• به نام خدا •

• تمرین سری اول - تئوری

- یادگیری عمیق -

" پوریا دادخواه _ ۴۰۱۲۰۱۳۸۱

سوال ① - برای طبقه بند Bayes m کلاسه داریم: (از فرمول بیز و طبقه بند اپتیمال)

$$\Omega_j = \arg\max_{\Omega_j} P(\Omega_j)\, P(x/\Omega_j) \qquad j=1..m$$

$\Delta =$ $m=3$ ، برای هر ۳ برچسب $P(\Omega_j)=\frac{1}{3}$ است ، بنابراین کافی است

$P(x|\Omega_j)$ برای هر ۳ برچسب حساب کرده و MAX را انتخاب کنیم. از آنجایی که

$j=1,2,3$

توزیع هر کدام نرمال است ، $n=2$ داریم: $(x|y=i) \sim N(\mu_i, \Sigma_i)$

$$\Rightarrow P(x;\mu_i,\Sigma_i) = \frac{1}{(2\pi)\,|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right)$$

$$\mu_1 = \begin{pmatrix}0\\0\end{pmatrix},\ \Sigma_1 = \begin{pmatrix}0.7 & 0\\ 0 & 0.7\end{pmatrix} \Rightarrow |\Sigma_1| = 0.49,\ \Sigma_1^{-1} = \frac{100}{49}\begin{pmatrix}0.7 & 0\\ 0 & 0.7\end{pmatrix} : \underline{i=1}$$

$$\mu_2 = \begin{pmatrix}1\\1\end{pmatrix},\ \Sigma_2 = \begin{pmatrix}0.8 & 0.3\\ 0.3 & 0.2\end{pmatrix} \Rightarrow |\Sigma_2| = 0.07,\ \Sigma_2^{-1} = \frac{100}{9}\begin{pmatrix}0.2 & -0.3\\ -0.3 & 0.8\end{pmatrix} : \underline{i=2}$$

$$\mu_3 = \begin{pmatrix}1\\1\end{pmatrix},\ \Sigma_3 = \begin{pmatrix}0.7 & 0.2\\ 0.2 & 0.8\end{pmatrix} \Rightarrow |\Sigma_3| = 0.52,\ \Sigma_3^{-1} = \frac{100}{52}\begin{pmatrix}0.8 & -0.2\\ -0.2 & 0.7\end{pmatrix} : \underline{i=3}$$

Scanned by CamScanner

الف) $x=\begin{pmatrix}50\\0.5\end{pmatrix}$

$$P(x|\Omega_1) = \frac{1}{2\pi\times 0.7} \times \exp\left(-\frac{1}{2}\times\frac{100}{49}\begin{pmatrix}50 & 0.5\end{pmatrix}\begin{pmatrix}0.7 & 0\\0 & 0.7\end{pmatrix}\begin{pmatrix}50\\0.5\end{pmatrix}\right)$$

[~ تقریبا]

$$= \frac{1}{1.4\pi}\exp\left(-\frac{50}{49}\times 1750.175\right) \Rightarrow \ln(P_1) = -1787.3$$

$$P(x|\Omega_2) = \frac{1}{0.2\pi\times\sqrt{7}}\exp\left(-\frac{1}{2}\times\frac{100}{9}\begin{pmatrix}49 & -0.5\end{pmatrix}\begin{pmatrix}0.2 & -0.3\\-0.3 & 0.8\end{pmatrix}\begin{pmatrix}49\\-0.5\end{pmatrix}\right)$$

$$= \frac{1}{0.2\pi\times\sqrt{7}}\exp\left(\frac{-50}{9}\times 495.1\right) \Rightarrow \ln(P_2) = -2751.06$$

$$P(x|\Omega_3) = \frac{1}{0.2\pi\times\sqrt{52}}\exp\left(-\frac{1}{2}\times\frac{100}{52}\begin{pmatrix}49 & -0.5\end{pmatrix}\begin{pmatrix}0.8 & -0.2\\-0.2 & 0.7\end{pmatrix}\begin{pmatrix}49\\-0.5\end{pmatrix}\right)$$

$$= \frac{1}{0.2\pi\times\sqrt{52}}\exp\left(-\frac{50}{52}\times 1930.78\right) \Rightarrow \ln(P_3) = -1858.03$$

⇐ $x\in\Omega_1$ است.

