درس یادگیری ماشین

استاد سید عباس حسینی

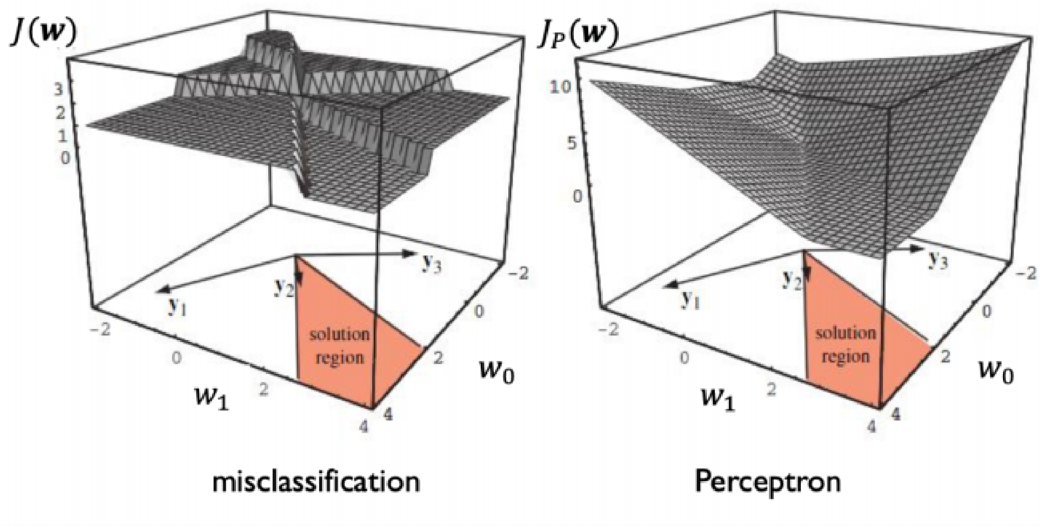
دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

تمرین پنجم | Classification | مصطفی قدیمی

سؤال ۱.

- آ) همان‌طور که در کلاس درس و اسلایدها [۱] نیز اشاره شد، SSE معیار خوبی اندازه‌گیری خطا برای دسته‌بندی نیست؛ زیرا حتی برای داده‌هایی که درست دسته‌بندی شده‌اند اما فاصله‌ی زیادی با خط دارند، جریمه‌ی [۲] سنگینی برای آن‌ها نیز در نظر می‌گیرد و معیار نزدیکی به دسته‌بند [۳] برای آن اهمیت دارد. در حالی‌که ممکن است دسته‌بندی داده‌ها به خوبی انجام گرفته باشد و خطای SSE بزرگ باشد.

- ب)



شکل ۱: مقایسه‌ی عمل‌کرد perceptron و شمردن داده‌های misclassified

با توجه به شکل، [۴] استفاده از روش perceptron criterion به ما این امکان را می‌دهد که به یک فضای محدب، [۵] و پیوسته برسیم. به همین خاطر از روش‌های iterative مانند gradient descent برای رسیدن به جواب بهینه می‌توان استفاده کرد. این در حالی است که همان‌طور که مشاهده می‌شود، در روش شمردن تعداد داده‌های اشتباه دسته‌بندی‌شده، به شکلی می‌رسیم که مشتق [۶] آن، صفر است و نمی‌توان از روش‌های iterative استفاده کرد.

---

[۱] اسلاید ۷، صفحه‌ی ۲۱.
[۲] penalty
[۳] classifier
[۴] ر.ک. اسلاید ۷، صفحه‌ی ۲۴.
[۵] convex
[۶] gradient

- پ) در حالت کلی سه روی‌کرد برای دسته‌بندی وجود دارد که دو نوع آن احتمالاتی و یکی نوع آن discriminant است. تابع خطای logistic regression از نوع احتمالاتی و perceptron از نوع discriminant است. مزایای logistic regression:
  - در مدل احتمالاتی، استنتاج [۷] از تصمیم‌گیری [۸] جدا می‌شود. [۹]
  - از لحاظ محاسباتی کارا است.
  - این امکان را به ما می‌دهد تا:
    * ریسک را حتی اگر ماتریس وزن‌دهی خطا [۱۰] عوض شد، کمینه کنیم.
    * Reject option داشته باشیم.
    * unbalanced class priors داشته باشیم.
    * مدل‌ها را ترکیب کنیم.

