תרגיל בית מספר 3 – למידת מכונה – רועי מסילתי - 315253336

1. Regularized polynomial regression

We derived in class the solution for a zero-degree polynomial regression. Consider the problem of regularized polynomial regression.

$$Err(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(h_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 + \lambda||\mathbf{w}||^2 \quad .$$

1. Derive the solution for a polynomial of degree 0: $h_{\mathbf{w}}(\mathbf{x}) = w_0$. Analyze the solution in the limit of $\lambda \to \infty$ and $\lambda \to 0$.

אנחנו רוצים למצוא את $h_w(x)$ שיביא למינימום את $Err(w)$.
עבור פולינום מדרגה 0 מתקיים $h_w(x) = w_0$ ולכן $\|w\|^2 = w_0^2$, כלומר עלינו להביא למינימום את:

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}(w_0 - y_i)^2 + \lambda w_0^2$$

נגזור את הפונקציה לפי w_0 ונביא לנקודת מינימום. כפי שראינו בהרצאה, כאשר גוזרים ומשווים ל-0, אכן מקבלים נקודת מינימום ותוספת של λw_0^2 לא משנה את המינימליות של נקודה זו:

$$\frac{\partial Err(w)}{\partial w} = \left(\frac{1}{n}\sum_{i=1}^{n}2(w_0 - y_i) * 1\right) + \lambda * 2w_0 = \left(\frac{2}{n}\sum_{i=1}^{n}w_0 - y_i\right) + 2\lambda w_0$$
$$= \frac{2}{n} * n * w_0 - \frac{2}{n}\sum_{i=1}^{n}y_i + 2\lambda w_0 = 2w_0 - \left(\frac{2}{n}\sum_{i=1}^{n}y_i\right) + 2\lambda w_0$$

נשווה ל-0:

$$2w_0 - \left(\frac{2}{n}\sum_{i=1}^{n}y_i\right) + 2\lambda w_0 = 0 \Leftrightarrow 2w_0 + 2\lambda w_0 = \frac{2}{n}\sum_{i=1}^{n}y_i \Leftrightarrow w_0(1+\lambda) = \overset{y\ average}{\bar{y}}$$
$$\Leftrightarrow w_0 = \frac{\bar{y}}{1+\lambda}$$

כעת נראה שכאשר $\lambda \to \infty$ נקבל:

$$w_0 = \underset{\lambda\to\infty}{\frac{\bar{y}}{1+\lambda}} \Longrightarrow w_0 = 0$$

וכאשר $\lambda \to 0$ נקבל:

$$w_0 = \underset{\lambda\to 0}{\frac{\bar{y}}{1+\lambda}} = \frac{\bar{y}}{1} = \bar{y}$$

מסקנה – כאשר λ גדול מאוד, הרגיולריזציה גדולה ותגרום ל- w_0 להיות שווה ל-0.
וכאשר λ קטן מאוד (שואף ל-0), נקבל שאין ל- λ השפעה וזה יגרום ל- w_0 להיות שווה לממוצע של y_iים.

2. Derive the solution for a polynomial of degree 1: $h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1 x$, by computing the derivatives w.r.t. w_0 and w_1 and writing a system of two linear equations in w_0 and w_1. No need to solve the system. Analyze the solution in the limit of $\lambda \to \infty$ and $\lambda \to 0$.

נראה כי:

$$Err(w) = \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1x_i - y_i)^2 + \lambda(w_0^2 + w_1^2)$$

$$\frac{\partial Err(w)}{\partial w_0} = \left(\frac{1}{n}\sum_{i=1}^{n} 2(w_0 + w_1x_i - y_i)\right) + \lambda(2w_0) = \left(\frac{2}{n}\sum_{i=1}^{n} w_0 + w_1x_i - y_i\right) + 2\lambda w_0$$

$$= \left(\frac{2}{n}\sum_{i=1}^{n} w_1x_i - y_i\right) + \frac{2}{n} * w_0 * n + 2\lambda w_0 = \left(\frac{2}{n}\sum_{i=1}^{n} w_1x_i - y_i\right) + 2w_0 + 2\lambda w_0$$

