תרגיל 3

אור שמואל 211195184

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 $\lambda \to 0$.
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 $\lambda \to 0$.

1

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

1

$$\Downarrow \qquad H_w(x) = w_0$$

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

W [illegible] $\Downarrow$

$$Err(w)' = \frac{1}{n}\cdot 2\sum_{i=1}^{n}(w_0 - y_i) + 2\lambda w_0 = 0$$

$$\Downarrow$$

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

$$\Downarrow \quad / \frac{n}{2}$$

$$\sum_{i=1}^{n}(w_0 - y_i) = -\lambda n w_0$$

$$\Downarrow$$

$$n w_0 - \sum_{i=1}^{n} y_i = -\lambda n w_0$$

$$\Downarrow$$

$$n w_0 + \lambda n w_0 = \sum_{i=1}^{n} y_i$$

$$\Downarrow$$

$$w_0 (n + \lambda n) = \sum_{i=1}^{n} y_i$$

$$\Downarrow$$

$$w_0 = \frac{\sum_{i=1}^{n} y_i}{n + \lambda n}$$

$\lambda \to \infty$ [illegible] [illegible]

$\theta \to (0,1)$

$\lambda \to 0$ [illegible]

$$w_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

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 $\lambda \to 0$.

(2)

$$Err_L(w) = \frac{1}{n}\sum_{i=1}^{n} \left(h_w(x_i) - y_i\right)^2 + \lambda \|w\|^2$$

$$h_w(x_i) = w_0 + w_1 x_i$$

$$\Downarrow \quad \|w\|^2 = w_0^2 + w_1^2$$

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

נגזר חלקית לפי w_0

$$\frac{1}{n}\sum_{i=1}^{n} 2\left(w_0 + w_1 x_i - y_i\right) + 2\lambda w_0 = 0$$

$$\Downarrow \cdot \frac{n}{2}$$

$$\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right) + \lambda n w_0 = 0$$

$$\Downarrow$$

$$n w_0 + w_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i + \lambda n w_0 = 0$$

$$\Downarrow$$

$$w_0(n+\lambda n) = \sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i$$

$$\Downarrow$$

$$w_0 = \frac{\sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i}{n(1+\lambda)}$$

כאשר

$\lambda \to 0$

נקבל

$$w_0 = \frac{\sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i}{n}$$

וכאשר

$\lambda \to \infty$

נקבל

$w_0 = 0$

כעת נגזור לפי w_1

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

$$Err'(w) = \frac{1}{n}\sum_{i=1}^{n} 2 \cdot x_i (w_0 + w_1 x_i - y_i) + 2\lambda w_1 = 0$$

$$\Downarrow \cdot \frac{n}{2}$$

$$\sum_{i=1}^{n} x_i (w_0 + w_1 x_i - y_i) + \lambda n w_1 = 0$$

$$\Downarrow$$

$$w_0 \sum_{i=1}^{n} x_i + w_1 \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \lambda n w_1$$

$$\Downarrow$$

$$w_0 \sum_{i=1}^{n} x_i + w_1 \cdot \left(\sum_{i=1}^{n} x_i^2 + \lambda n\right) - \sum_{i=1}^{n} x_i y_i = 0$$

$$\Downarrow$$

$$w_1 \cdot \left(\sum_{i=1}^{n} x_i^2 + \lambda n\right) = \sum_{i=1}^{n} x_i y_i - w_0 \sum_{i=1}^{n} x_i$$

$$\Downarrow$$

$$w_1 = \frac{\sum_{i=1}^{n} x_i y_i - w_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 + \lambda n}$$

כאשר $\lambda \to \infty$

נקבל $w_1 = 0$

וכאשר $\lambda \to 0$ נקבל:

$$w_1 = \frac{\sum_{i=1}^{n} x_i y_i - w_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

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 $\mathbf{w}_a = -\mathbf{w}_b$.

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 $1 \leq i \leq K$.

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):

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

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:

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

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

   (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.

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

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 $\frac{1}{2}||\mathbf{w}||^2$.

2

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 $\mathbf{w}_a = -\mathbf{w}_b$.

1

$w_a = -w_b$ [illegible]

$w' = 2w_a$ [illegible] Logistic regression [illegible]

$P(y = \text{"}a\text{"} \mid x) = \sigma(2w_a^T x)$

$P(y = \text{"}b\text{"} \mid x) = 1 - \sigma(2w_b^T x)$

. Soft max [illegible]

$$P(y=\text{"a"} \mid X) = \frac{e^{w_a^T x}}{e^{w_a^T x} + e^{w_b^T x}}$$

$$\Downarrow \quad w_a = -w_b$$

$$\frac{e^{w_a^T x}}{e^{w_a^T x} + e^{-w_a^T x}}$$

$$\Downarrow$$

$$\frac{1}{e^{-2w_a^T x} + 1} = \sigma(2w_a^T X)$$

ולכן $c = \sigma(2w_a^T x) = \sigma(2w_a^T X)$ כנדרש.

2

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

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

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

$$\mathrm{softmax}_i(z-b) = \frac{e^{z_i - b}}{\sum_{i=1}^{K} e^{z_k - b}} = \frac{e^{z_i} \cdot e^{-b}}{\sum_{i=1}^{K} e^{z_k} \cdot e^{-b}} =$$

$$\frac{e^{z_i} \cdot \cancel{e^{-b}}}{\cancel{e^{-b}} \sum_{i=1}^{K} e^{z_k}} = \frac{e^{z_i}}{\sum_{i=1}^{K} e^{z_k}} = \mathrm{softmax}_i(z)$$

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):

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

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:

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

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

(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.

