הטכניון-מכון טכנולוגי לישראל **אביב תשפ"א**
הפקולטה להנדסת חשמל ומחשבים **26.7.2021**

למידה עמוקה - 046211
בחינה סופית - מועד א'

משך המבחן: *3* שעות

1. ניתן להשתמש בדפי נוסחאות ומחשבון.
2. יש לענות על כל השאלות.
3. משקל כל שאלה מצוין בטופס הבחינה. סך הנקודות הוא 100.
4. נא לכתוב בצורה ברורה ומסודרת.
5. יש לכלול פירוט מלא של דרך הפתרון. תשובות לא מנומקות לא יזכו בניקוד, אם לא נאמר אחרת.

ב ה צ ל ח ה !!!

שאלה מס' 1: רגולריזציית Dropout ובחירת מאפיינים (35%)

בזמן שאמיר חגג ביום הסטודנט עלה לו רעיון להשתמש ב-dropout כדי לבחור מאפיינים בכניסה: כלומר ליצור מסיכת dropout בה יש הסתברות p_i לאיפוס הרכיב ה- i בכניסה, ולעשות אופטימיזציה על p_i כך שנקבל שיעודד בחירה דטרמינסטית של מאפיינים (כלומר $p_i \to 0 \text{ or } 1$). אמיר החליט תחילה לבחון את השיטה ברגרסיה לינארית:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)} - \mathbf{w}^T\mathbf{D}^{(n)}\mathbf{x}^{(n)}\right)^2$$

כאשר $\mathbf{w}\in\mathbb{R}^d$ זהו וקטור הפרמטרים, $\mathbf{x}^{(n)}\in\mathbb{R}^d$ הן הדוגמאות בסט האימון, $y^{(n)}\in\mathbb{R}$ הן התגיות של הדוגמאות, ו- $\mathbf{D}^{(n)}\in\{0,1\}^{d\times d}$ היא מסיכת ה-dropout אקראית אלכסונית, כאשר כל איבר במטריצה מוגרל בצורה בלתי תלויה לפי

$$D_{ii}^{(n)} = \frac{1}{1-p_i}\cdot\begin{cases}1 \text{ w.p. } 1-p_i\\ 0 \text{ w.p. } p_i\end{cases}$$

כאשר $p_i\in[0,1]$ היא ההסתברות לאיפוס הרכיב ה- i בכניסה, ונסמן $\mathbf{p}=[p_1,\dots,p_d]^T$.

א. מצאו את $ED_{ii}^{(n)}$ והראו כי $E\left[D_{ii}^{(n)}D_{jj}^{(n)}\right]=1+\delta_{ij}\frac{p_i}{1-p_i}$.

ב. הראו כי פונקציית המחיר הממוצעת $\bar{\mathcal{L}}(\mathbf{w},\mathbf{p})$ (כאשר הממוצע הוא על המסכות) היא

$$\bar{\mathcal{L}}(\mathbf{w},\mathbf{p}) \triangleq E\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)} - \mathbf{w}^T\mathbf{x}^{(n)}\right)^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{p_i}{1-p_i}c_i w_i^2$$

כאשר $c_i \triangleq \sum_{n=1}^{N}\left(x_i^{(n)}\right)^2$. מכאן ואילך ניתן תמיד להניח כי $\forall i: c_i>0$.

ג. הסבירו איכותית מה ההבדל בין $\bar{\mathcal{L}}(\mathbf{w},\mathbf{p})$ ופונקציית המחיר ברגרסיה לינארית ללא dropout.

ד. נשים לב ש- $\bar{\mathcal{L}}(\mathbf{w},\mathbf{p})$ תלויה ב- $\mathbf{p}$, אבל כזכור $p_i\in[0,1]$. הציעו פונקציה $\mathbf{p}=f(\mathbf{u})$ כך שנוכל להשתמש ב-GD ללא אילוצים על $\bar{\mathcal{L}}(\mathbf{w},f(\mathbf{u}))$.

ה. כזכור אנו רוצים שעבור כמה מאפיינים $p_i\to 1$ והשאר $p_i\to 0$. ניתן להניח שקיים פתרון $\mathbf{w}_0$ דליל (המכיל אפסים), ופתרון לא דליל $\mathbf{w}_*$ כך ש- $\forall n: y^{(n)} = \mathbf{w}_0^T\mathbf{x}^{(n)} = \mathbf{w}_*^T\mathbf{x}^{(n)}$. האם בהכרח נקבל את $\mathbf{w}_0$ ע"י הגעה למינימום של $\bar{\mathcal{L}}(\mathbf{w},\mathbf{p})$ ב- $\mathbf{w},\mathbf{p}$?

