دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

بهار ۱۴۰۲

استاد: محمدحسین رهبان
گردآورندگان: محمدرضا دویران، امیررضا میرزایی و محمد جواد هزاره

تمرین پنجم **آشنایی با یادگیری ماشین، رگرسیون، درخت تصمیم‌گیری** مهلت ارسال: ۱۹ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۷ روز و در مجموع ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۲ درصد از نمره تمرین به صورت ساعتی 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 (۱۲۰ نمره)

۱. (۲۰ نمره) با توجه به مفاهیم فراگرفته شده در درس به سوالات زیر پاسخ دهید.

الف) برای مدل‌هایی که اریبی زیادی دارند دست کم دو راهکار ارائه دهید که مقدار این اریبی کاهش یابد.

ب) فرض کنید تعدادی از ویژگی‌های مدل دوبه‌دو با یکدیگر هم‌بسته باشند. این اتفاق را از دیدگاه bias - variance بررسی کنید. اگر ویژگی‌های هم‌بسته را حذف کنیم بیان کنید که چگونه bias و variance تغییر می‌کنند.

پ) کدام یک از گزاره‌های زیر درست می‌باشند؟ چرا؟

- اگر مقدار بایاس زیاد باشد، افزایش تعداد داده‌های آموزش می‌تواند باعث کاهش بایاس شود.
- افزایش پیچیدگی مدل در رگرسیون همواره باعث کاهش خطای آموزش و افزایش خطای تست می‌شود.

۲. (۳۰ نمره)

(آ) فرض کنید برای داده‌های جدول ۱ یک درخت تصمیم آموزش می‌دهیم تا X را به وسیله‌ی A, B, C پیش‌بینی کنیم. درصد خطای مدل پس از آموزش بر روی داده‌های آموزش چقدر خواهد بود؟

(ب) فرض کنید روی مجموعه‌ی داده‌ی دلخواهی، درخت تصمیمی برای دسته‌بندی بین k کلاس، آموزش می‌دهیم. حداکثر خطایی که ممکن است این مدل روی داده‌های آموزش داشته باشد چقدر خواهد بود؟ (پاسخ را به صورت کسری بنویسید)

C	B	A	X
۰	۰	۰	۰
۱	۰	۰	۰
۱	۰	۰	۰
۰	۱	۰	۰
۱	۱	۰	۰
۱	۱	۰	۱
۱	۱	۰	۱
۰	۰	۱	۰
۱	۰	۱	۱
۰	۱	۱	۱
۰	۱	۱	۱
۱	۱	۱	۰
۱	۱	۱	۱

جدول ۱: داده‌های مدل درخت تصمیم

۳. (۴۰ نمره) در رابطه با الگوریتم Logistic regression به سوالات زیر پاسخ دهید.

الف) این الگوریتم را برای حالت K کلاسه تغییر دهید و احتمالات آن را بنویسید.

ب) همانطور که در قسمت الف به دست آوردید، در Logistic regression برای K کلاس، احتمال پسین به روش زیر محاسبه می‌شود:

$$P(Y=k|X=x)=\frac{e^{w_k^T x}}{۱+\Sigma_{i=۱}^{K-۱} e^{w_i^T x}}, k=۱,۲,۳,۴,...,K-۱$$
$$P(Y=K|X=x)=\frac{۱}{۱+\Sigma_{i=۱}^{K-۱} e^{w_i^T x}}$$

برای راحتی فرض کردیم که $w_k=۰$ می‌باشد. کدام یک از پارامترها باید تخمین زده شوند؟

پ) حال log-likelihood زیر را برای n نمونه‌ی زیر ساده کنید:

$$Samples: (x_۱,y_۱),(x_۲,y_۲),...,(x_n,y_n)$$
$$L(w_۱,...,w_{k-۱})=\Sigma_{i=۱}^{n} lnP(Y=y_i|X=x_i)$$

ت) گرادیان L را نسبت به هریک از w_k ها بیاید و آن را ساده کنید.

ث) تابع هدف زیر را در نظر بگیرید. گرادیان f را با توجه به هریک از w_k ها بیابید.

$$f(w_۱,...,w_{k-۱})=L(w_۱,...,w_{k-۱})-\frac{\lambda}{۲}\Sigma_{j=۱}^{K-۱}||w_j||_۲^۲$$

۴. (۳۰ نمره) فرض کنید n داده آموزش با m ویژگی داریم که که ماتریس این داده‌ها را $X_{n\times m}$ در نظر می‌گیریم. بردار مقدار هدف نیز برابر $y=[y^{(۱)},...,y^{(n)}]$ می‌باشد. در ادامه منظور از x_j، j امین ستون ماتریس X است. حال با توجه به توضیحات داده شده به سوالات زیر پاسخ دهید.

الف) ابتدا ثابت کنید اگر رگرسیون را فقط بر روی یکی از m ویژگی موجود آموزش دهیم آنگاه خواهیم داشت:

$$w_j=\frac{x_j^T y}{x_j^T x_j}$$

ب) فرض کنید ستون‌های ماتریس X متعامد باشد. ثابت کنید که پارامترهای بهینه از آموزش رگرسیون بر روی همه ویژگی‌ها با پارامترهای بهینه حاصل از آموزش روی هر ویژگی به طور مستقل یکسان است.

پ) فرض کنید می‌خواهیم یک رگرسیون بر روی بایاس و یکی از ویژگی‌های نمونه داده‌ها آموزش دهیم. ($w = [w_j, w_.]$) با توجه به اطلاعات داده شده عبارات زیر را اثبات کنید:

$w_j = \frac{cov[x_j, y]}{var[x_j]}$

$w_. = E[y] - w_j E[x_j]$

سوالات عملی (۲۰۰ نمره)

۱. (۱۰۰ نمره) برای حل سوال عملی اول به نوت‌بوک decision_tree.ipynb مراجعه کنید.

۲. (۱۰۰ نمره) برای حل سوال عملی دوم به نوت‌بوک logistic_regression.ipynb مراجعه کنید.