



## یادگیری عمیق

۱۴۰۲ پاییز

استاد: دکتر فاطمی زاده

گردآورندها: علیرضا خالقی، امیررضا حاتمی

دانشگاه صنعتی شریف  
دانشکده‌ی مهندسی برق

تمرين اول مفاهيم پايه مهلت ارسال: جمعه ۵ آبان (با احتساب تاخیر)

تمرين اول

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.

• در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمرین‌ها سقف ۵ روز و در مجموع ۲۰ روز وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال شده پذیرفته نخواهد بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.

• هم‌کاری و همفکری شما در انجام تمرین مانع ندارد اما پاسخ ارسالی هر کس حتماً باید توسط خود او نوشته شده باشد.  
(دقت کنید در صورت تشخیص مشابهت غیرعادی برخورد جدی صورت خواهد گرفت.)

• 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.

• لطفاً تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

• نتایج و پاسخ‌های خود را در یک فایل با فرمت zip به نام StudentNumber-Name HW1 قرار دهید. برای بخش عملی تمرین نیز لینک گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار بدهید. دقت کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید. لطفاً تمامی سوالات خود را از طریق کوئی‌ای درس مطرح بکنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترها دیگر پاسخ داده نخواهد شد).

• دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطای هنگام اجرای کدتان، حتی اگه خطای بدلیل اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

## سوالات نظری (۱۰۰ نمره)

۱. (۱۰ نمره) در مدل Naive Bayes فرض می‌کنیم که با شرط داشتن برچسب داده‌ها، تمامی ویژگی‌ها از هم دیگر مستقل هستند. اما در واقعیت با مشخص بودن برچسب داده‌ها باز هم ممکن است که ارتباطی بین ویژگی‌ها وجود داشته باشد. فرض کنید که یک مسئله‌ی classification داریم. برچسب یک داده را با  $Y$  و ویژگی‌های آن را با  $X$  نشان می‌دهیم.

فرض کنید که برچسب داده‌ها می‌توانند ۳ مقدار مختلف اختیار کند و داریم:

$$P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}$$

در این مسئله فرض کنید  $n = 2$  هم چنین توزیع توانمند این دو متغیر تصادفی، توزیع نرمال ۲ متغیره است.

$$1 \leq i \leq 3 : (X|Y = i) \sim N(\mu_i, \Sigma_i)$$

هم چنین میدانیم که:

$$\mu_1 = [0, 0]^T, \mu_2 = [1, 1]^T, \mu_3 = [1, 1]^T$$

$$\Sigma_1 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.7 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 0.8 & 0.3 \\ 0.3 & 0.2 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 0.7 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$$

$\textcircled{1} \quad \mathbf{x}_1 = \begin{bmatrix} 50 \\ 0.5 \end{bmatrix}$ 
 $\textcircled{2} \quad \mathbf{x}_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ 
 $\propto p(\mathbf{x}_j | \mathbf{x}) \frac{1}{2\pi|C|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_j - \mu_j)^T C^{-1} (\mathbf{x}_j - \mu_j)\right)$

$\Rightarrow \mathbf{r}^* = \underset{\mathbf{r}_j}{\operatorname{argmax}} \underbrace{p(\mathbf{r}_j) p(\mathbf{x} | \mathbf{r}_j)}_{\text{is } 1/3 \text{ for all classes}} = \underset{\mathbf{r}_j}{\operatorname{argmax}} p(\mathbf{x} | \mathbf{r}_j)$

$p(\mathbf{x}_j | \mathbf{r}_j) = \begin{cases} \mathbf{r}_j = 1 \rightarrow \frac{1}{2\pi|C_1|^{1/2}} \times e^{-\frac{1}{2}(35\gamma_1 - \dots)} \\ \mathbf{r}_j = 2 \rightarrow \dots \times \exp(-\frac{1}{2}\mathbf{x}^T \mathbf{C}_2 \mathbf{x}_j) \\ \mathbf{r}_j = 3 \rightarrow \dots \times \exp(-\frac{1}{2}\mathbf{x}^T \mathbf{C}_3 \mathbf{x}_j) \end{cases}$

It seems like an estimation! but the  $\boxed{Y=2}$  is argmax!