נגזור לפי w_1:

$$\frac{\partial Err(w)}{\partial w_1} = \left(\frac{1}{n}\sum_{i=1}^{n} 2(w_0 + w_1x_i - y_i) * x_i\right) + \lambda * 2w_1$$

נשים לב כי כאשר $\lambda \to 0$:

$$\frac{\partial Err(w)}{\partial w_0} = \left(\frac{2}{n}\sum_{i=1}^{n} w_1x_i - y_i\right) + 2w_0$$

$$\frac{\partial Err(w)}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1x_i - y_i) * x_i$$

כלומר אין ל- λ חשיבות ובחירת המקדמים תושפע בעיקר מהמדגם.

באופן כללי, ככל שניקח λ קטן יותר, בחירת המקדמים לפונקציה תסתמך יותר על הדגימות, ואם שונות המדגם גדולה יחסית זה עלול להיות בעייתי עבורנו כיוון שאולי נקבל פונקציית שגיאה לא מספיק טובה.

וכאשר $\lambda \to \infty$, נחלק למקרים:

- $w_0 = 0, w_1 = 0$:

$$\frac{\partial Err(w)}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(-y_i)$$

$$\frac{\partial Err(w)}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n} x_i(-y_i)$$

- $w_0 = 0, w_1 \neq 0$:

$$\frac{\partial Err(w)}{\partial w_0} = \frac{2}{n}\sum_{(i=1)}^{n}(w_1 x_i - y_i)$$
$$\frac{\partial Err(w)}{\partial w_1} \to (\infty\ OR - \infty)$$

- $w_0 \neq 0, w_1 = 0$

$$\frac{\partial Err(w)}{\partial w_0} \to (\infty\ OR - \infty)$$
$$\frac{\partial Err(w)}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n} x_i(w_0 - y_i)$$

- $w_0 \neq 0, w_1 \neq 0$:

$$\frac{\partial Err(w)}{\partial w_0} \to (\infty\ OR - \infty)$$
$$\frac{\partial Err(w)}{\partial w_1} \to (\infty\ OR - \infty)$$

מסקנה – ככל שהנגזרת של הפונקציה גדולה יותר, כך השיפוע שלה גדול יותר.
כאשר λ הולך וגדל עד אינסוף, ניתן לראות שרק עבור $w_0 = w_1 = 0$ השינויים בפונקציית השגיאה לא קיצוניים.
במצב זה השגיאה תלויה רק במדגמים.
באופן כללי, ככל ש- λ גדלה, המקדמים w_0, w_1 צריכים להיות קרובים יותר ל-0 כדי להביא את פונקציית השגיאה למינימום. קירוב המקדמים ל-0 מקטין את מידת ההשפעה של מחלקת ההיפותזות, שכן הביטוי $h_w(x) = w_0 + w_1 x$ נהיה קטן יותר. ביטוי זה מייצג את ה"מרחק" שלנו מהדגימות.

2. Logistic regression

1. Prove that the logistic regression classifier is equivalent to a softmax over a linear multiclass classifier for two classes $y = \text{“}a\text{”}, y = \text{“}b\text{”}$, when their separating hyperplanes obey $\mathbf{w}_a = -\mathbf{w}_b$.

ניזכר כי סיווג לפי $logistic\ regression$ מוגדר כך:

קלט: $x \in \mathbb{R}^d$, פלט: $y \in \{0,1\}$ שניתן להקביל ל- $y \in \{\text{"a"},\text{"b"}\}$ והמודל שלנו ילמד את הפרמטר w.

$$Sigmoid: \sigma(w^T x) = \frac{1}{1 + e^{-(w^T x)}} = \frac{e^{w^T x}}{e^{w^T x} + 1}$$

$$P(y = \text{"a"}\ |x) = \sigma(w^T x) = \frac{e^{w^T x}}{e^{w^T x} + 1}$$

$$P(y = \text{"b"}\ |x) = 1 - \sigma(w^T x) = 1 - \frac{e^{w^T x}}{e^{w^T x} + 1} = \frac{e^{-w^T x}}{e^{-w^T x} + 1}$$