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

③

Ⓐ

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

$$\lim_{T\to\infty}(f_i(z), \ldots, f_K(z)) = \arg\max(z_1, \ldots, z_K) =$$

נניח עבור i ; כלומר ערך שווה למקסימום

נניח -

$$f_i(z) = \frac{\exp(Ti)}{\sum_{k=1}^{K} \exp(Tz_k)} = \frac{\exp(Ti)}{\exp(Ti) + \sum_{\substack{k=1 \\ k\neq i}}^{K} \exp(Tz_k)}$$

$$\frac{\cancel{\exp(T_i)}}{\cancel{\exp(T_i)}\left(1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{\exp(T_{2k})}{\exp(T_i)}\right)} =$$

$$\frac{1}{1+\sum_{\substack{k=1\\k\neq i}}^{K}\frac{\exp(T_{2k})}{\exp(T_i)}} = \frac{1}{1+\sum_{k=1}^{K}\exp(T(2k-i))} \underset{\substack{i \text{ גדול}\\ \text{מספיק}}}{=} \frac{1}{\infty} = 0$$

כאשר עבור $i = max$

$$f_i(2) = \frac{\exp(T_{max})}{\sum_{k=1}^{K}\exp(T_{2k})} = \frac{\exp(T_{max})}{\exp(T_{max}) + \sum_{\substack{k=1\\k\neq max}}^{K}\exp(T_{2k})} =$$

$$\frac{\cancel{\exp(T_{max})}}{\cancel{\exp(T_{max})}\left(1+\sum_{\substack{k=1\\k\neq max}}^{K}\frac{\exp(T_{2k})}{\exp(T_{max})}\right)} = \frac{1}{}$$

$$\frac{1}{1+\sum\limits_{\substack{k=1\\k\neq max}}^{n}\frac{\exp(T_{2k})}{\exp T_{max}}} = \frac{1}{1+\sum\limits_{k=1}^{K}\exp(T(2k-2max))} =$$

$2max$

[illegible]

[illegible]

$$\frac{1}{1+0} = 1$$

(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.

ב

נוכיח ומקרים,

כאשר יש מקסימום 1 יחידי מספר א.

כלומר נראה מה קורה שיש כמה.

נניח שיש m מקסימומים לכן שווים ל $\exp(Tz_{max})$.

נסמן $i = z_{max}$.

$$f_i(z) = \frac{\exp(Tz_{max})}{\sum_{k=1}^{K} \exp(Tz_k)} = \frac{\exp(Tz_{max})}{m\cdot\exp(Tz_{max}) + \sum_{\substack{k=1 \\ k\neq z_{max}}}^{K} \exp(Tz_k)} =$$

נניח כי יש m מקסימומים

$$\frac{\exp(Tz_{max})}{m\cdot\exp(Tz_{max})\cdot\left(1 + \frac{\sum_{\substack{k=1 \\ k\neq z_{max}}}^{K} \exp(Tz_k)}{m\exp(Tz_{max})}\right)} = \frac{1}{m+0} = \frac{1}{m}$$

[illegible] [illegible] [illegible] ס כמו [illegible]

[illegible]

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

$$f_i(z) = \frac{\exp(T z_i)}{\sum_{k=1}^{K} \exp(T z_k)}$$

לכן נראה כאשר $T \to 0$ נקבל -

$$\frac{\exp(T z_i)}{\sum_{k=1}^{K} \exp(T z_k)} = \frac{\exp(T z_i)}{\exp(T z_i) + \sum_{\substack{k=1 \\ k \neq i}}^{K} \exp(T z_k)} =$$

$$\frac{\cancel{\exp(T z_i)}}{\cancel{\exp(T z_i)} \cdot \left(1 + \frac{\sum_{\substack{k=1 \\ k \neq i}}^{K} \exp(T z_k)}{\exp(T z_i)} \right)} = \frac{1}{1 + \frac{\sum_{\substack{k=1 \\ k \neq i}}^{K} \exp(T z_k)}{\exp(T z_i)}} \underset{\substack{T \to 0 \\ e^0 = 1}}{=}$$

$$\frac{1}{1 + \sum_{\substack{k=1 \\ k \neq i}}^{K} 1} = \frac{1}{K}$$

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 $\frac{1}{2}||\mathbf{w}||^2$.

4

w - weight vector

$x^{(i)}$ - feature vector for the sample

$y^{(i)}$ - true label for the i sample

n - number of training samples

λ - regularization parameter

$\hat{y}^{(i)} = \sigma(w^T x^{(i)}) = \frac{1}{1+e^{-w^T x^i}}$ logistic regression

$$J(w) = L(w) + \frac{\lambda}{2} ||w||^2 = -\frac{1}{n} \sum_{i=1}^{n} y^{(i)} \cdot \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)}).$$

$$\Delta_w J(w) = -\frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)}) \cdot x^{(i)} + \lambda w$$

Gradient update rule !!

$$w \leftarrow w - \alpha \nabla_w J(w)$$

α = learning rate !!

$$w \leftarrow w + \alpha \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right) \cdot x^{(i)} - \alpha \lambda w$$

ומגל -