$p(\mathbf{x}_2 | \mathbf{r}_j) = \begin{cases} \mathbf{r}_j = 1 \rightarrow 0.159 \\ \mathbf{r}_j = 2 \rightarrow 0.29 \\ \mathbf{r}_j = 3 \rightarrow 0.169 \end{cases} \Rightarrow \boxed{\text{label is } Y=2.}$

حال اگر ورودی های زیر را داشته باشیم، برچسب داده ها را بدست آورید:

$$(آ) \quad x = [50, 0/5]$$

$$(ب) \quad x = [0/5, 0/5]$$

(۱۵ نمره) یک مدل خطی به فرم زیر را در نظر بگیرید:

$$y(x_n, \omega) = \omega_0 + \sum_{i=1}^D \omega_i x_{ni}$$

خطای آن را به صورت زیر در نظر میگیریم:

$$E_D(w) = \frac{1}{n} \sum_{n=1}^N [y(x_n, w) - y_n]^2$$

حال فرض کنید که یک نویز گوسی  $\epsilon_i \sim N(0, \sigma^2 I)$  به هر ورودی  $x_i$  اضافه شده است.  $\epsilon_i$  ها به صورت i.i.d تولید شده اند.

اگر  $\tilde{E}_D(w)$  خطای مدل وقتی از  $\epsilon_i + x_i$  استفاده می کنیم باشد آنگاه امید ریاضی این عبارت را پیدا کنید.

۳. (۱۵ 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) حال Log likelihood زیر را برای n نمونه‌ی زیر ساده کنید:

$$\text{Samples : } (x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

$$L(w_1, w_2, \dots, w_{k-1}) = \sum_{i=1}^n \ln P(Y = y_i | X = x_i)$$

پ) گرادیان L را نسبت به هریک از  $w_k$  ها بباید و آن را ساده کنید.

ت) تابع هدف زیر را در نظر بگیرید. گرادیان f را با توجه به هریک از  $w_k$  ها بباید.

$$f(w_1, \dots, w_{k-1}) = L(w_1, w_2, \dots, w_{k-1}) - \frac{\lambda}{2} \sum_{j=1}^{k-1} |w_j|^2$$

(۱۵ نمره)

فرض کنید n داده آموزش با m ویژگی داریم که ماتریس این داده ها را  $X_{nm}$  در نظر میگیریم. برچسب داده ها نیز به صورت  $[y_1, \dots, y_n] = y$  میباشد. در ادامه منظور از  $x_i$  ستون i ام ماتریس X است.

حال با توضیحات داده شده به سوالات زیر پاسخ دهید.

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

$$\omega_j = \frac{(x_j^T y)}{(x_j^T x_j)}$$

ب) فرض کنید ستون های ماتریس X متعامد باشد. ثابت کنید که پارامتر های بهینه از آموزش رگرسیون بر روی همه ویژگی ها با پارامتر های بهینه حاصل از آموزش روی هر ویژگی به طور مستقل یکسان است.

پ) فرض کنید میخواهیم یک رگرسیون بر روی بایاس و یکی از ویژگی های نمونه داده ها آموزش دهیم.

(ج) با توجه به اطلاعات داده شده عبارت زیر را اثبات کنید:

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

$$\omega_0 = E[y] - \omega_j E[x]$$

۴. (۱۵ نمره)

(آ) با در نظر گرفتن متغیر تصادفی نامنفی X، نامساوی زیر را اثبات کنید (نامساوی مارکوف)

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

(ب) با در نظر گرفتن نتیجه بخش الف، نشان دهید برای متغیر تصادفی دلخواه Z با امید ریاضی  $\mu$  و

واریانس  $\sigma^2$  نامساوی زیر برقرار است (نامساوی چیشف):

$$P(|Z - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

(ج) می خواهیم مقدار عدد  $\pi$  را تخمین بزنیم. برای این کار روی صفحه مختصات دو بعدی، دایره ای به شعاع واحد و مریع محیطی را در نظر بگیرید. مساحت این دایره  $\pi$  و مساحت مریع محیطی آن ۴ است.

برای تخمین مقدار  $\pi$  تعدادی نقاط تصادفی داخل این مریع تولید کرده و نسبت تعداد نقاطی که داخل

دایره قرار می گیرند را به تعداد کل به عنوان مقدار عدد  $\pi$  در نظر می گیریم.

