

10
מוס' שאלון - 572
ביוולי 2025

"ד בتمודז תשפ"ה

סמסטר 2025
22961 / 4

86 מוס' מועד

שאלון בחינת גמר

22961 - למידה عمוקה

משך בחינה: 3 שעות

בשאלון זה 10 עמודים

מבנה הבחינה:

בחינה 20 שאלות רב ברירה.

יש לענות על כלן.

לכל שאלה יש רק תשובה אחת נכונה.

הקפידו לנחל את הזמן נכון, כרך שתסייעו לענות על כל השאלות.

את התשובות עלייכם לסמן על גבי גיליון התשובות לשאלות הרוב ברירה

שבגביה מחברת הבחינה.

בסוף הבחינה מצורף ריכוז של נוסחאות והגדרות לשימושכם.

בהצלחה !!!

חומר עזר:

כל חומר עזר אסור לשימוש.

**איןכם חייבים
להחזיר את השאלון לאוניברסיטה הפתוחה**

שאלה 1

איזה מההתשובות הבאות נכונה ביחס לעובר פועלות אלגוריתם מורד הגרדיאנט (gradient descent) לתחילה המינימיזציה של פונקציית המחיר (cost function).

- א. בכל נקודה במרחב הפרמטרים, תנוצה קטנה בכיוון השלילי של הגרדיאנט תוביל לירידה בערך פונקציית המחיר.
- ב. אם גודל הצעד (learning rate) קטן מדי, האלגוריתם עשוי להתקדם באיטיות רבה ואף לא להתכנס גם לאחר מספר רב של איטרציות.
- ג. גודל צעד גדול מדי יוביל לאי-יציבות בתהליכי ויתכן שנDELג מעל נקודת המינימום, וכך ערך פונקציית המחיר לא ייריד.
- ד. כל התשובות נכונות.

שאלה 2

נתונות הטענות הבאות עברו פונקציית softmax:

- 1. מניבה מספרים חיוביים שסכומם אחת.
- 2. אינואրיאנטית לתוספת קבוע לכל אחד מהאיברים, כלומר $\text{softmax}(x_1 + c, x_2 + c, \dots, x_n + c) = \text{softmax}(x_1, x_2, \dots, x_n)$.
- 3. פונקציה גזירה.
- 4. תמיד מחזירה ערכים בינאריים (0 או 1).
- 5. אינה מושפעת מסדר האיברים בוקטור הקלט.

בחרו בתשובה הנכונה ביותר:

- א. אף טענה אינה נכונה.
- ב. רק טענה אחת נכונה.
- ג. רק שתי טענות נכונות.
- ד. רק שלוש טענות נכונות.
- ה. רק ארבע טענות נכונות.
- ו. כל חמש הטענות נכונות.

שאלה 3

השימוש בשיטת הגרדיאנט האקראי (Stochastic Gradient Descent - SGD) נועד להתמודד עם הבעיה הבאה (בחרו בתשובה הנכונה ביותר):

- א. היעדר גרדיאנט בנקודות לא גזירות של פונקציית העלות.
- ב. חישוב איטי של הגרדיאנט באימון רשותן מרכיבות על אוסף נתונים (dataset) גדול.
- ג. חוסר יכולת למצוא מינימום גלובלי של פונקציית העלות.
- ד. תשובות א', ב', ו-ג' נכונות.
- ה. תשובות א', ב', ו-ג' אינן נכונות.

שאלה 4

נתונה רשת נוירונים מסווג Fully connected FeedForward הכוללת שתי שכבות, כאשר פלט הרשות מוגדר ע"י:

$$y = h(U^T g(W^T x + b) + c)$$

עם פונקציות אקטיבציה $g()$ ו- $h()$ בשכבה החוביה ו- $x \in \mathbb{R}^d$, הפרמטרים הנלמדים הם מטריצות המשקלות $W \in \mathbb{R}^{d \times k}$, $U \in \mathbb{R}^{k \times d}$, וקטור הפלט הוא $y \in \mathbb{R}^d$, וקטור הטיות $b \in \mathbb{R}^k$ ו- $c \in \mathbb{R}^d$. נתון גם ש-

בחרו בתשובה הנכונה ביותר, בהינתן **שפונקציות האקטיבציה הן פונקציות זהות (identity)** **כלומר:**

$$g(z) = z, h(z) = z$$

א. רשת זו מסוגלת לייצר כל טרנספורמציה אפינית (affine) מהצורה $y = P^T x + d$ בין הפלט

$$y \in \mathbb{R}^d$$
 לבין הפלט $x \in \mathbb{R}^d$

ב. רשת זו מאפשרת ייצוג של הפלט $y \in \mathbb{R}^d$ באמצעות מרחב נמוך מימד.

