تمرین ۱۵ ML

آرمین سعادت
۹۶۱۰۵۸۲۹

سوال ۱)

a: فرض کنید یک داده به کلاس $\begin{bmatrix}0\\1\\0\end{bmatrix}$ تعلق داشته باشد و مقدار به دست آمده برای آن $\begin{bmatrix}5\\150\\5\end{bmatrix}$ باشد.
مشخص است که مقدار این داده درست تعیین شده و به کلاس درستی تخصیص داده شده است. اما بازهم SSE بسیار بالایی خواهد داشت.
بنابراین SSE اصلا ملاک خوبی برای classification نیست و تنها سعی می‌کند که داده ها را تا نزدیک خط جداکننده نگه دارد. یعنی نه تنها به داده های نادرست و نویز خیلی حساس است، بلکه اگر یک داده به درستی در کلاس 1 باشد اما از صفحه جداکننده فاصله زیادی داشته باشد، بازهم خطای خیلی زیادی SSE خروجی می دهد به صورتی که داده کاملا دقیق و درست دسته بندی شده است.
به عبارتی ممکن است تمام داده ها خیلی خوب و درست و با اطمینان دسته بندی شده باشند اما خطای SSE مقدار خیلی زیادی باشد.

b: برای رسیدن به جواب بهینه، از روش های iterative مانند gradiant decent استفاده می شود.
استفاده از Perceptron Criterion یک فضای پیوسته و محدب (convex) ایجاد می کند که می توان با استفاده از روش های iterative به جواب بهینه رسید.
اما اگر از error-function به این صورت استفاده شود که تعداد داده‌های misclassified را بشمارد، تقریبا همه جای فضا گرادیان صفر است و فضای صاف داریم. پس عملا الگوریتم های iterative نمی‌تواند دقیق و خیلی زود متوقف می‌شود چرا که گرادیان صفر است اما هنوز به جواب بهینه نرسیده است.

c: logistic regression یک روش احتمالی است درصورتی که Perceptron یک تابع Discriminant است.
بنابراین با استفاده از logistic regression از مزایای مدل احتمالاتی بهره‌مند می‌شویم. مثلا این که فرایند inference از decision making جدا می‌شود، می‌توانیم یک rejection area تعریف کنیم و سایر مزیت های مدل احتمالاتی نسبت به مدل غیراحتمالاتی.
علاوه براین ها، Perceptron خیلی صعب به حالت چندکلاسه تعمیم می‌یابد درصورتی که در روش احتمالاتی تابع Softmax به راحتی جایگزین می‌شود.
همچنین Perceptron بین صفحه‌های جداکننده که margin متفاوت اما error یکسانی دارند تفاوتی قائل نیست که منجر به افزایش خطا

روی داده های تست می شود. (به دلیل در نظر نگرفتن margin)

اگر ملاک ها و وزنهای misclassified کردن عوض شود، در روش Perceptron همه ی کارها باید از اول انجام شود. امّا در روش logistic regression، $P(C_k|x)$ به دست آمده همچنان معتبر است و باید از این Decision making به روز رسانی شود.

d: در روش IRLS، عمدتاً از ماتریس Hessian استفاده می شود. (با توجه به الگوریتم نیوتون-رافسون) بنابراین هر step آن هزینه محاسباتی بیشتری نسبت به step در GD دارد.
امّا در هر مرحله، تخمین بهتری ارائه می دهد و در نهایت در تعداد مراحل کمتری جواب بهینه به دست می آید. بین یک trade-off بین هزینه هر مرحله، تعداد مراحل رخ می دهد که در برخی موارد GD در زمان کمتری به جواب می رسد و در برخی موارد IRLS.
دلیل این که هر مرحله IRLS پیچیده تر طول می کشد این است که هر دفعه باید $(X^TRX)^{-1}$ را حساب کند که این کار هزینه بر تر است نسبت به محاسبه ی گرادیان. اما خب در تعداد مراحل کمتری به جواب می رسد و معلوم است در کل زمان رسیدن به جواب را کاهش دهد.

e: Probit regression نسبت به outliers حساس تر است.
انتهای (tails) تابع logistic regression که همان logistic sigmoid است خیلی شبیه به تابع $\exp(-x)$ عمل می کند. در صورتی که انتها (tails) در تابع Probit regression خیلی شبیه به تابع $\exp(-x^2)$ عمل می کند. بنابراین تابع Probit regression احتمال خیلی کمی برای وجود outliers قائل شده است و در صورت مشاهده آن، بسیار جریمه بالایی در نظر گرفته و در نتیجه به outlier حساس تر است.

