یادگیری ماشین

نیم‌سال دوم ۱۴۰۲-۱۴۰۱
مدرس: دکتر سید ابوالفضل مطهری

دانشکده‌ی مهندسی کامپیوتر

تمرین دوم

۱. فرض کنید ما یک مدل رگرسیون لجستیک را آموزش داده‌ایم و احتمال کلاس‌ها را می‌توان از طریق فرمول زیر محاسبه کرد:

$$z(x) = \sigma(wx + b)$$

که در آن $(w_k, w_{k,0})$ پارامترهای مدل هستند و می‌توانیم هر داده جدید را با استفاده از رابطه زیر دسته‌بندی کنیم:

$$y(x) = [z(x) > 0.5]$$

نشان دهید این رابطه متناظر با یک مرز تصمیم خطی در فضای ورودی است.

۲. مدل‌های رگرسیون لجستیک زیر را برای عمل دسته‌بندی دوتایی با استفاده از تابع سیگموید $g(z) = \frac{1}{1+e^{-z}}$ در نظر بگیرید.

- مدل ۱: $P(Y = 1 \mid X, w_1, w_2) = g(w_1 X_1 + w_2 X_2)$
- مدل ۲: $P(Y = 1 \mid X, w_1, w_2, w_0) = g(w_0 + w_1 X_1 + w_2 X_2)$

فرض کنید سه داده آموزشی با ویژگی‌های زیر داریم:

$$x^{(1)} = [1, 1]^T \qquad x^{(2)} = [1, 0]^T \qquad x^{(3)} = [0, 0]^T$$
$$y^{(1)} = 1 \qquad y^{(2)} = -1 \qquad y^{(3)} = 1$$

(آ) آیا برچسب داده سوم $x^{(3)}$ در مدل ۱ اهمیتی دارد؟(آیا مقدار یادگرفته شده برای $\mathbf{w} = (w_1, w_2)$ تغییری خواهد کرد اگر $y^{(3)}$ را برابر با -1 قرار دهیم؟). آیا تفاوتی برای مدل ۲ خواهد داشت؟ جواب خود را توجیه کنید.(راهنمایی: از مرز تصمیم در فضای دو بعدی استفاده کنید) [1]

(ب) حال فرض کنید که ما مدل رگرسیون لجستیک (مدل ۲) را بر روی n داده آموزشی $x^{(1)}, \ldots, x^{(n)}$ با برچسب‌های $y^{(1)}, \ldots, y^{(n)}$ را از طریق بیشینه کردن log-likelihood با جمله regularization زیر آموزش دهیم

$$\sum_i \log P(y^{(i)} \mid x^{(i)}, \mathbf{w}) - \frac{\lambda}{2}||\mathbf{w}||^2 = \sum_i \log\ g(y^{(i)\mathbf{w}^{\mathbf{T}}x^{(i)}}) - \frac{\lambda}{2}||\mathbf{w}||^2$$

[1] decision boundary in 2D space

برای مقادیر بالای λ، جمله‌های log-likelihood را می‌توان به صورت تابع‌هایی خطی از $\mathbf{w}$ به صورت زیر در نظر گرفت:

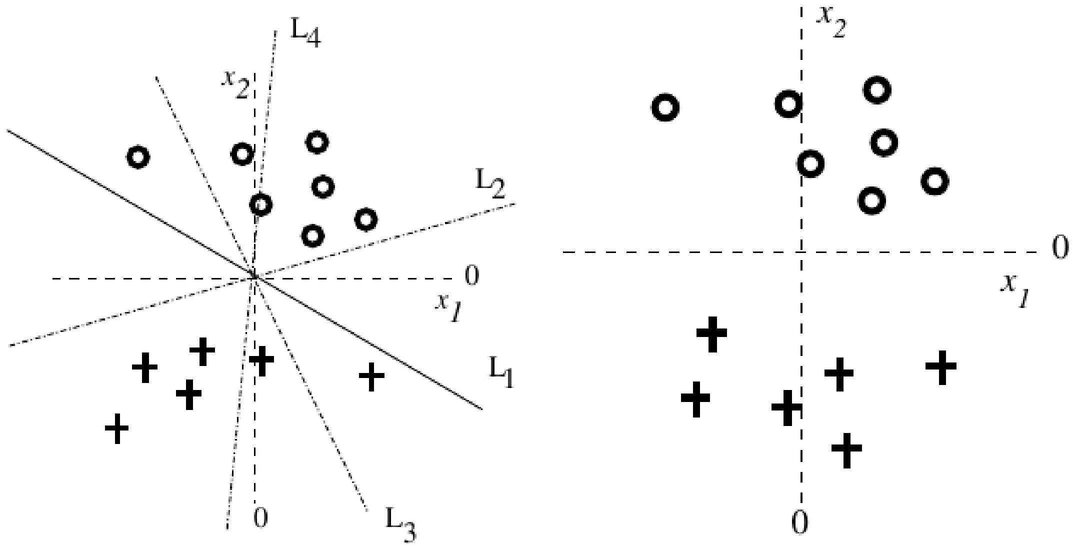
$$\log\ g(y^{(i)}\mathbf{w}^{\mathbf{T}}x^{(i)}) \approx \frac{۱}{۲}y^{(i)}\mathbf{w}^{\mathbf{T}}x^{(i)}$$

