واریانس $\hat{\beta}$ نیز به صورت زیر به دست می‌آید:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\big((X^T X)^{-۱} X^T \epsilon\big) = (X^T X)^{-۱} X^T \mathrm{Var}(\epsilon)\, X (X^T X)^{-۱}.$$

با توجه به $\mathrm{Var}(\epsilon) = \sigma^۲ I$ داریم:

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-۱} X^T (\sigma^۲ I)\, X (X^T X)^{-۱} = \sigma^۲ (X^T X)^{-۱} X^T X (X^T X)^{-۱} = \sigma^۲ (X^T X)^{-۱}.$$

۷. (۱۰ نمره) همان‌طور که می‌دانید اعتبارسنجی متقاطع[۳] با حذف یک نمونه یا LOOCV را می‌توان بدین گونه نشان داد:

$$e(S_n) = \frac{۱}{n} \sum_{i=۱}^{n} \big(y_i - \hat{f}_{-i}(\mathbf{x}_i)\big)^۲$$

که در آن $S_n = \{(\mathbf{x}_۱, y_۱), \ldots, (\mathbf{x}_n, y_n)\}$ (مجموعه داده‌های آموزش)، $S_n^{-i} = S_n \setminus \{(\mathbf{x}_i, y_i)\}$ (مجموعه داده‌ها پس از حذف نمونهٔ iام)، و $\hat{f}_{-i}$ مدل آموزش‌دیده روی S_n^{-i} می‌باشند.

(آ) با فرض اینکه فقط نمونهٔ اول حذف شده است، خطا به شکل

$$e_۱(S_n) = \big(y_۱ - \hat{f}_{-۱}(\mathbf{x}_۱)\big)^۲$$

است. نشان دهید:

$$\mathbb{E}\big[e_۱(S_n)\big] = \mathbb{E}\Big[\big(y - \hat{f}_{S_{n-۱}}(\mathbf{x})\big)^۲\Big]$$

که در آن $(\mathbf{x}, y)$ یک نمونهٔ تصادفی از توزیع داده‌هاست و $\hat{f}_{S_{n-۱}}$ مدل آموزش‌دیده روی $S_{n-۱}$ است.

(ب) با استفاده از نتیجهٔ بالا، نشان دهید که:

$$\mathbb{E}\big[e(S_n)\big] = \mathbb{E}\Big[\big(y - \hat{f}_{S_{n-۱}}(\mathbf{x})\big)^۲\Big]$$

[۳] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

حل.

(آ) اگر A و B متغیرهای تصادفی (RVs) با توزیع احتمال یکسان باشند، آنگاه $\mathbb{E}[f(A)] = \mathbb{E}[f(B)]$ خواهد بود. این امر را می‌توان با نوشتن انتگرال متناظر، روشن‌تر بیان کرد:

$$\mathbb{E}[f(A)] = \int f(x)\, p_A(x)\, dx = \int f(x)\, p_B(x)\, dx = \mathbb{E}[f(B)].$$

این حکم زمانی که A و B خود مجموعه‌ای از متغیرهای تصادفی باشند نیز برقرار است. به‌ویژه، اگر

$$A = \{S_n^{-1}, (\mathbf{x}_1, y_1)\}$$

باشد، به این معنا که روی S_n^{-1} آموزش می‌دهیم و روی $(\mathbf{x}_1, y_1)$ آزمون می‌گیریم، و این مجموعه همان توزیع

$$B = \{S_{n-1}, (\mathbf{x}, y)\}$$

را داشته باشد، که در آن روی مجموعه‌ی دیگری از $n-1$ نمونه به نام S_{n-1} آموزش می‌دهیم و روی $(\mathbf{x}, y)$ آزمون می‌گیریم (و همه‌ی این‌ها از یک توزیع زیربنایی مشترک نمونه‌برداری شده‌اند)، آنگاه همچنان خواهیم داشت:

$$\mathbb{E}[f(A)] = \mathbb{E}[f(B)].$$

(ب) این نتیجه اساساً از قسمت (آ) و با استفاده از خاصیت خطی بودن امید ریاضی به‌دست می‌آید. ابتدا توجه می‌کنیم که استدلال قسمت (آ) به حالت کلی زیر تعمیم می‌یابد:

$$\text{error}_1(S_n) = \text{error}_2(S_n) = \cdots = \text{error}_i(S_n) = \cdots = \text{error}_n(S_n),$$

که در آن $\text{error}_i(S_n)$ خطای مدلی است که روی $S_n \setminus \{(\mathbf{x}_i, y_i)\}$ یاد گرفته شده و روی $(\mathbf{x}_i, y_i)$ ارزیابی شده است.
بنابراین برای امید ریاضی خطای LOOCV خواهیم داشت:

$$\mathbb{E}\big[\text{error}_{\text{LOOCV}}(S_n)\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[(y_i - \hat{f}_{-i}(\mathbf{x}_i))^2\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\text{error}_i(S_n)\big] = \mathbb{E}\big[\text{error}_1(S_n)\big].$$

در برابری اول، از قانون خطی بودن امید ریاضی استفاده کرده‌ایم تا عملگر $\mathbb{E}$ را به داخل علامت جمع منتقل کنیم.