נקבע שבהינתן קלט $x \in \mathbb{R}^d$, נסווג אותו להיות "a" אם מתקיים:

$$P(y = \text{"a"}\ |x) > \frac{1}{2}$$

נקבע שבהינתן קלט $x \in \mathbb{R}^d$, נסווג אותו להיות "b" אם מתקיים:

$$P(y = \text{"b"}\ |x) > \frac{1}{2} \quad \Leftrightarrow \quad P(y = \text{"a"}\ |x) < \frac{1}{2}$$

נשים לב כי במקרה של שוויון נבחר למי לתת עדיפות באופן שרירותי.

נפשט את המסווג:

$$\frac{e^{w^T x}}{e^{w^T x} + 1} > \frac{1}{2} \quad \Leftrightarrow \quad \frac{2 \cdot e^{w^T x}}{e^{w^T x} + 1} > 1$$

מפני שמתקיים כי $\forall x \in \mathbb{R}: e^x > 0$, אז נסיק כי המכנה חיובי, ולכן התנאי מעל קורה אם ורק אם

$$2 \cdot e^{w^T x} > e^{w^T x} + 1 \quad \Leftrightarrow \quad e^{w^T x} > 1$$

באופן דומה נסווג דוגמא $x \in \mathbb{R}^d$ להיות "b" אם מתקיים:

$$e^{w^T x} < 1$$

כעת עבור סיווג לפי $softmax\ over\ a\ linear\ multiclass$ לשתי מחלקות $y = "a", y = "b"$ כך ש- $w_a = -w_b$:

יש רק שתי מחלקות, אז נסווג דוגמא $x \in \mathbb{R}^d$ להיות "a" אם מתקיים:

$$\frac{e^{w_a^T * x}}{e^{w_a^T * x} + e^{w_b^T * x}} > \frac{e^{w_b^T * x}}{e^{w_a^T * x} + e^{w_b^T * x}}$$

נראה כי אם נסמן $w = w_a$, אז $-w = -w_a = w_b$, לכן נתייחס אליהם כך במשוואות.
נסווג דוגמא $x \in \mathbb{R}^d$ להיות "a" אם מתקיים:

לכן סה"כ נסווג דוגמא $x \in \mathbb{R}^d$ להיות "a" אם מתקיים:

$$\frac{e^{w^T x}}{e^{w^T x} + e^{-w^T x}} > \frac{e^{-w^T x}}{e^{w^T x} + e^{-w^T x}}$$

באופן דומה נסווג דוגמא $x \in \mathbb{R}^d$ להיות "b" אם מתקיים:

$$\frac{e^{w^T x}}{e^{w^T x} + e^{-w^T x}} < \frac{e^{-w^T x}}{e^{w^T x} + e^{-w^T x}}$$

שוב נשים לב כי במקרה של שוויון נבחר למי לתת עדיפות באופן שרירותי.

נפשט את המסווג השני:

$$\frac{e^{w^T x}}{e^{w^T x} + e^{-w^T x}} > \frac{e^{-w^T x}}{e^{w^T x} + e^{-w^T x}}$$

שוב $e^x > 0$ לכל $x \in \mathbb{R}$, לכן המכנה חיובי, ומכיוון שהם שווים נקבל כי:

$$\frac{e^{w^T x}}{e^{w^T x} + e^{-w^T x}} > \frac{e^{-w^T x}}{e^{w^T x} + e^{-w^T x}} \Leftrightarrow e^{w^T x} > e^{-w^T x} \overset{\div\left(e^{-w^T x}\right)}{\Leftrightarrow} e^{2w^T x} > 1$$