ب) $$P(x|\Omega_1) = \frac{1}{1.4\pi}\exp\left(\frac{-50}{49}\begin{pmatrix}0.5 & 0.5\end{pmatrix}\begin{pmatrix}0.7 & 0\\0 & 0.7\end{pmatrix}\begin{pmatrix}0.5\\0.5\end{pmatrix}\right) = \frac{1}{1.4\pi}\exp(-0.357)$$

$$P(x|\Omega_2) = \frac{1}{0.2\pi\sqrt{7}}\exp\left(\frac{-50}{9}\begin{pmatrix}-0.5 & -0.5\end{pmatrix}\begin{pmatrix}0.2 & -0.3\\-0.3 & 0.8\end{pmatrix}\begin{pmatrix}-0.5\\-0.5\end{pmatrix}\right) = \frac{1}{0.2\pi\sqrt{7}}\exp(-0.56)$$

$$P(x|\Omega_3) = \frac{1}{0.2\pi\sqrt{7}}\exp\left(-\frac{50}{52}\begin{pmatrix}-0.5 & -0.5\end{pmatrix}\begin{pmatrix}0.8 & -0.2\\-0.2 & 0.7\end{pmatrix}\begin{pmatrix}-0.5\\-0.5\end{pmatrix}\right) = \frac{1}{0.2\pi\sqrt{7}}\exp(-0.264)$$

⇐ $P(x|\Omega_3)$ بیشتر است ⇐ $x\in\Omega_3$.

707 Note Book

Scanned by CamScanner

سؤال ۴

$$\tilde{E}_D(w) = \frac{1}{2}\sum_n \left(w_0 + \sum_i w_i x_{ni} + \sum_i w_i \epsilon_{ni} - y_n \right)^2$$

$$= \frac{1}{2}\sum_n \left(w_0 + \sum_i w_i x_{ni} - y_n \right)^2 + \frac{1}{2}\sum_n \sum_i w_i^2 \epsilon_{ni}^2$$

$$+ \sum_n \left(\sum_i w_i \epsilon_{ni} \right)\left(w_0 + \sum_i w_i x_{ni} - y_n \right)$$

⇐ از عبارت فوق امید ریاضی می‌گیریم (به دلیل خطی بودن E، امید هر ترم جدا حساب شده و جمع می‌شود.

$$E[\tilde{E}_D(w)] = \underbrace{E_D(w)}_{\text{ترم اول}} + \underbrace{\frac{1}{2}\sum_n \sum_i w_i^2 \sigma^2}_{\text{ترم دوم}} + \underbrace{\sum_n \left(\sum_i w_i \underbrace{E(\epsilon_{ni})}_{0} \right)(\cdot)}_{\text{ترم سوم}}$$

$$= E_D(w) + \frac{N\sigma^2}{2}\sum_i w_i^2 + 0$$

سایر ترم‌ها ثابت بوده و امید نویز هم صفر است.

$$\Rightarrow \boxed{E[\tilde{E}_D(w)] = E_D(w) + \frac{N\sigma^2}{2}\sum_i w_i^2}$$