- ت) به دلیل غیرخطی بودن تابع sigmoid، برای logistic regression راه‌حل بسته‌ای [۱۱] وجود ندارد. روش IRLS [۱۲] می‌تواند سریع‌تر باشد. برای مثال اگر تابع $log-likelihood$ تقریبا درجه دو باشد، ممکن است فقط در چند مرحله به نقطه‌ی بهینه همگرا شود. به همین علت در کتاب Bishop نوشته شده است که: «اگر چه چنین روشی ممکن است منطقی به‌نظر برسد اما در حقیقت یک الگوریتم ضعیف است.» [۱۳]

  با توجه به دلایل مذکور، احتمال پیدا کردن جواب بهینه، در روش IRLS با تعداد عملیات کم‌تر، بیش‌تر است. [۱۴]

- ث) Probit regression نسبت به داده‌های پرت حساس‌تر است؛ زیرا اگر انتهای هر کدام از تابع‌ها را بررسی کنیم ، به نتیجه‌ی زیر می‌رسیم:

$$\begin{cases} logistic\,regression\,tails & \approx e^{-x} \\ probit\,regression & \approx e^{-x^2} \end{cases}$$

به همین دلیل اگر داده‌ی پرتی به تابع probit regression داده شود، وزن بیش‌تری به آن می‌دهد؛ بنابراین به داده‌های پرت نیز حساس‌تر است.

---

[۷] inference
[۸] decision
[۹] در استنتاج به دنبال تخمین $p(t|x)$ هستیم اما در تصمیم‌گیری برای x داده شده به دنبال t بهینه هستیم.
[۱۰] loss matrix
[۱۱] closed form solution
[۱۲] Iterative Reweighted Logistic Regression
[۱۳] سوال پرسیده شده در stackexchange
[۱۴] اسلاید ۷، صفحه‌ی ۶۰.

سؤال ۲.

همان‌طور که در اسلایدها [۱۵] دیدیم، داریم:

تعریف محدب بودن: به ازای هر دو نقطه دل‌خواه مانند A و B، که متعلق به ناحیه‌ی R_C هستند، خط متصل‌کننده‌ی آن دو نقطه نیز باید کاملا در آن ناحیه قرار داشته باشد.

$$\forall \alpha \in [0,\ 1] \rightarrow \hat{x} = \alpha x_A + (1-\alpha) x_B$$

$$F_C(x) = F_C(\alpha A + (1-\alpha)B) \xRightarrow{linearity\, of\, F} \alpha F_C(A) + (1-\alpha) F_C(B)$$

$$\rightarrow \forall d \in 1,\ ...,\ k : F_C(A) \geq F_d(A),\ F_C(B) \geq F_d(B) \implies \alpha F_C A \geq \alpha F_d(A),\ (1-\alpha)F_C(B) \geq (1-\alpha)F_d(B)$$

با توجه به نتیجه‌ی بالا، از جمع کردن دو طرف نامساوی داریم:

$$\forall d \in 1,\ ...,\ k : \alpha F_C(A) + (1-\alpha)F_C(B) \geq \alpha F_d(A) + (1-\alpha)F_d(B)$$

پس ثایت می‌شود که به ازای هر نقطه‌ی روی خط متصل‌کننده‌ی A و B داریم:

$$F_C(\theta) \geq F_d(\theta)$$

پس ثابت شد که θ نیز در ناحیه‌ی R_C است. بنابراین محدب بودن به این شکل ثابت می‌شود.

[۱۵] اسلاید ۷، صفحه ۱۷

سؤال ۳.

$$Ak \leq ||w^{k+1}|| \rightarrow w^{k+1}.w^* = (w^k + t^{(i)}x^{(i)})w^* \implies w^{(k)}.w^{(*)} + t^{(i)}(w^* x^{(i)}) \geq w^{(k)}w^* + \lambda$$

$$\implies w^{(k+1)}w^* \geq k\lambda$$

همان‌طور که در صورت سوال گفته شده است، اگر w را از صفر شروع کنیم، به کمک استقرا داریم:

$$||w^{(k+1)}|| \geq k\lambda$$

$$||w^{(k+1)}|| \leq \beta\sqrt{k},\ ||w^{k+1}||^2 = ||w^{(k)} + t^i x^i||^2$$

$$\rightarrow ||w^{(k)}||^2 + ||x^{(i)}||^2(t^{(i)})^2 + 2t^{(i)}(w^{(k)}x^{(i)}) \leq ||w^{(k)}||^2 + ||x^{(i)}||^2(t^{(i)})^2$$

$$\begin{cases} t^{(i)}(x^{(k)}x^{(i)}) \leq 0 \\ (t^{(i)})^2||x^{(i)}||^2 = ||x^{(i)}||^2 = R^2 \\ (t^{(i)})^2 = 1 \end{cases} \implies ||w^{(k+1)}||^2 \leq kR^2$$

$$(w^{(i)})^2 = 0 \rightarrow ||w^{(k+1)}|| \leq \sqrt{k}R \rightarrow k\lambda \leq ||w^{(k+1)}|| \leq \sqrt{k}R \rightarrow k \leq (\frac{R}{\lambda})^2$$

سؤال ۴.

$$p(C_1) = \alpha,\ p(C_2) = 1 - \alpha$$

حال اگر یک مجموعه‌ی داده [۱۶] با n داده داشته باشیم داریم:

$$p(x_n, C_1) = p(C_1)p(x_n|C_1) = \alpha N(x_n|\mu_1, \Sigma)$$

$$p(x_n, C_2) = p(C_2)p(x_n|C_2) = (1-\alpha)N(x_n|\mu_2, \Sigma)$$

حال اگر تابع likelihood را بنویسیم، داریم:

$$p(T|\alpha, \mu_1, \mu_2, \Sigma) = \Pi_{n=1}^{N}[\alpha N(x_n|\mu_1, \Sigma)]^{t_n}[(1-\alpha)N(x_n|\mu_2, \Sigma)]^{1-t_n}$$

با لگاریتم و سپس مشتق گرفتن داریم:

$$\alpha = \frac{1}{N}\Sigma_{n=1}^{N} t_n$$

حال اگر با توجه به μ_1 مشتق (گرادیان) بگیریم داریم:

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

$$\implies \begin{cases} \mu_1 = \frac{1}{N}\Sigma_{n=1}^{N} t_n x_n \\ \mu_2 = \frac{1}{N}\Sigma_{n=1}^{N}(1-t_n)x_n \end{cases}$$

$$-\frac{1}{2}\Sigma_{n=1}^{N} t_n ln|\Sigma| - \frac{1}{2}\Sigma_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1}(x_n - \mu_1) - \frac{1}{2}\Sigma_{n=1}^{N}(1-t_n) ln|\Sigma| - \frac{1}{2}\Sigma_{n=1}^{N}(1-t_n)(x_n - \mu_2)^T \Sigma_{-1}(x_n - \mu_2)$$

$$= -\frac{N}{2} ln|\Sigma| - \frac{N}{2} tr(\Sigma_{-1} S)$$

$$S = \frac{N_1}{N}S_1 + \frac{N_2}{N}S_2 \implies \begin{cases} S_1 = \frac{1}{N_1}\Sigma_{n\in C_1}(x_n - \mu_1)(x_n - \mu_1)^T \\ S_2 = \frac{1}{N_2}\Sigma_{n\in C_1}(x_n - \mu_2)(x_n - \mu_2)^T \end{cases} \quad \to \Sigma = S$$

---

[۱۶] dataset

سؤال ۵.

همان‌طور که در آخر صفحه ۲۰۶ کتاب Bishop توضیح داده است، داریم:

طبق تعریف اگر مجموعه داده، به‌صورت خطی جداپذیر باشد، می‌توان یک w پیدا کرد که برای بعضی از نقاط $w^T\phi(x_n) > 0$ و برای دیگر نقاط $w^T\phi(x_m) < 0$ باشد (مقدار آن برابر با صفر نمی‌تواند باشد، چون کلا دو کلاس داریم که بر اساس مثبت یا منفی بودن، دسته‌بندی می‌شوند.).