אנו יודעים כי $e^x > 0$ לכל $x \in \mathbb{R}$, לכן נוציא שורש משני האגפים ונקבל:

$$e^{w^T x} > 1$$

סה"כ נקבל שנסווג דוגמא $x \in \mathbb{R}^d$ להיות "a" אם מתקיים:

$$e^{w^T x} > 1$$

באופן דומה נסווג דוגמא $x \in \mathbb{R}^d$ להיות "b" אם מתקיים (סימטרי):

$$e^{w^T x} < 1$$

סה"כ קיבלנו שבהינתן דוגמא $x \in \mathbb{R}^d$ נסווג אותה להיות "a" לפי המודל הראשון אם ורק אם סיווגנו אותה להיות "a" לפי המודל השני, ונסווג אותה להיות "b" לפי המודל הראשון אם ורק אם סיווגנו אותה להיות "a" לפי המודל השני, ולכן סיווג על פי כל אחד מהמודלים הוא שקול, כנדרש.

■

2. For a vector $\mathbf{z} \in \mathbb{R}^K$, consider the regular softmax function:

$$\text{softmax}_i(\mathbf{z}) = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \tag{1}$$

For any vector $\mathbf{b} = (b, \ldots, b) \in \mathbb{R}^K$ for some $b \in \mathbb{R}$, prove that $\text{softmax}_i(\mathbf{z}) = \text{softmax}_i(\mathbf{z} - \mathbf{b})$ for any $1 \leq i \leq K$.

יהי $b = (b, \ldots, b) \in \mathbb{R}^K$.

$$softmax_i(z) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

$$\boldsymbol{softmax_i(z - b) = \frac{e^{z_i - b}}{\sum_{k=1}^{K} e^{z_k - b}}}$$

$$\frac{e^{z_i - b}}{\sum_{k=1}^{K} e^{z_k - b}} = \frac{e^{z_i} * e^{-b}}{\sum_{k=1}^{K} e^{z_k} * e^{-b}}$$

e^{-b} הוא קבוע, לכן נוציא אותו מהסכום במכנה:

$$= \frac{e^{-b} * e^{z_i}}{e^{-b} \sum_{k=1}^{K} e^{z_k}} = \frac{\boldsymbol{e^{z_i}}}{\boldsymbol{\sum_{k=1}^{K} e^{z_k}}} = \boldsymbol{softmax_i(z)}$$

3. For a vector $\mathbf{z} \in \mathbb{R}^K$, consider the softmax function that is scaled by a constant $T \in \mathbb{R}$:

$$f_i(\mathbf{z}) = \frac{\exp(Tz_i)}{\sum_{k=1}^{K} \exp(Tz_k)} \tag{2}$$

Further, for a vector $\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$, instead of considering the $\arg\max(z_1, \ldots, z_K)$ function as a function with categorical output $1, \ldots, K$ (corresponding to the index of a vector's largest element), consider the arg max function with **one-hot** representation of the output (assuming there is a unique maximum element):

$$\arg\max(z_1, \ldots, z_K) = (y_1, \ldots, y_K) = (0, \ldots, 0, 1, 0, \ldots, 0) \qquad (3)$$

where $y_i = 1$ if and only if $i = \arg\max(z_1, \ldots, z_K)$, meaning that z_i is the unique maximum value of $\mathbf{z} = (z_1, \ldots, z_K)$.

(a) For any vector $\mathbf{z} \in \mathbb{R}^K$ whose maximum element is unique, show that the softmax converges to the arg max function as $T \to \infty$, i.e., prove that:

$$\lim_{T\to\infty} (f_1(\mathbf{z}), \ldots, f_K(\mathbf{z})) = \arg\max(z_1, \ldots, z_K) \qquad (4)$$

when arg max is in **one-hot** encoding.