Scanned by CamScanner

سوال ③: الف)

class 1: $\ln\left(\frac{P_1}{P_K}\right) = \beta_1^T x$

class 2: $\ln\left(\frac{P_2}{P_K}\right) = \beta_2^T x$

class $_{K-1}$: $\ln\left(\frac{P_{K-1}}{P_K}\right) = \beta_{K-1}^T x$

class K: $1 - \sum_{i=1}^{K-1} P_i$

$\xrightarrow{\Pr(\cdot)}$

class 1: $P_1 = \dfrac{e^{\beta_1^T x_i}}{1 + \sum_{i=1}^{K-1} e^{\beta_i^T x_i}}$

class 2: $P_2 = \dfrac{e^{\beta_2^T x_i}}{1 + \sum_{i=1}^{K-1} e^{\beta_i^T x_i}}$

class $_{K-1}$: $P_{K-1} = \dfrac{e^{\beta_{K-1}^T x_i}}{1 + \sum_{i=1}^{K-1} e^{\beta_i^T x_i}}$

class K: $P_K = \dfrac{1}{1 + \sum_{i=1}^{K-1} e^{\beta_i^T x_i}}$

با ارسیت الف احتمال یا بند.

تخمین پارامتر از B_i استفاده کردیم.

$$L(\beta_i^T, x) = \sum_{i=1}^{N} \log(P_i(x_i, \beta_i^T)) = \sum_{i=1}^{N} \log\left(\frac{e^{\beta_i^T x_i}}{1 + e^{\beta_i^T x_i}}\right)$$ ب)

ج)

$$\frac{\partial}{\partial \beta} L(\beta, x_i) = \sum_{i=1}^{N} \frac{\partial}{\partial \beta} y_i \beta^T x_i - \frac{\partial}{\partial \beta} \ln(1 + e^{\beta^T x_i})$$

$$= \sum_{i=1}^{N} y_i x_i - \frac{x_i e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} = \sum_{i=1}^{N} x_i (y_i - P_i)$$

$$\nabla L(\beta, X) = \tilde{X}^T (Y - P)$$

ح

Scanned by CamScanner

سوال ۴ - الف) اگر ستون ها را فقط روی یک ویژگی آموزش دهیم:
ویژگی j که آموزش دهیم.

باید w ها را بیابیم که هزینه را min کند ← $\hat{y} = w_j x_j + \varepsilon$: از رابطه مدل

$$\Rightarrow \arg\min \sum (y_i - w_j x_{ij})^2$$ (جمع مربعات در نقطه مینیمم)

$$\longrightarrow \frac{\partial}{\partial w_j}(\cdot) = 0 \Rightarrow \sum 2x_{ij}(y_i - w_j x_{ij}) = 0$$

$$\Rightarrow 2\sum(x_{ij} y_i) - 2w_j \sum x_{ij}^2 = 0 \Rightarrow w_j = \frac{\sum x_{ij} y_i}{\sum x_{ij}^2}$$

ترم $x_{ij}\, y_i$، حاصل ضرب داخلی ستون ویژگی x_j و ضریب y است و

ترم x_{ij}^2، جمع مربعات ستون ویژگی x_{ij} است که ضرب داخلی با خودش است ⇐

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

ب) متعامد بودن ستون ها، یعنی مستقل بودن ویژگی ها داده است.

در سناریوی آموزش روی همه ویژگی ها و آموزش مستقل هر ویژگی را بررسی می کنیم و نشان می دهیم نتیجه یکسان است.

ادامه صفحه بعد

Scanned by CamScanner

۱) آموزش روی همه ویژگی‌ها (همزمان)

در این حالت رگرسیون خطی چند متغیره را فیت می‌کنیم که y ترکیب خطی از تمام ویژگی‌ها است:

ماتریس ویژگی‌ها

$y = X \cdot \beta + \varepsilon$ → ترم اخلال   برای یافتن بردار β که هزینه جمع مربعات اخلال →

بردار هدف   بردار ضرایب

را کمینه کند باید دستگاه معادلات عادی روبه‌رو را حل کنیم:

$$X^T \cdot X \cdot \beta = X^T \cdot y$$

۲) آموزش مستقل هر ویژگی:

در این حالت یک رگرسیون خطی تک متغیره مستقل را برای هر ویژگی فیت می‌کنیم. برای ویژگی j داریم:

(ستون j از ماتریس کل) بردار ویژگی j

$y = w_j x_j + \varepsilon$

بردار هدف   ضریب   → برای یافتن پارامتر بهینه w_j باید عبارت زیر که از قسمت قبل

استخراج شده استفاده می‌کنیم:

$$w_j = \frac{x_j^T \cdot y}{x_j^T \cdot x_j} \quad \forall j$$

اکنون باید نشان دهیم پارامترهای بهینه هر دو روش یکسان اند:

در دستگاه معادله $X^T \cdot X \cdot \beta = X^T \cdot y$ → $i \neq j$ so $x_i^T \cdot x_j$ → چون ستون X متعامدند

تمام درایه‌های غیر قطر اصلی $X^T \cdot X$ صفر اند، پس ماتریس قطری است و هر معادله β مستقل از بقیه است.

و معادله حل برای w_j جداگانه است.

Scanned by CamScanner

وزن‌ها آموزشی

$$y = w_j \times x_j + w_0 + \varepsilon$$

ج

ویژگی انتخاب شده

$$\Rightarrow w_0, w_j = \underset{w_0, w_j}{\arg\min}\left(\sum \left(y_i - (w_j x_{ij} + w_0)\right)^2 \right)$$

w_j: $\xrightarrow{\partial / \partial w_j (\cdot) = 0}$ $\sum -2\left(y_i - (w_j \cdot x_{ij} + w_0)\right) \times (-x_{ij}) = 0$

$$\Rightarrow \sum \left(y_i x_{ij} - w_j x_{ij}^2 - w_0 x_{ij}\right) = 0$$

$$\Rightarrow w_j \sum x_{ij}^2 = \sum y_i \times x_{ij} - \sum w_0 \times x_{ij}$$

$$\Rightarrow w_j = \frac{\sum y_i x_{ij} - \sum w_0 x_{ij}}{\sum x_{ij}^2} = \frac{Cov(x_j, y)}{Var(x_j)} - \frac{w_0 \cancel{\sum} x_{ij}}{\cancel{\sum} x_{ij}^2}$$

✓

w_0: $\xrightarrow{\partial / \partial w_0 (\cdot) = 0}$ $\sum -2(y_i - w_j x_{ij} + w_0) \times (-1) = 0$

$$\Rightarrow \sum y_i - w_j \times \sum x_{ij}^2 - \sum w_0 = 0 \Rightarrow w_0 = \frac{\sum y_i}{N} - w_j \times \frac{\sum x_{ij}}{N}$$

$$\Rightarrow w_0 = E(y) - w_j E(x)$$ ✓

Scanned by CamScanner

سوال ۵ :

حکم) $P(X \ge \alpha) \le \dfrac{E(X)}{\alpha}$

الف) فرض) $0 \le X$ که طبیعتاً α هم باید نامنفی باشد وگرنه احتمال اعمال نمی‌شود.

اثبات) $$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx \xrightarrow{0 \le X} E(X) = \int_{0}^{\infty} x f(x)\,dx$$

$$\Rightarrow E(X) = \int_{0}^{\alpha} x f(x)\,dx + \int_{\alpha}^{\infty} x f(x)\,dx \ge \int_{\alpha}^{\infty} x f(x)\,dx \ge \int_{\alpha}^{\infty} \alpha f(x)\,dx$$

$$\Rightarrow E(X) \ge \int_{\alpha}^{\infty} \alpha f(x)\,dx = \alpha \int_{\alpha}^{\infty} f(x)\,dx = \alpha \Pr(X \ge \alpha)$$

$$\Rightarrow \boxed{\Pr(X \ge \alpha) \le \frac{E(X)}{\alpha}} \quad \checkmark$$