Q2

Q3

Q 4

Q5

a)  $P(X \geq \alpha) \leq \frac{E[X]}{\alpha}$   $X \geq \alpha$

Proof:  $P(X \geq \alpha) = \int_{\alpha}^{\infty} f(x) dx \leq \int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx$

because  $x > \alpha$  for all  $x \in (\alpha, \infty)$  ( $\frac{x}{\alpha} > 1$ )

$$\Rightarrow \int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx \leq \int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx$$


---

$$\hookrightarrow \underbrace{\int_{\alpha}^{\alpha} \frac{x}{\alpha} f(x) dx}_{=0} + \underbrace{\int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx}_{\geq 0}$$

$$\Rightarrow P(X \geq \alpha) = \int_{\alpha}^{\infty} f(x) dx \leq \int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx \stackrel{x \geq 0}{\leq} \int_{\alpha}^{\infty} \frac{x}{\alpha} f(x) dx$$

$$\Rightarrow P(X \geq \alpha) \leq \frac{E[X]}{\alpha}$$

$$= \frac{1}{\alpha} \int_{\alpha}^{\infty} x f(x) dx = \frac{1}{\alpha} E[X]$$

Q5  
b)

$$P(|Z-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

$(Z-\mu)^2$  is a non-negative random variable.  
So we can apply

Markov's inequality  
with  $\alpha = \varepsilon^2$

$$\Rightarrow P((Z-\mu)^2 \geq \varepsilon^2) \leq \frac{E[(Z-\mu)^2]}{\varepsilon^2}$$

$$(Z-\mu)^2 \geq \varepsilon^2 \Rightarrow |Z-\mu| \geq \varepsilon$$

$$\overline{E[(Z-\mu)^2]} = \text{Var}(z) = \sigma^2 \quad \boxed{P(|Z-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}}$$

Q5

c) Assume  $X_i$  as a random variable with Bernoulli distribution  $P(X_i = 1) = \frac{\pi}{4} \leftarrow$  Being in the Circle

$$P(X_i = 0) = 1 - \frac{\pi}{4} \leftarrow \text{O.W.}$$

$\leftarrow \frac{\text{Area of the Circle}}{\text{Area of the Square}} = \frac{\pi}{4}$

if we generate  $n$  points we have :

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

$$E[X_i] = \frac{\pi}{4} \times 1 + \left(1 - \frac{\pi}{4}\right) \times 0 = \frac{\pi}{4}$$

$$\text{Var}[X_i] = \frac{\pi}{4} \cdot \left(1 - \frac{\pi}{4}\right) \rightarrow V(\hat{\pi}) = V\left(\frac{4}{n} \sum_{i=1}^n X_i\right) = \frac{16}{n^2} \sum V(X_i) = \frac{\pi(4-\pi)}{n}$$

the estimation error ( $\delta$ ) should be less than 0.07 with the probability at least 95% ( $\varepsilon$ ).

$$P(|\hat{\pi}(n) - \pi| < \delta) > \varepsilon$$

$$\Rightarrow P(|\hat{\pi}(n) - \pi| \geq \delta) \leq 1 - \varepsilon$$

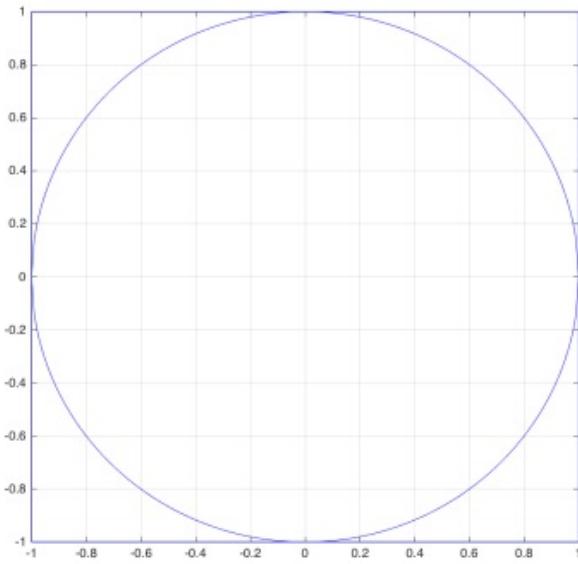
$$P(|\hat{\pi}(n) - \pi| \geq 0.01) \leq 0.05$$

From Chebyshev we have  $\frac{\text{Var}(\hat{\pi})}{\delta^2} \leq 0.05$

$$\Rightarrow \boxed{\frac{\text{Var}(\hat{\pi})}{\delta^2} = \frac{\pi(4-\pi)}{n(0.01)^2} \leq 0.05}$$

$$\frac{n \times 10^{-4}}{\pi(4-\pi)} \geq \frac{1}{5 \times 10^{-2}} \Rightarrow n \geq \frac{\pi(4-\pi)}{5} \times 10^6$$

$$n \geq 539353.24$$



با استفاده از نامساوی چیشیف تعداد عدهای تصادفی ای که باید تولید کنیم تا با قطعیت ۹۵ درصد بدانیم که خطای تخمین از ۱ درصد کمتر است را مشخص کنید.