$$f_i(z) = \frac{e^{Tz_i}}{\sum_{k=1}^{K} e^{Tz_k}}$$

נסמן z_m להיות האיבר המקסימלי בוקטור z, ואת m להיות האינדקס שלו, כלומר:

$$z_m = \max(z_1, z_2, \ldots, z_K)$$

נראה שעבור $T \to \infty$ נקבל כי:

$$\sum_{k=1}^{K} e^{Tz_k} \approx e^{Tz_m}$$

נראה כי עבור $\boldsymbol{i = m}$:

$$\lim_{T\to\infty} \boldsymbol{f_i(z)} = \lim_{T\to\infty} \frac{e^{Tz_m}}{\sum_{k=1}^{K} e^{Tz_k}} = \lim_{T\to\infty} \frac{1}{\sum_{k=1}^{K} e^{Tz_k} - e^{Tz_m}} = \frac{1}{1} = \mathbf{1}$$

ועבור כל $\boldsymbol{i \neq m}$:

$$\lim_{T\to\infty} \boldsymbol{f_i(z)} = \lim_{T\to\infty} \frac{e^{Tz_i}}{\sum_{k=1}^{K} e^{Tz_k}} \le \lim_{T\to\infty} \frac{e^{Tz_i}}{e^{Tz_m}} = \lim_{T\to\infty} e^{Tz_i - Tz_m} = e^{T(z_i - z_m)}$$

$$z_m = \max(z_1, \ldots, z_K)$$

ולכן $z_m > z_i$, כלומר:

$$\underset{T\to\infty}{T(z_i - z_m)} = -\infty$$

ולכן: $\underset{T\to\infty}{e^{T(z_i - z_m)}} = e^{-\infty} = \mathbf{0}$

כלומר כאשר $T \to \infty$ נקבל כי כאשר $i = m$: $f_i(z) \to 1$ ולכל $i \neq m$ מתקיים: $f_i(z) \to 0$.

מסקנה – כאשר $T \to \infty$ נקבל כי הפונקציית $softmax$ שלנו שואפת לביטוי $one-hot$ של האינדקס המקסימלי ב- z, כלומר ל- $[0,0,\ldots,\underset{m'th\ index}{1},0,\ldots,0]$.

ולכן $\lim_{T\to\infty} \left(f_1(z), f_2(z), \ldots, f_K(z)\right) = argmax(z_1, z_2, \ldots, z_K)$

(b) For any vector $\mathbf{z} \in \mathbb{R}^K$ whose maximum element is **not necessarily** unique, compute $\lim_{T\to\infty}(f_1(\mathbf{z}),\ldots,f_K(\mathbf{z}))$ and provide a literal interpretation for your result.

נסמן $M = \{m_1, m_2, \ldots, m_r\}$ קבוצת האינדקסים של כל האיברים המקסימליים.

כלומר:

$$\max(z_1, z_2, \ldots, z_K) = z_{m_1} = z_{m_2} = \cdots = z_{m_r}$$

כעת נראה כי בדומה לסעיף הקודם, מתקיים כאשר $T \to \infty$:

$$\sum_{k=1}^{K} e^{Tz_k} \approx \sum_{m\in M}^{r} e^{Tz_m} = |M| * e^{Tz_{max}}$$

לכן עבור כל אינדקס $i \notin M$:

$$f_i(z) = \frac{e^{Tz_i}}{\sum_{k=1}^{K} e^{Tz_k}} \approx \frac{e^{Tz_i}}{|M| * e^{Tz_{max}}} = e^{T(z_i - |M|z_{max})}$$

$z_i - |M|z_{max} < 0$ ולכן $\underset{T\to\infty}{T(z_i - |M|z_{max})} = -\infty$, ולכן:

$$\underset{T\to\infty}{f_i(z)} = e^{-\infty} = 0$$

ועבור $i \in M$ מתקיים:

$$f_i(z) = \frac{e^{Tz_{max}}}{\sum_{k=1}^{K} e^{Tz_k}} \approx \frac{e^{Tz_{max}}}{|M|e^{Tz_{max}}} = \frac{\mathbf{1}}{|\boldsymbol{M}|}$$

כאשר הערך המקסימלי בוקטור z אינו ייחודי, אז עבור: $\lim_{T\to\infty}(f_1(z), \ldots, f_K(z))$ מתקיים:

$\boldsymbol{f_i}(\boldsymbol{z}) = \frac{\mathbf{1}}{|\boldsymbol{M}|}$ כאשר $i \in M$ ו- M זה קבוצת האינדקסים שהם בעלי ערך מקסימלי בוקטור z.