ג. הודות לשימוש ביותר משכבה אחת, רשת זו עשויה לייצג קשרים מורכבים יותר מאשר אפיניים (affine) בין הפלט לפלט.

ד. תשובה א', ב', ו-ג' אינן נכונות.

ה. תשובה א', ב', ו-ג' נכונות.

שאלה 5

מה ההבדל העיקרי בין (SGD) Stochastic Gradient Descent עם מומנטום לבין Adam ?

א. SGD עם מומנטום משתמש רק במידע על כיוון הגרדיינט, בעוד Adam מתאים את כיוון וגודל העדכון לפי גראדינטים קודמים.

ב. Adam משלב מומנטום עם התאמת של קצב הלמידה לכל פרמטר בנפרד.

ג. בעוד SGD עם מומנטום משתמש בגרדיינט ממוצע לאורץ זמן, Adam מחשב עדכון מבוסס על סטיית תקן של הגרדיינטים בלבד.

ד. Adam משתמש רק בגרדיינט הנוכחי, ללא מידע היסטורי כמו מומנטום.

שאלה 6

מהם היתרונות המרכזים של רשתות עצביות כונבולציוניות (Convolutional Neural Networks) לעומת רשתות עצביות כונבולציוניות (Fully connected FeedForward networks)?

- א. רשתות כונבולוציה דורשות יותר פרמטרים ולכון מסוגלות ללמידה תבניות מורכבות יותר.
- ב. רשתות כונבולוציה משתמשות בשכבות כונבולוציה המאפשרות שיתוף משקלים ויזיהו תבניות מקומיות, דבר המפחית את מספר הפרמטרים ומשפר את יכולת ההכללה.
- ג. רשתות כונבולוציה אינן תלויות בסדר הפיקסלים ולכון אין מושפעות ממיקום האובייקטים בתמונה.
- ד. רשתות כונבולוציה מועלמות מבנה התמונה ולכון אין מנצלות מידע מרחבי לצורך זיהוי תכונות.
- ה. תשובה א', ב', ג' ו-ד' נכונות.
- ו. תשובה א', ב', ג' ו-ד' אינן נכונות.

שאלה 7

במסגרת אימון רשת נוירונית ליזיהו עצמים בתמונות, המודל מפגין דיקוגרף גבו על קבוצת האימון אך ביצועים ירודים על קבוצת הווילידציה. איזו מהאפשרויות הבאות פחות מתאימה להתמודדות עם מצב זה?

- א. עצירה אוטומטית של תהליך האימון כאשר אין שיפור מתחשך במדדי הביצוע על סט הווילידציה.
- ב. השמטה אקראית של יחידות פנימיות בשכבות הנסתורות במהלך האימון.
- ג. שילוב של פלטים ממספר עותקים שונים של הרשת שאומנו בנפרד.
- ד. הגדלת מספר שכבות הרשת והארכת זמן האימון עד למצוי מוחלט של הדאטה.
- ה. יצירת גרסאות נוספות של הדאטה באמצעות סיבובים, חיתוכים ושינויים חזותיים.

שאלה 8

רשת כונבולוציה מקבלת קלט טנזור קלט ממימד 128×256 העובר דרך שכבה כונבולוציה בעלת 64 פילטרים עם גודל גרעין 3×3 , ולא ריפוד באפסים. מה יהיה גודל טנзор הפלט?

בחרו את התשובה הנכונה:

- א. $64 \times 40 \times 50$.
- ב. $64 \times 40 \times 51$.
- ג. $64 \times 39 \times 50$.
- ד. $64 \times 41 \times 52$.
- ה. תשובה א', ב', ג' ו-ד' אינן נכונות.