ו. לאחר ניסוי וטעייה, אמיר השיג שיפור כשהוסיף רעש לכניסה ורגולרזציה על $\mathbf{p}$, וקיבל

$$\tilde{\mathcal{L}}(\mathbf{w},\mathbf{p}) \triangleq E\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)} - \mathbf{w}^T\mathbf{x}^{(n)}\right)^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{1-p_i}w_i^2\sum_{n=1}^{N}\left(x_i^{(n)}\right)^2 + \mu\sum_{i=1}^{d}(1-p_i)$$

הראו, ע"י חישוב $\tilde{\mathcal{L}}(\mathbf{w}) \triangleq \min_{\mathbf{p}\in\mathbb{R}^d}\tilde{\mathcal{L}}(\mathbf{w},\mathbf{p})$, כי ניתן היה לוותר על dropout, ולהוסיף במקום זה רגולריזציה $R(\mathbf{w})$ ישירות לרגרסיה לינארית. חשבו את הרגולריזציה $R(\mathbf{w})$ והסבירו כיצד היא עוזרת בבחירת מאפיינים.

ז. קיימות הרחבות שונות ל-dropout הנמצאות בשימוש כיום ברשתות קונוולוציה. בחרו הרחבה אחת כזו, תארו כיצד היא עובדת, ומה המוטיבציה מאחוריה. (סעיף זה לא קשור לסעיפים קודמים).

שאלה מס׳ 2: רגולריזציית L2 (25%)

נבחן את ההשפעה של רגולריזציית L2 בנוירון בודד עם אקטיבציה לינארית, ופונקציית מחיר ריבועית – כלומר, רגרסיה לינארית עם שגיאת אימון:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2}\mu\|\mathbf{w}\|^2$$

כאשר $\mathbf{w} \in \mathbb{R}^d$ זהו וקטור הפרמטרים, השורות של $\mathbf{X} = \left[\mathbf{x}^{(1)},...,\mathbf{x}^{(N)}\right] \in \mathbb{R}^{N\times d}$ הן הדוגמאות בסט האימון, $\mathbf{y} \in \left[y^{(1)},...,y^{(N)}\right] \in \mathbb{R}^N$ Wזהו וקטור התגיות של הדוגמאות, ו- $\mu > 0$ זה מקדם הרגולריזציה. נרצה למזער את השגיאה $\mathcal{L}(\mathbf{w})$ באמצעות GD מאתחול $\mathbf{w}(0)$ עם גודל צעד η.

נסמן ב- $\mathbf{\Lambda}$ את המטריצה האלכסונית של הערכים העצמיים $\lambda_1 \le \lambda_2 \le ... \le \lambda_d$ של המטריצה $\mathbf{H}_0 = \mathbf{X}^T\mathbf{X}$ ב- $\mathbf{U}$ את המטריצה של את הוקטורים העצמיים $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_d$, כך ש- $\mathbf{H}_0 = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$.

כזכור, סימנו בכיתה $\mathbf{H}_\mu = \mathbf{X}^T\mathbf{X} + \mu\mathbf{I}$, $\mathbf{w}_\mu = \mathbf{H}_\mu^{-1}\mathbf{X}^T\mathbf{y}$, ו- $\mathbf{z}(0) = \mathbf{U}^T\left(\mathbf{w}(0) - \mathbf{w}_\mu\right)$. והראנו כי

כי השגיאה המתקבלת לאחר t צעדים של GD

$$(*) \quad \mathcal{L}(\mathbf{w}(t)) = \frac{1}{2}\mathbf{z}(0)^T\left(\mathbf{I} - \eta\left(\mathbf{\Lambda} + \mu\mathbf{I}\right)\right)^{2t}\left(\mathbf{\Lambda} + \mu\mathbf{I}\right)\mathbf{z}(0) + C$$

א. קצב ההתכנסות הוגדר בתור $\text{rate}(\eta) = \max_i \left|1 - \eta\left(\lambda_i + \mu\right)\right|$. השתמשו ב- $(*)$ להסביר מדוע.