ب) در نامساوی مارکوف متغیر تصادفی هدف (X) را برابر $Y = (Z-\mu)^2$ قرار می‌دهیم که یک متغیر تصادفی نامنفی است و شرط فرض را برآورده می‌کند و مقدار α را برابر $\alpha = \varepsilon^2$ قرار می‌دهیم. خواهیم داشت:

$$\boxed{\Pr(|Z-\mu| \ge \varepsilon) = \Pr\left((Z-\mu)^2 \ge \varepsilon^2\right) \underset{\text{مارکوف}}{\le} \frac{E\left((Z-\mu)^2\right)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}}$$

$$\Rightarrow \Pr(|Z-\mu| \ge \varepsilon) \le \frac{\sigma^2}{\varepsilon^2} \quad \checkmark$$

Scanned by CamScanner

ج) طبق مسائل تعریف شده ، x_i را متغیر تصادفی نقطه i ام انتخابی می‌نامیم که

از این رو احتمال انتخاب می‌شود اگر داخل دایره بود $x_i = 1$ و در غیر این صورت $x_i = 0$

در این صورت پس از انتخاب n نقطه می‌توانیم تخمین مقدار π را به صورت زیر در نظر بگیریم:

$$\hat{\pi}(n) = 4 \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

چرا که نسبت مساحت دایره به کل برابر $\frac{\pi}{4}$ می‌باشد.

$$\begin{cases} E(x_i) = \overbrace{\frac{\pi}{4}}^{p} \times 1 + \overbrace{\left(1 - \frac{\pi}{4}\right)}^{q = 1-p} \times 0 = \frac{\pi}{4} \\ V(x_i) = \frac{\pi}{4} \times \left(1 - \frac{\pi}{4}\right) \end{cases}$$

(احتمال موفقیت از هر x_i برابر $\frac{\pi}{4}$ است.)

می‌توان نشان داد این تخمین بدون بایاس است و تخمین صحیحی می‌باشد:

$$E(\hat{\pi}) = E\left(\frac{4}{n} \sum_{i=1}^{n} x_i\right) = \frac{4}{n} \sum_{i=1}^{n} E(x_i) = \pi$$

$$V(\hat{\pi}) = V\left(\frac{4}{n} \sum_{i=1}^{n} x_i\right) = \frac{16}{n^2} \sum_{i=1}^{n} V(x_i) = \frac{\pi(4-\pi)}{n}$$

حال می‌خواهیم با احتمال حداقل $\varepsilon = 95\%$ ، خطای تخمین کمتر از $\delta = 1\%$ باشد:

$$P\left(|\hat{\pi}(n) - \pi| < \delta\right) > \varepsilon \xrightarrow[\text{چبیشف}]{\text{معادل}} P\left(|\hat{\pi}(n) - \pi| \geq 0.01\right) \leq 1 - 0.95$$

یعنی از چبیشف داریم ، $\frac{V(\hat{\pi})}{\delta^2} \leq 0.05$ ← ادامه صفحه بعد

Scanned by CamScanner

داریم

$$v(\hat{\pi}) = \frac{\pi(4-\pi)}{n}$$

$$\Rightarrow \frac{\pi(4-\pi)}{n(0.01)^2} \leq 0.05$$

$$\Rightarrow n \geq 5.393532 \times 10^5$$

$$\Rightarrow \boxed{n \geq 539353}$$

Scanned by CamScanner

سوال ۶- الف)

چون A ماتریس مربعی وارون پذیر است، مقادیر تکین A، ریشه مربعی مقادیر ویژه

AA^T یا به طور معادل A^TA است. از آن جایی که $A^{-1}(A^{-1})^T = (A^TA)^{-1}$

مقادیر تکین A^{-1}، عکس مقادیر تکین A هستند و اگر σ_1، بیشینه مقدار تکین A و

σ_n کمینه آن باشد، $\sigma_{MAX}(A^{-1})\,\sigma_{MAX}(A) = \frac{\sigma_1}{\sigma_n}$ می‌شود و چون $\sigma_1 \geq \sigma_n$

