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) درستی و نادرستی گزاره‌های زیر را همراه یک استدلال کوتاه بیان کنید.

(الف) 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 که واریانس زیاد و بایاس کمی دارند انتخاب می‌کنیم و نتیجه نهایی با یک اجماع از نتیجه آن‌ها به دست می‌آید.

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

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

۲. (۱۵ نمره) روش تخمین چگالی در $k$nn برآوردی از چگالی یک نقطه از فضای ویژگی‌ها را نسبت به چگالی k همسایهٔ نزدیک‌تر می‌دهد. نشان دهید که مدل چگالی $k$nn 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.

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

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 $\hat{h}_۱(\mathbf{x}), \hat{h}_۲(\mathbf{x}), \ldots, \hat{h}_m(\mathbf{x})$ نشان می‌دهیم. اکنون قرار دهید:

$$\hat{H}_m(\mathbf{x}) = \frac{۱}{m}\sum_{i=۱}^{m} \hat{h}_i(\mathbf{x})$$

ثابت کنید که:

$$\mathbb{E}_{\mathbf{x}}\left[(\hat{H}_m(\mathbf{x}) - h(\mathbf{x}))^۲\right] \leq \frac{۱}{m}\sum_{i=۱}^{m} \mathbb{E}_{\mathbf{x}}\left[(\hat{h}_i(\mathbf{x}) - h(\mathbf{x}))^۲\right]$$

مفهوم این حکم چیست؟

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

(الف) 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 $\mathbf{X} = \{\mathbf{x}_i\}_{i=۱}^{n}$ با میانگین و واریانس $\mu, \sigma^۲$ داده شده‌اند. بگیرید:

$$\mathbf{z}_m = \frac{۱}{m}\Big(\sum_{i=۱}^{m} \mathbf{a}_i\Big)$$

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.

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

(الف) 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 $\theta_۱, \theta_۲, \ldots, \theta_K$ یک بردار ثابت $\mathbf{c}$ اضافه کنیم، مقادیر احتمالات تغییری نمی‌کنند.

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

(الف) 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.

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

- 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 $\frac{1}{2}\|\mathbf{w}\|^2$ با توجه به شروط داده شده، این margin افزایش می‌یابد؟

(ب) فرض کنید مدل Linear Regression را داریم:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma\mathbf{I})$$

- 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.

۷. (۱۰ نمره) همان‌طور که می‌دانید اعتبارسنجی متقاطع[3] با حذف یک نمونه یا LOOCV را می‌توان بدین گونه نشان داد:

$$e(S_n) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \hat{f}_{-i}(\mathbf{x}_i)\big)^2$$

که در آن $S_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ (مجموعه داده‌های آموزش)، $S_n^{-i} = S_n \backslash \{(\mathbf{x}_i, y_i)\}$ (مجموعه داده‌ها پس از حذف نمونهٔ iام)، و $\hat{f}_{-i}$ مدل آموزش‌دیده روی S_n^{-i} می‌باشند.

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

$$e_1(S_n) = \big(y_1 - \hat{f}_{-1}(\mathbf{x}_1)\big)^2$$

است. نشان دهید:

$$\mathbb{E}\big[e_1(S_n)\big] = \mathbb{E}\Big[\big(y - \hat{f}_{S_{n-1}}(\mathbf{x})\big)^2\Big]$$

که در آن $(\mathbf{x}, y)$ یک نمونهٔ تصادفی از توزیع داده‌هاست و $\hat{f}_{S_{n-1}}$ مدل آموزش‌دیده روی S_{n-1} است.

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

$$\mathbb{E}\big[e(S_n)\big] = \mathbb{E}\Big[\big(y - \hat{f}_{S_{n-1}}(\mathbf{x})\big)^2\Big]$$

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

(الف) آنتروپی[4] اولیه ستون مربوط به خوشایند بودن غذا چند است؟

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

(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.

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