ו- $f_i(z) = 0$ כאשר $i \notin M$.

כלומר כאשר הערך המקסימלי בוקטור z אינו יחיד, פונקציית ה- $scaled\ softmax$ מחלקת את מסת ההסתברות באופן שווה בין האינדקסים של הערכים המקסימליים בוקטור z וזה משקף התפלגות אחידה על פני האינדקסים של הערכים המקסימליים, כלומר כל ערך מקסימלי הוא בעל סיכוי שווה להיבחר בגבול כאשר $T \to \infty$.

(c) For any vector $\mathbf{z} \in \mathbb{R}^K$, what happens when $T \to 0$? Namely, compute the limit $\lim_{T\to 0}(f_1(\mathbf{z}),\ldots,f_K(\mathbf{z}))$.

כאשר $T \to 0$ אז לכל $1 \le i \le K$ מתקיים:

$e^{Tz_i} \approx e^{0z_i} = e^0 = 1$, ולכן:

$$f_i(z) = \frac{e^{Tz_i}}{\sum_{k=1}^{K} e^{Tz_k}} \approx \frac{e^{0*z_i}}{\sum_{k=1}^{K} e^{0z_k}} = \frac{1}{K}$$

כלומר כאשר $T \to 0$ מתקיים:

$$\lim_{T\to 0}\left(f_1(z),\dots,f_K(z)\right)=\left(\frac{1}{K},\frac{1}{K},\dots,\frac{1}{K}\right)$$

כלומר הפונקציית $softmax$ שלנו תתפלג יוניפורמית לכל איבר, כלומר כל איבר ייבחר בצורה שווה לא משנה מה ערכו.

4. Write the gradient update rule for a logisitic regression model, when the usual loss of the negative log likelihood is now regularized with the square of the L_2 norm over the weight vector $\frac{1}{2}||\mathbf{w}||^2$.

במקרה ללא תקנון ($regularization$) אנו מגדירים את פונקציית ה- $loss$ להיות:

$$Err=-\log(L)=\sum_{i=1}^{n}y_i\log\left(\sigma(w^Tx)\right)-\sum_{i=1}^{n}(1-y_i)\log\left(1-\sigma(w^Tx)\right)$$

וכפי שראינו הגרדיאנט של פונקציית ה- $loss$ היא:

$$\nabla Err=\sum_{i=1}^{n}x_i[\hat{y_i}-y_i]$$

ולכן כלל העדכון במקרה זה הוא:

$$\boldsymbol{w^{t+1}=w^t-\eta x_i[\hat{y_i}-y_i]}$$

וכאשר נשתמש בתקנון הנתון, נקבל:

$$Err_{reg}=Err+\frac{1}{2}\|w\|^2$$

כלומר:

$$Err_{reg}=\sum_{i=1}^{n}y_i\log\left(\sigma(w^Tx)\right)-\sum_{i=1}^{n}(1-y_i)\log\left(1-\sigma(w^Tx)\right)+\frac{1}{2}\|w\|^2$$

נחפש בנפרד את הגרדיאנט של $\frac{1}{2}\|w\|^2$:

הגרדיאנט של $\frac{1}{2}\|w\|^2$ לפי w הוא:

$$\nabla\left(\frac{1}{2}\|w\|^2\right)=w$$

לכן הגרדיאנט של Err_{reg} הוא:

$$\nabla Err_{reg}=\nabla Err+\nabla\left(\frac{1}{2}\|w\|^2\right)$$

כלומר הכלל עדכון של הגרדיאנט הוא:

$$\boldsymbol{w^{t+1}=w^t-\eta\left(x_i[\hat{y_i}-y_i]+w^t\right)}$$

■