۶. (۱۵ نمره)

- فرض کنید برای ماتریس مربعی معکوس پذیر  $A$  داریم :  $A^{-1} = V\Sigma^{-1}U^T$  ، مقادیر تکین ماتریس  $A^{-1}$  را بدست آورده و درست نمایی رابطه زیر را نمایش دهید.

$$1 \leq \sigma_{\max}(A)\sigma_{\max}(A^{-1})$$

- با توجه به تعریف نرم ماتریسی رابطه زیر را برای  $A \in R^{m*n}$  ثابت کنید:

$$\|A\|_F \leq \|A\|_2 \leq \sqrt{\text{rank}(A)}\|A\|_2$$

راهنمایی: از تجزیه SVD ماتریس  $A$  استفاده کنید.

۷. (۱۵ نمره) نشان دهید یک ترکیب خطی کلی از توابع sigmoid به فرم زیر:

$$y(x, w) = w_0 + \sum_{j=1}^n [w_j \sigma(\frac{x-\mu_j}{s})]$$

با یک ترکیب خطی از توابع tanh به فرم زیر برابر است:

$$y(x, u) = u_0 + \sum_{j=1}^n [u_j \tanh(\frac{x-\mu_j}{s})]$$

## سوالات عملی (۳۰۰ نمره)

- (۱۰۰ نمره) فایل نوتبوکی PCA\_questions در اختیار شما قرار داده شده است. راهنمایی های لازم برای نحوه انجام تمرین در فایل نوتبوک انجام شده است. در این تمرین قرار است داده های mnist را با استفاده از PCA کاهش بعد بدھیم. داده های mnist را نیز از کتابخانه keras.datasets لود میکنیم.

۲. (۱۰۰ نمره)

- فایل نوتبوک decision\_tree که در اختیارتان قرار داده شده است را کامل کنید. در این تمرین با درخت تصمیم گیری و الگوریتم آن و نحوه پیاده سازی اش بیشتر آشنا میشوید.

۳. (۱۰۰ نمره)

- در این سوال می خواهیم با استفاده از یک سری ویژگی های بیمار های قلبی مختلف، در معرض خطر بودن یا نبودن آنها را با استفاده از الگوریتم Support Vector Machines (SVM) بررسی کنیم. بیماری های قلبی

$$\frac{Q_6}{a})$$

$$A = U \Sigma V^T$$

$$A^{-1} = V \Sigma^{-1} V^T$$

$$AA^{-1} = U \Sigma \underbrace{V^T V}_{I} \Sigma^{-1} V^T$$

$$I = U \Sigma \Sigma^{-1} V^T$$

$$I \leq \frac{C}{\lambda_{\min}} \begin{pmatrix} \sigma_{\max}(A) \sigma_{\max}(A^{-1}) & 0 \\ 0 & \dots \\ \vdots & \vdots \\ 0 & \dots \\ 0 & 1 \end{pmatrix} V^T$$

$$U = \begin{bmatrix} u_1 & \dots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 \sigma_{A^{-1}} & & \\ & \ddots & \\ & & \sigma_m \sigma_{A^{-1}} \end{bmatrix}$$

$$V = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$$

$$U = \begin{bmatrix} u_1 & u_2 & \dots & u_m \end{bmatrix} \begin{bmatrix} \sigma_1 \sigma_{A^{-1}} & & & \\ & \ddots & & \\ & & \sigma_m \sigma_{A^{-1}} & \\ & & & \end{bmatrix} V^T$$

$$= u_1 u_1 \sigma_1 \sigma_{A^{-1}} + u_2 u_2 \sigma_2 \sigma_{A^{-1}} + \dots + u_m u_m \sigma_m \sigma_{A^{-1}}$$



$$Q7 \quad \sigma(n) = \frac{1}{1+e^{-n}}, \tanh(n) = \frac{e^n - e^{-n}}{e^n + e^{-n}}$$

$$\sigma\left(2\frac{n-\mu_j}{s}\right) = \frac{1}{1+e^{-\frac{2(n-\mu_j)}{s}}}$$