رابطه log-likelihood با regularization را با استفاده از این تقریب (برای مدل ۱) بدست آورید و رابطه MLE[۲] برای $\hat{\mathbf{w}}$ نسبت به λ و داده آموزشی $\{x^{(i)}, y^{(i)}\}$ بدست آورید. با توجه به نتایج بدست آمده توضیح دهید رفتار $\mathbf{w}$ با افزایش λ چگونه خواهد بود. (فرض کنید $x^{(i)} = (x_۱^{(i)}, x_۲^{(i)})^T$ و $y^{(i)}$ دو مقدار ۱− و ۱ را به خود می‌گیرد).

۳. در یک مسئله دسته‌بندی که برچسب $y \in \{۰, ۱\}$ است، تابع هزینه cross-entropy برای رگرسیون لجستیک را بنویسید و محدب‌بودن یا نبودن آن را اثبات کنید.

۴. در این سوال سعی می‌کنیم مسئله دسته‌بندی دوکلاسه شکل (۱آ) را به وسیله رگرسیون لجستیک خطی زیر حل کنیم:

$$\hat{P}(y = ۱|x, w_۱, w_۲) = \frac{۱}{۱ + e^{-w_۱x_۱ - w_۲x_۲}}$$



(ب) داده‌ها می‌توانند به وسیله $L_۱$ با خطای صفر جدا شوند. چند مرز تصمیم ممکن دیگر نیز با $L_۲, L_۳$ و $L_۴$ نشان داده‌شده‌اند.

(آ) مجموعه داده دو بعدی

شکل ۱

[۲] Maximum Likelihood Estimation

(آ) جهت regularization سعی می‌کنیم برای مقدار C بزرگ، مقدار

$$\sum_{i=1}^{n} \log p(y_i|x_i, w_1, w_2) - \frac{C}{2} w_2^2$$

را بیشینه کنیم. برای هر یک از خطوط L_2، L_3 و L_4 توضیح دهید که چرا می‌توانند یا نمی‌توانند حاصل regularization بالا باشند.

(ب) اگر عبارت regularization را به نرم-۱ تغییر دهیم و علاوه بر w_2، w_1 را نیز در آن دخیل کنیم به penalized log-likelihood زیر می‌رسیم:

$$\sum_{i=1}^{n} \log p(y_i|x_i, w_1, w_2) - \frac{C}{2}(|w_1| + |w_2|) \ .$$

با در نظر گرفتن دادگان شکل (۱آ) و مدل گفته‌شده برای $\hat{P}(y = 1|x, w_1, w_2)$، با زیاد کردن ضریب C:

i. ابتدا w_1 و سپس w_2 صفر می‌شود.

ii. w_1 و w_2 به صورت همزمان صفر می‌شوند.

iii. ابتدا w_2 و سپس w_1 صفر می‌شود.

iv. هیچکدام.

راجع به انتخاب خود توضیح دهید.

۵. فرض کنید داده‌های تعدادی از دانشجویان یک کلاس را جمع آوری می‌کنیم که X_1 در آن ساعت مطالعه، X_2 معدل (GPA) آن‌ها و Y این است که آن شخص نمره کامل این درس را گرفته است یا نه. یک مدل رگرسیون لجستیک را روی این اطلاعات آموزش داده‌ایم که پارامترهای آن به صورت زیر است.

$\beta_0 = -6, \beta_1 = 0/05, \beta_2 = 1$

β_1 به مربوط X_1 و β_2 به مربوط X_2

الف) احتمال اینکه شخصی که ۴۰ ساعت درس می‌خواند و معدل ۳.۵ دارد، از این درس نمره کامل بگیرد را حساب کنید.

ب) دانش آموز بخش قبل چند ساعت باید درس بخواند که با احتمال ۵۰ درصد نمره کامل از درس بگیرد؟

۶. فرض کنید یک مسئله دسته بندی با سه دسته داریم که برچسب‌های آن‌ها $y \in (0, 1, 2)$ است و هر ورودی ویژگی داده باینری $X_1, X_2, X_3 \in (0, 1)$ دارد. برای اینکه داده‌هایی به این شکل را با دسته‌بند بیز ساده [3] دسته‌بندی کنید به دانستن چند پارامتر نیاز دارید؟

[3] Naive Bayes Classifier