שאלה 9

צוות חוקרים מבקש לאמן מודל לזיהוי סוגים נדירים של סרטן באמצעות תמונות מיקרוסקופיות. ברשותם רק 300 תמונות מתוויות, ומספר הקטגוריות לשיווג גבוהה יחסית.

באייזו מהגישות הבאות סביר ביותר להשתמש כדי לשפר את ביצועי המודל בתנאים אלו?

- א. לאמן רשת عمוקה מאוד מאפס (from scratch) על סט התמונות הקטן.
- ב. להשתמש ב- Transfer Learning עם מודל שאומן מראש על תמונות כלליות (כגון ImageNet) ולבצע fine-tuning.
- ג. לשכפל את התמונות המקוריות כדי להרחיב את סט האימון.
- ד. לאמן רשת عمוקה מאוד מאפס (from scratch) , ולהחיל Dropout על כל שכבות המודל כדי לצמצם תופעת אימון היתר (overfitting).

שאלה 10

אייזו מהדריכים הבאות אינה נחשבת לשיטה מקובלת ליישום Transfer Learning במודל רשת عمוקה ?

- א. הקפאת (freeze) השכבות המוקדמות במודל שאומן מראש ואימון רק של שכבות האחרוניות על הדאטה החדש.
- ב. שכפול ארכיטקטורת המודל המאומן מראש, אך אתחול אקראי של כל המשקלות ואימון מחדש על הדאטה החדש כדי למנוע הטיות ממשימה קודמת.
- ג. טעינת משקלות ממודל שאומן מראש וביצוע fine-tuning על חלק מהשכבות לפי הדאטה החדש.
- ד. החלפת השכבות האחרוניות במודל המאומן בשכבות חדשות המותאמות למשימה חדשה.

שאלה 11

בעת תכנון רשת נוירוניים, קיימת התאמה טبيعית בין סוג הבעייה, פונקציית האקטיבציה בשכבה הפלט, ופונקציית המחיר (cost function). אייזו מההתאמות הבאות אינה אידיאלית לסוג הבעייה?

- א. בעיית גרסיה , פונקציית פלט : זהות (identity), פונקציית המחיר : סכום ריבועים.
- ב. סיוג בינארי, פונקציית פלט : סיגמוואיד, פונקציית המחיר : Cross Entropy .
- ג. סיוג רב-מחלקות (multi-class), פונקציית פלט : softmax, פונקציית המחיר : רב-מחלקטית.
- ד. סיוג בינארי, פונקציית פלט: tanh, פונקציית המחיר : סכום ריבועים.

שאלה 12

נתונה פונקציה ריבועית קמורה מהצורה:

$$f(x) = \frac{1}{2}x^T Ax$$

כאשר A היא מטריצה סימטרית וחיבורית מוגדרת. נניח כי קיים פער משמעותי בין הערך העצמי הקטן ביותר לערך העצמי הגדול ביותר של A.

אופטימיזציה של הפונקציה מתבצעת באמצעות שיטת מורד הגראדיינט (Gradient Descent) עם וקטור התחלת שרירותי ועם גודל צעד קבוע ותקין (שאינו גדול מדי כך שהשיטה מתכנסת תאורטית). הסבירו מהו הגורם העיקרי שמשפיע על קצב ההתקנסות של השיטה, ובחירה בתשובה הנכונה ביותר מתוך האפשרויות הבאות:

- א. ההתקנסות מהירה בגלל שהגרדיינטים גדולים בכיוון של הערך העצמי הגדול של A.
- ב. ההתקנסות איטית, בעיקר בגלל שהעדכניםים בכיוון של הערך העצמי הגדול של A גדולים יותר.
- ג. ההתקנסות איטית, בעיקר בגלל שהעדכניםים בכיוון של הערך העצמי הקטן של A קטנים יותר.
- ד. שיטת מורד הגראדיינט אינה רגישה כלל לערבים העצמיים של A.
- ה. תשובה א', ב', ג' ו-ה אינן נכונות.

שאלה 13

מהי מטרת השימוש ב- **Dataset Augmentation** באימון רשות נוירונים?