$$\tanh\left(\frac{n-\mu_j}{s}\right) = \frac{e^{\frac{n-\mu_j}{s}} - e^{-\frac{n-\mu_j}{s}}}{e^{\frac{n-\mu_j}{s}} + e^{-\frac{n-\mu_j}{s}}} =$$

$$\frac{e^{\frac{n-\mu_j}{s}} \left(1 - e^{-2\frac{n-\mu_j}{s}}\right)}{e^{\frac{n-\mu_j}{s}} \left(1 + e^{-2\frac{n-\mu_j}{s}}\right)} = \left(1 - e^{-2\frac{n-\mu_j}{s}}\right) \sigma\left(e^{\frac{n-\mu_j}{s}}\right)$$

$$\tanh\left(\frac{n-\mu_j}{s}\right) = \sigma\left(2\frac{n-\mu_j}{s}\right) \left(1 - e^{-2\frac{n-\mu_j}{s}}\right)$$

$$\tanh(z_j) = \sigma(2z_j) \left(1 - e^{-2z_j}\right)$$

$$y_j \tanh(z_j) = y_j \sigma(2z_j) \left(1 - e^{-2z_j}\right)$$

$$\sum_{j=1}^n u_j \tanh(z_j) = \sum_{j=1}^n u_j \sigma(2z_j) \left(1 - e^{-2z_j}\right)$$

$\xrightarrow{w_j = u_j (1 - e^{-2z_j})}$

$$\sum_{j=1}^n u_j \tanh(z_j) = \sum_{j=1}^n w_j \sigma(2z_j)$$

$$\Rightarrow \sum_{j=1}^n [u_j \tanh\left(\frac{x - u_j}{s}\right)] = \sum_{j=1}^n [w_j \sigma\left(2 \frac{x - u_j}{s}\right)]$$

در جهان سالانه منجر به حدود ۱۸ میلیون مرگ می شوند.  
ابتدا برای آشنایی بیشتر با دیتاست، فایل Dataset\_Description.pdf را مطالعه بفرمایید.

(آ) ابتدا باید اطلاعات کلی دیتا را ارائه کنید. در این قسمت موارد زیر را بررسی کنید:  
i. اندازه دیتا

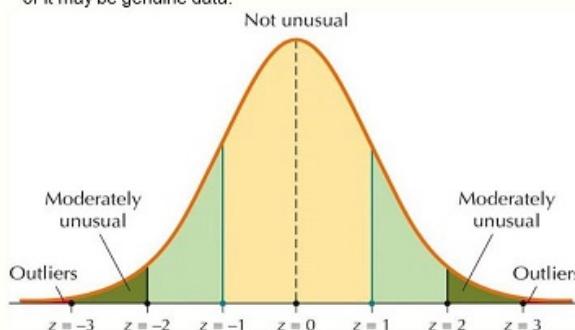
ii. بررسی اینکه دیتا در هر نمونه آیا ویژگی حذف شده دارد یا نه.

iii. بررسی بالانس بودن دو کلاس

iv. رسم نمودار توزیع سن و توزیع جنسیت برای هر کلاس (مجموعاً چهار نمودار)

(ب) در این بخش با استفاده از Z-test داده های پرت را حذف کنید.

An outlier is an extremely large or extremely small data value relative to the rest of the data set. It may represent a data entry error, or it may be genuine data.



برای این بخش، ترشهولد Z-test را روی ۳ در نظر بگیرید و گزارش کنید کدام داده ها، داده پرت (outlier) هستند. همچنین سایز نهایی بعد از حذف این داده ها را هم گزارش کنید.

(ج) حال برای این قسمت، باید داده هایی که numerical هستن را نرمال کنید(بین صفر و یک قرار دهید)

(د) در این قسمت، ابتدا ۷۰ درصد داده ها را برای آموزش و مابقی را برای تست جدا کنید. حال می خواهیم با سه kernel مختلف با استفاده از الگوریتم SVM آموزش را انجام بدیم:

i. کرنل linear

ii. کرنل RBF

iii. کرنل Polynominal

برای هر کرنل، پارامترها رو به گونه ایی تغییر دهید تا بالاترین درصد را بگیرید. برای کرنل RBF پارامتر gamma را به گونه ایی تغییر دهید تا به حداقل دقت ۸۵ درصد برسید. همچنین برای کرنل ploynomial میتوانید پارامتر degree را تغییر دهید.  
برای هر کرنل، نتایج F1score ، Recall ، Precision ، Accuracy را گزارش کنید.