پس داریم: $\sigma_{MAX}(A^{-1})\,\sigma_{MAX}(A) \geq 1$ ✓

ب) $\|A\|_2 \leq \|A\|_F \leq \sqrt{rank(A)}\,\|A\|_2$ (صحیح)

قسمت اول (نامساوی) $\|A\|_2 \leq \|A\|_F$ ؟

$\|A\|_2 =$ بزرگترین مقدار تکین A

$\|A\|_F = \sqrt{\sum |a_{ij}|^2}$ روی همه المان‌های A

که U و V متعامدند و Σ ماتریس قطری از مقادیر تکین $\Leftarrow$ $A = U\Sigma V^T$: بازنویسی A

$\sigma_1 \dots \sigma_k$ است که $rank(A)=k$ است.

$\Rightarrow \|A\|_F = \sqrt{\sum_{j=1}^{k}(\sigma_j)^2}$, $\|A\|_2 = \sigma_1$

ادامه صفحه بعد.

Scanned by CamScanner

از دو عبارت درست آمده واضح است که $\sigma_1^2 \le \sum_{j=1}^{k} \sigma_j^2 \Rightarrow \sigma_1 \le \sqrt{\sum(\sigma_j^2)}$

چرا که ترم سمت راست جمع σ_1^2 و مقادیر نامنفی بوده است.

قسمت دوم نامساوی) $\|A\|_F \le \sqrt{\underbrace{rank(A)}_{k}} \cdot \|A\|_2$ ؟

در قسمت قبل دیدیم: $\|A\|_F = \sqrt{\sum_{j=1}^{k} \sigma_j^2} \le \sqrt{k} \times \sigma_1 = \sqrt{rank(A)} \times \|A\|_2$

σ_1 بیشتر از σ_j هاست و اگر همه را با σ_1 جایگزین کنیم عبارت زیر رادیکال بزرگتر خواهد شد. تعریف $\|A\|_2$.

بنابراین کل نامساوی اثبات شد. $\|A\|_2 \le \|A\|_F \le \sqrt{rank(A)} \cdot \|A\|_2$ ■

(لم)

سوال ۷ : ابتدا رابطه زیر را بین تابع sigmoid و tanh ثابت می‌کنیم

$$\tanh(a) = 2\sigma(2a) - 1, \text{ where: } \sigma(z) = \frac{1}{1+e^{-z}}, \quad \tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$$

اثبات)

$$2\sigma(2a) - 1 = \frac{2}{1+e^{-2a}} - 1 = \frac{2-1-e^{-2a}}{1+e^{-2a}}$$

$$= \frac{1-e^{-2a}}{1+e^{-2a}} = \frac{\cancel{e^{-a}}(e^{a}-e^{-a})}{\cancel{e^{-a}}(e^{a}+e^{-a})} = \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}} = \tanh(a) \checkmark$$

حال با لم ثابت شده از رابطه خطی داده شده برای sigmoid شروع کرده و به

رابطه خطی با tanh می‌رسیم : هم چنین $y_j = \frac{x-\mu_j}{s}$ را داریم.

$$\boxed{y(x,w)} = w_0 + \sum_{j=1}^{n} w_j\, \sigma(2y_j) \xrightarrow{\sigma(2a) = (\tanh(a)+1)\times\frac{1}{2}}$$

$$= w_0 + \sum w_j \left(\frac{\tanh(y_j)+1}{2}\right)$$

$$= \underbrace{w_0 + \tfrac{1}{2}\sum_{j=1}^{n} w_j}_{u_0} + \sum_{j=1}^{n} \underbrace{\frac{w_j}{2}}_{u_j} \cdot \tanh(y_j)$$

$$= u_0 + \sum_{j=1}^{n} u_j \tanh(y_j) = y(x,u) \checkmark$$