سؤال ۲)

۲ نقطه دلخواه مانند a و b را در C_j (یعنی متعلق به ناحیه j) در نظر می‌گیریم. باید اثبات کنیم خط وصل کننده این ۲ نقطه نیز در ناحیه j قرار دارد.

$$\forall \lambda \in (0,1): x = \lambda a + (1-\lambda) b$$

زیرا f_j خطی است $\rightarrow$ $f_j(x) = f_j(\lambda a + (1-\lambda) b) = \lambda f_j(a) + 1-\lambda f_j(b)$

$$\forall i \in \{1,2,\dots,K\}: f_j(a) \geq f_i(a),\ f_j(b) \geq f_i(b)$$

زیرا a و b در ناحیه j است.

$$\Rightarrow \lambda f_j(a) \geq \lambda f_i(a),\ (1-\lambda) f_j(b) \geq (1-\lambda) f_i(b)$$

$$\Rightarrow \forall i \in \{1,2,\dots,K\}: \lambda f_j(a) + (1-\lambda) f_j(b) \geq \lambda f_i(a) + (1-\lambda) f_i(b)$$

$$\Rightarrow f_j(x) \geq f_i(x) \Rightarrow$$ x نیز در ناحیه‌ی j است.

به ازای تمام نقاط x روی خط واصل a و b، اثبات می‌شود که x هم درون ناحیه j است.

پس decision regions شرط convex بودن را دارند. چون برای هر ناحیه دلخواه مانند j اثبات شد.

سوال ۳)

$w^1 = 0$

$\forall x^j \to$ misclassified: $w^{k+1} \cdot w^* = (w^k + t^j x^j) \cdot w^*$

$= w^k \cdot w^* + t^j (x^j \cdot w^*)$

$= w^k \cdot w^* + \underbrace{w^{*T} x^j t^j}_{\geq \gamma} \geq w^k \cdot w^* + \gamma$

به روش مشابه داریم:

$$\left.\begin{array}{l} w^{k+1} \cdot w^* > w^k \cdot w^* + \gamma \\ w^k \cdot w^* > w^{k-1} \cdot w^* + \gamma \\ \vdots \\ w^2 \cdot w^* > \gamma \end{array}\right\} \text{از استقرا و جایگزینی نتیجه می‌شود که} \rightarrow \boxed{w^{k+1} \cdot w^* > k\gamma} \rightarrow \text{اثبات}$$

$$w^{k+1} \cdot w^* \leq \|w^{k+1}\| \, \|w^*\| = \|w^{k+1}\| \Rightarrow \boxed{\|w^{k+1}\| > k\gamma} \quad \text{lower bound}$$

$$\underbrace{w^{k+1} = w^k + t^j x^j}_{\text{update rule}} \Rightarrow \|w^{k+1}\|^2 = \|w^k + t^j x^j\|^2 = \|w^k\|^2 + \|t^j x^j\|^2 + 2(w^k \cdot x^j) t^j$$

$$= \|w^k\|^2 + \|x^j\|^2 + 2(w^k \cdot x^j) t^j$$

$$\leq \|w^k\|^2 + \|x^j\|^2 \leq \|w^k\|^2 + R^2$$

$$\Rightarrow \|w^{k+1}\|^2 \leq \|w^k\|^2 + R^2 \xrightarrow{\text{از استقرا مشابه قسمت قبل}} \boxed{\|w^{k+1}\|^2 \leq kR^2} \quad \text{upper bound}$$

$$k^2\gamma^2 \leq \|w^{k+1}\|^2 \leq kR^2 \Rightarrow k^2\gamma^2 \leq kR^2 \Rightarrow \boxed{k \leq \frac{R^2}{\gamma^2}}$$

$P(C_1) = \pi \Rightarrow P(C_2) = 1-\pi$

$P(x|C_1) = N(x|\mu_1, \Sigma)$, $P(x|C_2) = N(x|\mu_2, \Sigma)$, $t_n \in \{1, 0\}$ ($1 \leftarrow C_1$, $0 \rightarrow C_2$)

$$P(t|X, \mu_1, \mu_2, \Sigma, \pi) = \prod_{n=1}^{N} P(t_n|x_n) = \prod_{n=1}^{N} P(x_n|t_n) P(t_n)$$

$$= \prod_{n=1}^{N} \left(\pi N(x_n|\mu_1, \Sigma)\right)^{t_n} \left((1-\pi) N(x_n|\mu_2, \Sigma)\right)^{1-t_n}$$