בחרו איזו מהtheses הבאות אינה נכונה.

- א. להרחיב את גודל קבוצת האימון על ידי יצרת דוגמאות נוספות מהדעת הקיימים.
- ב. לשפר את יכולת ההכללה של המודל ולצמצם סכנת Overfitting.
- ג. לקצר את זמן האימון הכלול של המודל, לאחר והמודל רוכש את התפלגות הרצiosa של הנתונים תוך שימוש בפחות דוגמאות אמיתיות.
- ד. לאלאץ את המודל ללמידה אינואריאנטיות מסוימות שההשפעה אמורה להיות אדירה להן.

שאלה 14

מה ההשפעה הצפואה של dropout על זמן האימון הנדרש להתקנסות? **בחרו בתשובה הנכונה ביותר.**

- א. מקוצר את זמן האימון מאוחר ופחות נוירונים פעילים.
- ב. מאריך את זמן האימון, כי בכל איטרציה מאmins תתר-שרות שונה.
- ג. אין כל השפעה על זמן האימון.
- ד. Dropout מפסיק את האימון ברגע שנוצר overfitting, ולכן מקוצר את זמן האימון.
- ה. תשובה א', ב', ג' ו-ד' אינן נכונות.

שאלות 15,16:

נתונה פונקציית מחיר (cost function) ריבועית מהצורה:

$$f(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T Q \mathbf{w} + \mathbf{b}^T \mathbf{w}$$

כאשר $Q \in \mathbb{R}^{2 \times 2}$ היא מטריצה סימטרית עם ערכים עצמיים $\lambda_1 = 10, \lambda_2 = 2.5$ ווקטורים עצמיים $\mathbf{b}^T = [-1, 5.5], \mathbf{v}_1^T = [1, 2]/\sqrt{5}, \mathbf{v}_2^T = [-2, 1]/\sqrt{5}$ מתאימים.

שאלה 15

חשבו את הנקודה \mathbf{w}^* המביאה למינימום את פונקציית המחיר.

- א. $\mathbf{w}^{*T} = [1, -1]$
- ב. $\mathbf{w}^{*T} = [-1, 1]$
- ג. $\mathbf{w}^{*T} = [1, 0]$
- ד. $\mathbf{w}^{*T} = [0, -1]$
- ה. תשובה א', ב', ג' ו-ד' אינן נכונות.

שאלה 16

מוסיפים לפונקציית המחיר איבר רגולרייזציה מסוג Ridge :

$$f(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T Q \mathbf{w} + \mathbf{b}^T \mathbf{w} + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

עם פרמטר $\lambda = 10$. נקודת המינימום \mathbf{w}_{reg} של פונקציית המחיר החדשה עם איבר הרגולרייזציה נתונה ע"י:

- א. $\mathbf{w}_{reg}^T \approx [-0.14, 0.32]$
- ב. $\mathbf{w}_{reg}^T \approx [-0.14, 0.14]$
- ג. $\mathbf{w}_{reg}^T \approx [0.14, -0.32]$
- ד. $\mathbf{w}_{reg}^T \approx [0.32, -0.14]$
- ה. $\mathbf{w}_{reg}^T \approx [-0.32, 0.14]$
- ו. $\mathbf{w}_{reg}^T \approx [0.14, -0.14]$

שאלה 17

מהו היתרונו העיקרי של פונקציית האקטיבציה ReLU בהשוואה ל- sigmoid ו- tanh בקשרות נוירוניים? בחרו בתשובה הנכונה ביותר.

- א. פונקציית ReLU מבטיחה שפלט האקטיבציה יהיה בין 0 ל- 1, ולכן מתאימה טוב יותר לביעות סיוג.
- ב. פונקציית ReLU אינה גורמת לעולם לבעה של גרדיאנט נעלם (vanishing gradients).
- ג. פונקציית ReLU חוסכת חישובים ונוטה להוביל להתקנסות מהירה יותר באימון רשתות נוירוניים.
- ד. פונקציית ReLU משמרת מידע מרחבי טוב יותר.