حال دسته‌ی اول را داده‌هایی فرض می‌کنیم که برای آن‌ها رابطه‌ی $w^T\phi(x_n) > 0$ برقرار است و برای دیگر داده‌ها فرض می‌کنیم که در دسته‌ی دوم هستند (مقدار منفی دارند.).

$$p(C_1|\phi) = y(\phi) = \sigma(w^T\phi)$$

حال اگر $|w| \to \infty$ باشد، داریم:

$$p(C_1|\phi(x_n)) = \sigma(w^T\phi(x_n)) \to 1$$

با توجه $\begin{cases} w^T\phi(x_n) \to +\infty \\ w^T\phi(x_m) \to -\infty \end{cases}$ داریم:

$$p(C_2|\phi(x_m)) = 1 - p(C_1|\phi(x_m)) = 1 - \sigma(w^T\phi(x_m)) \to 1$$

به بیان دیگر برای تابع likelihood اگر $|w| \to \infty$ باشد، همه‌ی داده‌ها به بیش‌ترین مقدارشان یعنی ۱ می‌رسند.

بنابراین برای مجموعه داده‌های خطی جداپذیر، فرآیند یادگیری ممکن است به سمت $|w| \to \infty$ سوق پیدا کند و از boundaryهای خطی برای برچسب‌زدن مجموعه داده استفاده کند که ممکن است باعث overfitting شود.

سؤال ۶.

به دلیل افزایش ویژگی‌ها کارایی در مدل D_2 کاهش می‌یابد. اما به دلیل یکسان بودن (هم‌بسنگی ۱) تغییری در دقت آن به وجود نمی‌آید.

سؤال ۷.

$$\int_{-\infty}^{+\infty} N(\theta|0,1)d\theta = 1 \xRightarrow{Symmetric} \begin{cases} \int_{-\infty}^{0} N(\theta|0,1)d\theta = \frac{1}{2} \\ \int_{0}^{\infty} N(\theta|0,1)d\theta = \frac{1}{2} \end{cases}$$

$$\phi(a) = \int_{-\infty}^{a} N(\theta|0,1)d\theta = \int_{-\infty}^{0} N(\theta|0,1)d\theta + \int_{0}^{a} N(\theta|0,1)d\theta = \frac{1}{2} + \int_{0}^{a} \frac{1}{2\pi} e^{\frac{-\theta^2}{2}d}\theta$$

با اعمال تغییر متغیر و انتگرال‌گیری خواهیم داشت:

$$\phi(a) = \frac{1}{2} + \frac{1}{2}\int_{0}^{\frac{a}{\sqrt{2}}} \frac{2}{\sqrt{\pi}} e^{-\theta^2} d\theta = \frac{1}{2}(1 + \frac{2}{\sqrt{\pi}}\int_{0}^{\frac{a}{\sqrt{2}}} e^{-\theta^2} d\theta) \implies \phi(a) = \frac{1}{2}(1 + erf(\frac{a}{\sqrt{2}}))$$

سؤال ۸.

$$\frac{\partial E}{\partial y_n} = \frac{y_n - t_n}{y_n(1-y_n)},\ \nabla a_n = \phi_n$$

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \phi(a_n)}{\partial a_n} = \frac{1}{\sqrt{2\pi}} e^{-a_n^2}$$

با توجه به دو معادله بالا و ترکیب آن‌ها داریم:

$$\nabla E = \Sigma_{n=1}^{N} \frac{\partial E \partial y_n}{\partial y_n \partial a_n} \nabla a_n = \Sigma_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)\sqrt{2\pi}} e^{-a_n^2} \phi_n$$

$$\frac{\partial}{\partial y_n} \frac{(y_n - t_n)}{y_n(1-y_n)} = \frac{y_n(1-y_n)}{y_n^2(1-y_n)^2} - \frac{(y_n - t_n)(1-2y_n)}{y_n^2(1-y_n)^2} = \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2}$$

$$\implies \nabla\nabla E = \Sigma_{n=1}^{N} \frac{\partial}{\partial y_n} [\frac{y_n - t_n}{y_n(1-y_n)}] \frac{1}{\sqrt{2\pi}} e^{-a_n^2} \phi_n \nabla y_n + \frac{y_n - t_n}{y_n(1-y_n)} \frac{e^{-a_n^2}(-2a_n)\phi_n \nabla a_n}{\sqrt{2\pi}}$$

$$\nabla\nabla E = \Sigma_{n=1}^{N} (\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \frac{1}{\sqrt{(2\pi)}} e^{-a_n^2} - 2a_n(y_n - t_n)) \frac{e^{-2a_n^2} \phi_n \phi_n^T}{\sqrt{2\pi} y_n(1-y_n)}$$