$$\Rightarrow -\ln P(t|X, \mu_1, \mu_2, \Sigma, \pi) = -\sum_{n=1}^{N} \{t_n \ln \pi + t_n \ln N(x_n|\mu_1, \Sigma) + (1-t_n)\ln(1-\pi) + (1-t_n)\ln N(x_n|\mu_2, \Sigma)\}$$

$$\Rightarrow \nabla_{\pi} \ln P(t|\mu_1, \mu_2, \Sigma, \pi) = \nabla_{\pi} \sum_{n=1}^{N} t_n(\ln \pi + (1-t_n)\ln(1-\pi)) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \frac{t_n}{\pi} - \frac{1-t_n}{1-\pi} = \sum_{n=1}^{N} \frac{t_n - t_n\pi - \pi + t_n\pi}{\pi(1-\pi)} = \sum_{n=1}^{N} \frac{t_n - \pi}{\pi(1-\pi)} = 0$$

$$\Rightarrow \frac{1}{\pi(1-\pi)} \sum_{n=1}^{N} (t_n - \pi) = 0 \Rightarrow \sum_{n=1}^{N} t_n = N\pi \Rightarrow \boxed{\frac{1}{N}\sum_{n=1}^{N} t_n = \pi}$$

$$\nabla_{\mu_1} \ln P(t|\mu_1, \mu_2, \Sigma, \pi) = \nabla_{\mu_1} \sum_{n=1}^{N} t_n \ln N(x_n|\mu_1, \Sigma) = 0$$

$$\Rightarrow \nabla_{\mu_1} \sum_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1} (x_n - \mu_1) + \text{Const} = 0 \Rightarrow \sum_{n=1}^{N} t_n x_n - \mu_1 = 0$$

$$\Rightarrow \sum_{n=1}^{N} t_n x_n = N\mu_1 \Rightarrow \boxed{\frac{1}{N}\sum_{n=1}^{N} t_n x_n = \mu_1} \xrightarrow{\text{Caso Similar}} \boxed{\frac{1}{N}\sum_{n=1}^{N} (1-t_n) x_n = \mu_2}$$

ادامه سوال ۲)

$$\ln p(t|\mu_1,\mu_2,\Sigma,\pi) = \sum_{n=1}^{N} t_n \ln\pi + t_n \ln N(x_n|\mu_1,\Sigma) + (1-t_n)\ln(1-\pi) + (1-t_n)\ln N(x_n|\mu_2,\Sigma)$$

بخش‌های مربوط به Σ عبارت است از:

$$\sum_{n=1}^{N} t_n \ln N(x_n|\mu_1,\Sigma) + (1-t_n)\ln N(x_n|\mu_2,\Sigma)$$

$$= -\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} t_n (x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1) - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\Sigma|$$

$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)(x_n-\mu_2)^T\Sigma^{-1}(x_n-\mu_2) + \text{const}$$

با حذف const از عبارت بالا داریم:

$$-\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\mathrm{Tr}\{\Sigma^{-1}S\}, \quad S = \frac{N_1}{N}S_1 + \frac{N_2}{N}S_2$$

$$S_1 = \frac{1}{N_1}\sum_{n\in C_1}(x_n-\mu_1)(x_n-\mu_1)^T, \quad S_2 = \frac{1}{N_2}\sum_{n\in C_2}(x_n-\mu_2)(x_n-\mu_2)^T$$

$$\Rightarrow \boxed{\Sigma^* = S}$$

خلاصه:

$$P(C_1) = \frac{1}{N}\sum_{n=1}^{N} t_n, \quad P(C_2) = 1-P(C_1) = \frac{1}{N}\sum_{n=1}^{N}(1-t_n)$$

$$\mu_1 = \frac{1}{N}\sum_{n=1}^{N} t_n x_n, \quad \mu_2 = \frac{1}{N}\sum_{n=1}^{N}(1-t_n)x_n$$

$$\Sigma = \frac{1}{N}\left(\sum_{n\in C_1}(x_n-\mu_1)(x_n-\mu_1)^T + \sum_{n\in C_2}(x_n-\mu_2)(x_n-\mu_2)^T\right)$$

سوال ۵)

داده ها به صورت خطی جدا پذیر هستند. پس داریم:

$$\forall x_i \in C_1: \quad w^T x_i > 0$$

$$\forall x_i \in C_0: \quad w^T x_i < 0$$

اگر عدد $a > 0$ را در w ضرب کنیم، روابط بالا همچنان برقرار خواهند بود. پس اگر w یک جواب درست باشد که داده ها را به درستی از هم جدا می کند، aw نیز یک جواب درست است.