۸. (۱۵ نمره) جدول دادگان زیر را برای یک مسئلهٔ دسته‌بندی غذا در نظر بگیرید:

دما	اندازه	مزه	خوشایندی
گرم	کوچک	شور	خیر
سرد	بزرگ	شیرین	خیر
سرد	بزرگ	شیرین	خیر
سرد	کوچک	ترش	بله
گرم	کوچک	ترش	بله
گرم	بزرگ	شور	خیر
گرم	بزرگ	ترش	بله
سرد	کوچک	شیرین	بله
سرد	کوچک	شیرین	بله
گرم	بزرگ	شور	خیر

(آ) آنتروپی[۴] اولیه ستون مربوط به خوشایند بودن غذا چند است؟

(ب) فرض کنید ویژگی مزه را ریشهٔ درخت گرفته‌ایم، اکنون information gain[۵] مربوط به این ویژگی را مشخص کنید.

(پ) درخت تصمیم کامل را برای اجرای پیش‌بینی رسم کنید و همه محاسبات و مراحل را بیاورید.

(ت) اگر یک غذا شیرین، کوچک، و سرد باشد، پیش‌بینی ما با درخت قسمت قبل چه خواهد بود؟

(ث) یک نمونه bootstrap[۶] از دادگان جدول گرفته و درخت تصمیم‌گیری را دوباره بسازید.

حل.

(آ)

$$p(\text{خوشایند}) = \frac{۵}{۱۰} = \frac{۱}{۲}, \qquad p(\text{ناخوشایند}) = \frac{۵}{۱۰} = \frac{۱}{۲}$$

با فرض استفاده از لگاریتم مبنای ۲ داریم:

$$H(y) = -\left(\tfrac{۱}{۲}\log\tfrac{۱}{۲} + \tfrac{۱}{۲}\log\tfrac{۱}{۲}\right) = -۲\left(\tfrac{۱}{۲}\log\tfrac{۱}{۲}\right) = \log ۲ = ۱.$$

(ب)

$$\mathrm{IG}(y, m) = H(y) - H(y \mid m).$$

برای ویژگی «مزه» داریم:

$$p(\text{شور}) = \frac{۳}{۱۰} \quad \Rightarrow \quad \mathbb{P}(\text{ناخوشایند} \mid \text{شور}) = ۱,$$

$$p(\text{شیرین}) = \frac{۴}{۱۰} \quad \Rightarrow \quad \mathbb{P}(\text{خوشایند} \mid \text{شیرین}) = \frac{۱}{۲},$$

$$p(\text{ترش}) = \frac{۳}{۱۰} \quad \Rightarrow \quad \mathbb{P}(\text{خوشایند} \mid \text{ترش}) = ۱.$$

(توجه: بر طبق قرارداد، $۰ \log ۰ = ۰$ در نظر گرفته می‌شود.)
بنابراین انتروپی شرطیِ برحسب مزه برابر است با:

$$H(y \mid m) = -\frac{۳}{۱۰}[۰] - \frac{۴}{۱۰}\left(\tfrac{۱}{۲}\log\tfrac{۱}{۲} + \tfrac{۱}{۲}\log\tfrac{۱}{۲}\right) - \frac{۳}{۱۰}[۰] = -\frac{۴}{۱۰}\log\tfrac{۱}{۲} = \frac{۴}{۱۰} = \frac{۲}{۵}.$$

پس:

$$\mathrm{IG}\big(y(\text{خوشایندی}),\, m(\text{مزه})\big) = H(y) - H(y \mid m) = ۱ - \frac{۲}{۵} = \frac{۳}{۵} = ۰/۶.$$

(پ) برای انتخاب بهترین ویژگی به‌عنوان ریشه، IG همهٔ ویژگی‌ها را محاسبه می‌کنیم و ویژگی با بیشترین IG را به‌عنوان ریشهٔ درخت تصمیم برمی‌گزینیم.