ב. הראו כי גודל הצעד האופטימלי $\eta^* = \frac{2}{\lambda_d + \lambda_1 + 2\mu}$, וזה גורר $\text{rate}\left(\eta^*\right) = \frac{\lambda_d - \lambda_1}{\lambda_d + \lambda_1 + 2\mu}$.

<u>הנחיה</u>: (1) אילו ערכים עצמיים משפיעים על $\text{rate}(\eta) = \max_i \left|1 - \eta\left(\lambda_i + \mu\right)\right|$? (2) ציירו את $\text{rate}(\eta)$.

ג. איזה ערך μ נבחר כדי להאיץ את קצב ההתכנסות? מה תהיה השגיאה הסופית $\mathcal{L}(\mathbf{w}(\infty))$?

ד. עבור $d = N$, עם $\mathbf{H}_0$ לא סינגולרית, מצאו את μ הממזער את השגיאה הסופית $\mathcal{L}(\mathbf{w}(\infty))$.

שאלה מס' 3: רשתות רב שכבתיות וסימטריות (40%)

שלי התקבלה לעבודה בחברת "Aperture", למחלקת בדיקות גנטיות. ביקשו ממנה לתכנן רשת עצבית המקבלת כקלט רצף של קוד דנ"א באורך d. הדנ"א הוא רצף של בסיסים, כאשר כל בסיס הוא אחד מארבע אפשרויות (C,T,G,A), ולכן הקלט מתואר ע"י המטריצה $\mathbf{X}\in\mathbb{R}^{4\times d}$, כאשר $X[j,i]$ מייצג את הערך שנמדד לריכוז הבסיס ה-j במיקום ה-i. הרשת אמורה להוציא סיווג בינארי $y\in\{-1,1\}$ עבור תכונה מסוימת אותה רוצים למצוא.

שלי החליטה לבדוק תחילה רשת עם שכבה נסתרת אחת בגודל $4\times d$ ואקטיבציות LReLU

$$f_{\mathbf{w}}(\mathbf{X})=\sum_{r=1}^{4}\sum_{k=1}^{d}W_2[r,k]\varphi\left(\sum_{j=1}^{4}\sum_{i=1}^{d}W_1[r,k,j,i]X[j,i]\right)$$

כאשר $\mathbf{w}=\{\mathbf{W}_1,\mathbf{W}_2\}$ הן השכבות של <u>טנזורי</u> משקולות.

א. ציירו איור סכמטי של החיבורים ברשת עבור $d=2$.

ב. איזה סימטריות (Invariances) קיימות בפרמטרים ברשת?

ג. כעת שלי שמה לב כי הכיוון שבו הנסרק הדנ"א הוא שרירותי, ולכן אם עבור שני קלטים $\mathbf{X},\tilde{\mathbf{X}}$ מתקיים $\forall i,j: X[j,i]=\tilde{X}[j,d+1-i]$, אזי שני הקלטים בהכרח זהים במשמעות. שלי נזכרה שרצוי שהרשת תהיה זהה במוצא שלה עבור שני קלטים זהים במשמעות (invariant), ותהיה equivariant בשכבה הנסתרת. אילו אילוצים נעשה על פרמטרי הרשת כדי לאלץ תכונות אלו?

הרשת מאומנת על הדוגמאות $\left\{\mathbf{X}^{(n)},\mathbf{y}^{(n)}\right\}_{n=1}^{N}$, עם פונקציית מחיר לוגיסטית $\ell(y,\hat{y})=\log\left(1+e^{-y\hat{y}}\right)$.
לאחר האימון הסיווג הבינארי יתבצע ע"י $\operatorname{sign}\left(f(\mathbf{X})\right)$.

ד. לפתע שלי שמה לב שעקב תהליך המדידה הרועש, כל דוגמא $\mathbf{X}^{(n)}$ מוכפלת בסקאלה שרירותית, ולכן אם עבור שני קלטים $\mathbf{X},\tilde{\mathbf{X}}$ מתקיים $\forall i,j: X[j,i]=c\tilde{X}[j.i]$, עבור איזשהו קבוע $c>0$, אזי שני הקלטים בהכרח זהים במשמעות.

(1) עבור רשת נתונה (שכבר אומנה) מה ההשפעה של הסקאלה c על תוצאת הסיווג?

(2) כיצד הסקאלה השרירותית יכולה לפגוע בתהליך האימון?

<u>רמז</u>: חישבו מה קורה לגרדיאנטים לכל דוגמא.

(3) כיצד שלי תוכל להיעזר במידע זה כדי לשפר את המסווג?