برای maximum likelihood داریم:

$$P(T|X) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{(1-t_n)}$$

$$\Rightarrow -\ln P(T|X) = -\sum_{n=1}^{N} t_n \ln y_n + (1-t_n) \ln (1-y_n)$$

$w^T x_n$ یک عبارت مثبت است، پس اگر $a > 1$ در $w^T x_n$ ضرب شود، y_n نیز افزایش می یابد. پس $t_n \ln y_n$ نیز زیاد می شود.

$$\begin{cases} t_n = 1 \Rightarrow w^T x_n > 0 \Rightarrow y_n = \dfrac{1}{1+e^{-w^T x_n}} \longrightarrow \\ t_n = 0 \Rightarrow w^T x_n < 0 \Rightarrow y_n = \dfrac{1}{1+e^{-w^T x_n}} \longrightarrow \end{cases}$$

$w^T x_n$ یک عبارت منفی است. پس اگر $a > 1$ در $w^T x_n$ ضرب شود، y_n کاهش می یابد پس $1-y_n$ افزایش می یابد. پس $(1-t_n)(1-y_n)$ زیاد می شود.

بنابراین با ضرب w در $a > 1$، $\ln P(T|X)$ افزایش یافته پس $-\ln P(T|X)$ کاهش می یابد.

ادامه سؤال (۵)

بنابراین اگر w را به عددی بزرگتر از ۱ ضرب کنیم، یعنی هرچه $|w|$ را از اینش دهیم، ادعا که همچنان داده‌ها را به درستی دسته‌بندی می‌کند، تابع ضرر ML آن کوچک می‌شود. پس در حالت حدی، بهترین گزینه این است که w در $+\infty$ ضرب شود. بنابراین اندازه w به بی‌نهایت میل خواهد کرد.

که البته این اتفاق بدی است چراکه w بیش از اندازه بزرگ شده و عملاً حسابی overfit می‌کند مسئله روی داده‌های train و ولی test بد می‌شود.

سؤال ۶)

برای حل مدل را به صورت $\sigma(W^T x)$ در نظر می‌گیریم چرا که logistic regression است.

برای داده‌های D_1، ویژگی‌های ما بردار x است پس به اندازه‌ی بعد x (که آن را d می‌نامیم) پارامتر داریم. (یا مثلاً با در نظر گرفتن bias اما پارامتر).

اما در مجموعه داده D_2، عموماً تعداد ویژگی‌ها ۲ برابر می‌شود. بنابراین تعداد پارامترهای w_i نیز ۲ برابر می‌شود. پس می‌توان گفت که زمان حل و میزان پردازش مورد نیاز برای به‌دست آوردن پارامترهای بهینه تقریباً ۲ برابر می‌شود. (به عبارتی سرعت دستیابی به جواب نصف می‌شود.)

از طرفی این ویژگی‌های جدید کاملاً مشابه هم هستند و اگر پارامترها همگی از مثلاً ۰ شروع کنند، عموماً نصف پارامترها در هر مرحله برابر با نصف دیگر است. یعنی هیچ دیتای جدیدی به مسئله اضافه نشده و صرفاً هزینه بالا رفته است.

سوال ۷)

$\int_{-\infty}^{\infty} \mathcal{N}(\theta|0,1)\,d\theta = 1$ ، $\mathcal{N}(\theta|0,1) \to$ متقارن نسبت به ۰

$$\Rightarrow \int_{-\infty}^{0} \mathcal{N}(\theta|0,1)\,d\theta = \frac{1}{2}$$

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,d\theta = \int_{-\infty}^{0} \mathcal{N}(\theta|0,1)\,d\theta + \int_{0}^{a} \mathcal{N}(\theta|0,1)\,d\theta$$

$$= \frac{1}{2} + \int_{0}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}\,d\theta \xrightarrow[\text{تغیر متغیر}]{\theta \to \sqrt{2}x} = \frac{1}{2} + \int_{0}^{\frac{a}{\sqrt{2}}} \frac{1}{\sqrt{2\pi}} e^{\frac{-2x^2}{2}} dx\sqrt{2}$$

$$= \frac{1}{2} + \int_{0}^{\frac{a}{\sqrt{2}}} \frac{1}{\sqrt{\pi}} e^{-\theta^2}\,d\theta = \frac{1}{2} + \frac{1}{2}\int_{0}^{\frac{a}{\sqrt{2}}} \frac{2}{\sqrt{\pi}} e^{-\theta^2}\,d\theta$$

$$= \frac{1}{2}\left(1 + \frac{2}{\sqrt{\pi}} \int_{0}^{\frac{a}{\sqrt{2}}} e^{-\theta^2}\,d\theta\right) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)\right)$$

$$\Rightarrow \Phi(a) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)\right)$$