اطلاعات متقابلِ ویژگی اندازه:

$$H(y \mid \text{اندازه}) = -\frac{۱}{۲}\left(\frac{۴}{۵}\log\frac{۴}{۵} + \frac{۱}{۵}\log\frac{۱}{۵}\right) - \frac{۱}{۲}\left(\frac{۴}{۵}\log\frac{۴}{۵} + \frac{۱}{۵}\log\frac{۱}{۵}\right).$$

[۴] https://en.wikipedia.org/wiki/Entropy
[۵] https://en.wikipedia.org/wiki/Information_gain_(decision_tree)
[۶] https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

بنابراین:

$$H(y \mid \text{اندازه}) = \frac{۴}{۵} \log \frac{۵}{۴} + \frac{۱}{۵} \log ۵ \approx ۰/۷۲۱۹,$$
$$\text{IG}(y, \text{اندازه}) = H(y) - H(y \mid \text{اندازه}) \approx ۱ - ۰/۷۲۱۹ = ۰/۲۷۸۱.$$

اطلاعات متقابلِ ویژگی دما:

$$H(y \mid \text{دما}) = -\frac{۱}{۲} \left(\frac{۳}{۵} \log \frac{۳}{۵} + \frac{۲}{۵} \log \frac{۲}{۵} \right) \times ۲,$$

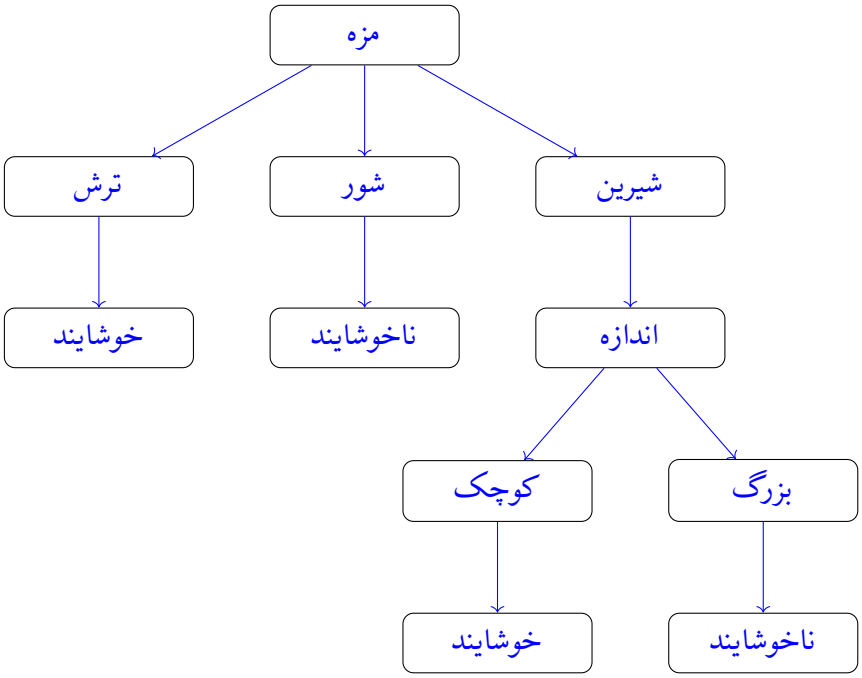
پس:

$$H(y \mid \text{دما}) = \frac{۳}{۵} \log \frac{۵}{۳} + \frac{۲}{۵} \log \frac{۵}{۲} \approx ۰/۹۷۱,$$
$$\text{IG}(y, \text{دما}) = H(y) - H(y \mid \text{دما}) \approx ۱ - ۰/۹۷۱ = ۰/۰۲۹.$$

از بخش (ب) داشتیم:

$$\text{IG}(y, \text{مزه}) = ۰/۶.$$

بنابراین بهترین ویژگی برای ریشه، ویژگی **مزه** است و درخت تصمیم کامل به صورت زیر است:



(ت) بر اساس درخت تصمیم کامل رسم‌شده در قسمت (پ)، پیش‌بینی برای نمونهٔ داده‌شده برابر است با:

خوشایند.

(ث) مجموعه‌دادهٔ bootstrap (نمونه‌گیری با جایگذاری) به صورت زیر است:
درخت تصمیم به‌دست‌آمده از این مجموعهٔ bootstrap:

اندازه	مزه	دما	خوشایند بودن غذا
کوچک	شور	گرم	خیر
بزرگ	شیرین	سرد	خیر
بزرگ	شیرین	سرد	خیر
بزرگ	شیرین	سرد	خیر
کوچک	ترش	سرد	بله
بزرگ	شور	گرم	خیر
بزرگ	ترش	گرم	بله
بزرگ	ترش	گرم	بله
کوچک	شیرین	سرد	بله
کوچک	شیرین	سرد	بله