שאלה 18

במהלך אימון רשת CNN הכוללת מספר שכבות convolution ו- MaxPooling, אתם מבחינים בכך שמידע מרחבי חשוב הולך לאיבוד כבר בשכבות הראשונות. מה הסיבה הסבירה ביותר לכך?

- א. ערך stride בשכבות הקונולוציה גדול מדי.
- ב. פונקציית האקטיבציה ReLU גורמת לאובדן מידע מרחבי ולכן אינה מתאימה לבניה כזו.
- ג. החסר בשכבה Fully Connected גורם לכך שהרשות לא מצליחה לשמור מידע בשכבות הראשונות.
- ד. אין מספיק training data, ולכן הרשות לא לומדת לשמר את המידע המרחבי החשוב.

שאלה 19

בעת חישוב הגרדיאנט של פונקציית המחיר (cost function) ביחס לפרמטרים של שכבה מסוימת ברשת נוירוניים, אילו רכיבים משפיעים עליו? בחרו בתשובה הנכונה ביותר:

- א. רק הקלט שהזון לשכבה זו.
- ב. רק השגיאה שחושבה ביציאת הרשות.
- ג. גם הקלט לשכבה וגם הגרדיאנט שהועבר מהשכבה הבאה.
- ד. רק גודל הרשות.
- ה. תשובה א', ב', ג', ו-ד' אינן נכונות.

שאלה 20

נתונה רשת נוירונים לSieving ביןاري עם שכבה אחת, שבה הפלט מחושב לפי :

$$\hat{y} = \sigma(\mathbf{w}^T \mathbf{x} + b)$$

כאשר σ היא פונקציית sigmoid, ופונקציית המחיר J היא cross-entropy, כלומר עבור דוגמה בודדת (\mathbf{x}, y) :

$$J(\mathbf{x}, y) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$$

מהו הגרדיינט של פונקציית המחיר J ביחס לוקטור הפרמטרים \mathbf{w} ?

א. $(2y - 1)\sigma((1 - 2y)(\mathbf{w}^T \mathbf{x} + b))\mathbf{x}$

ב. $(1 - 2y)\sigma((1 - 2y)(\mathbf{w}^T \mathbf{x} + b))\mathbf{x}$

ג. $(2y - 1)\left(1 - \sigma((1 - 2y)(\mathbf{w}^T \mathbf{x} + b))\right)\mathbf{x}$

ד. $(2y - 1)\sigma((2y - 1)(\mathbf{w}^T \mathbf{x} + b))\mathbf{x}$

ה. תשובה א', ב', ג' ו-ד' אינן נכונות.

ב ה צ ל ח ה !

רכיב הגדרות ונוסחאות:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Derivatives:

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T \mathbf{x}) = \mathbf{x}$$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T A \mathbf{w}) = (A + A^T)\mathbf{w}$$

$$\frac{\partial^2}{\partial \mathbf{w} \partial \mathbf{w}^T}(\mathbf{w}^T A \mathbf{w}) = A + A^T$$

The **directional derivative** in the direction of \mathbf{u} where \mathbf{u} is a unit vector:

$$\left. \frac{\partial}{\partial \alpha} f(\mathbf{x} + \alpha \mathbf{u}) \right|_{\alpha=0} = \mathbf{u}^T \nabla_{\mathbf{x}} f(\mathbf{x})$$

The **second directional derivative** in the direction of \mathbf{u} where \mathbf{u} is a unit vector:

$$\left. \frac{\partial^2}{\partial \alpha^2} f(\mathbf{x} + \alpha \mathbf{u}) \right|_{\alpha=0} = \mathbf{u}^T \mathbf{H}(\mathbf{x}) \mathbf{u}$$

where $\mathbf{H}(\mathbf{x})$ is the Hessian matrix: $H(\mathbf{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$.

A second-order Taylor approximation around the point $\mathbf{x}^{(0)}$:

$$f(\mathbf{x}) \approx f(\mathbf{x}^{(0)}) + (\mathbf{x} - \mathbf{x}^{(0)})^T \nabla_{\mathbf{x}} f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^{(0)}} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(0)})^T \mathbf{H}(\mathbf{x}^{(0)}) (\mathbf{x} - \mathbf{x}^{(0)})$$