سوال ۸)

$$\begin{cases} t_n = 1 & \text{if } a_n \geq \theta \\ t_n = 0 & \text{O.W.} \end{cases} \longrightarrow f(a) = \int_{-\infty}^{a} G(\theta)\, d\theta$$

که $G(\theta)$ تابع توزیع θ است،

$$P(t_n|x_n) = y_n^{t_n} \times (1-y_n)^{1-t_n}, \quad y_n = f(a_n) = \int_{-\infty}^{w^T x_n} G(\theta) d\theta$$

$$\Rightarrow P(T|X) = \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{(1-t_n)} \longrightarrow -\ln p(T|X) = -\sum_{n=1}^{N} t_n \ln y_n + (1-t_n)\ln(1-y_n) \quad \text{(E(w))}$$

$$\nabla_w E(w) = \nabla_y E(w) \times \nabla_w y = \sum_{n=1}^{N} \left\{ -\left(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right) \times \nabla_w y_n \right\}$$

$$= \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)} \nabla_w y_n = \sum_{n=1}^{N} \frac{(y_n - t_n)}{y_n(1-y_n)} x_n \times \frac{d f(w^T x_n)}{d(w^T x_n)} \longrightarrow \text{رابطی کراپیان}$$

متغیر q_i را اینچنین تعریف می‌کنیم:

$$\begin{cases} t_i = 1 \longrightarrow q_i = 1 \\ t_i = 0 \longrightarrow q_i = -1 \end{cases} \quad q_i = 2t_i - 1$$

پس رابطه به این شکل تبدیل می‌شود:

$$\sum_{n=1}^{N} \frac{q_n x_n}{f(q_n w^T x_n)} \times \frac{d f(q_n w^T x_n)}{d(q_n w^T x_n)}$$

با توجه به این نکته که $f(-a) = 1 - f(a)$ همچنین برای راحتی $\frac{df(a)}{d(a)}$ را به شکل $s(a)$ نمایش می‌دهیم.

پس برای Hessian داریم:

$$H = \nabla_w \nabla_w E(w) = \nabla_w \sum_{n=1}^{N} \frac{s(q_n w^T x_n) q_n}{f(q_n w^T x_n)} x_n$$

$$= \sum_{n=1}^{N} \nabla_w \left(\frac{s(q_n w^T x_n)}{f(q_n w^T x_n)} \right) q_n x_n = \sum_{n=1}^{N} \left\{ \frac{1}{-f(q_n w^T x_n)} \times \nabla_w s(q_n w^T x_n) + s(q_n w^T x_n) \nabla_w \left(\frac{1}{f(q_n w^T x_n)} \right) \right\} q_n x_n$$

$$= \sum_{n=1}^{N} \left(\frac{-1}{f(q_n w^T x_n)} s(q_n w^T x_n)(q_n w^T x_n) q_n x_n^T - \frac{s(q_n w^T x_n)^2 q_n x_n^T}{f(q_n w^T x_n)^2} \right) q_n x_n$$

$$= -\sum_{n=1}^{N} \left(\frac{s(q_n w^T x_n)(w^T x_n) q_n^2}{f(q_n w^T x_n)} + \frac{s(q_n w^T x_n)^2 q_n^2}{f(q_n w^T x_n)^2} \right) x_n^T x_n$$

$$= -\sum_{n=1}^{N} \left\{ \frac{s(q_n w^T x_n) q_n}{f(q_n w^T x_n)} (w^T x_n) + \left(\frac{s(q_n w^T x_n) q_n}{f(q_n w^T x_n)} \right)^2 \right\} x_n^T x_n \qquad q_n^2 = q_n \text{ :چون}$$

$$= -\sum_{n=1}^{N} \lambda_n (w^T x_n + \lambda_n) x_n^T x_n \; , \quad \lambda_n = \frac{s(q_n w^T x_n)}{f(q_n w^T x_n)} q_n \longrightarrow \underline{\text{Hessian}}$$

اگر فرض کنیم θ از $N(0,1)$ می‌آید، خواهیم داشت:

$$\nabla_w E(w) = \sum_{n=1}^{N} \frac{(y_n - t_n) x_n}{y_n(1-y_n)} \times \frac{1}{\sqrt{2\pi}} e^{-(w^T x_n)^2}$$

$$\text{Hessian} = \sum_{n=1}^{N} \left\{ \frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \times \frac{1}{\sqrt{2\pi}} e^{-a_n^2} - \frac{2a_n(y_n - t_n)}{y_n(1-y_n)} \times x_n x_n^T \times \frac{1}{\sqrt{2\pi}} e^{-2a_n^2} \right\}$$

$x_n^T x_n$:داریم