Machine Learning

Homework 3

Ari Feiglin

Exercise 3.1

Consider the total loss function for polynomial fitting:

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

(1) Derive a solution for a zero-degree polynomial. Analyze this solution as $\lambda \to 0, \infty$.

(2) Derive a solution for a one-degree polynomial. Analyze this solution as $\lambda \to 0, \infty$.

(1) Err is the ridge loss function, i.e. it is just

$$Err(\mathbf{w}) = \frac{1}{n}\|X\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$$

where k is the degree of the polynomial we are fitting and

$$X = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^k \\ \vdots & \vdots & & \vdots \\ x_n^0 & x_n^1 & \cdots & x_n^k \end{pmatrix}$$

and we saw in lecture that this takes a minimum when

$$\widehat{\mathbf{w}}_{\text{ridge}} = (X^\top X + \lambda n I)^{-1} X^\top \mathbf{y}$$

where $I = I_n$. Here, because $k = 0$ we have that

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

and so

$$\widehat{\mathbf{w}}_{\text{ridge}} = (n + \lambda n)^{-1}\sum_{i=1}^{n} y_i = \frac{1}{n(1+\lambda)}\sum_{i=1}^{n} y_i$$

Thus when $\lambda \to 0$, $\widehat{\mathbf{w}}_{\text{ridge}} \to \frac{1}{n}\sum_{i=1}^{n} y_i$, and when $\lambda \to \infty$ it approaches 0.

(2) Here we have that

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

So

$$X^\top X + \lambda n I = \begin{pmatrix} n(1+\lambda) & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \qquad X^\top \mathbf{y} = \begin{pmatrix} \sum_j y_j \\ \sum_j x_j y_j \end{pmatrix}$$

Thus

$$\widehat{\mathbf{w}}_{\text{ridge}} = \frac{1}{n(1+\lambda)\sum_i x_i^2 - \sum_{i,j} x_i x_j}\begin{pmatrix} \sum_{i,j} y_j x_i (x_i - x_j) \\ -\sum_{i,j} y_i (x_j + (1+\lambda)x_i) \end{pmatrix}$$

So as $\lambda \to 0$:

$$\widehat{\mathbf{w}}_{\text{ridge}} \longrightarrow \frac{1}{\sum_{i,j} x_i(x_i - x_j)}\begin{pmatrix} \sum_{i,j} y_j x_i (x_i - x_j) \\ -\sum_{i,j} y_i (x_i + x_j) \end{pmatrix}$$

And as $\lambda \to \infty$ by LHopital:

$$\widehat{\mathbf{w}}_{\text{ridge}} \longrightarrow \begin{pmatrix} 0 \\ -\frac{\sum_i y_i x_i}{\sum_i x_i^2} \end{pmatrix}$$

Exercise 3.2

(1) For a vector $\mathbf{z} \in \mathbb{R}^K$ define the softmax function

$$\text{softmax}_i(\mathbf{z}) = \frac{\exp(\mathbf{z}_i)}{\sum_{k=1} \exp(\mathbf{z}_k)}$$

for a vector $b\mathbf{1} \in \mathbb{R}^K$, show that $\text{softmax}_i(\mathbf{z} + b\mathbf{1}) = \text{softmax}_i(\mathbf{z})$.

(2) Define

$$f_i(\mathbf{z}) = \text{softmax}_i(T\mathbf{z})$$

and consider the **one-hot** representation of the argmax function:

$$\text{argmax}(\mathbf{z}) = e_{\text{argmax}\,\mathbf{z}}$$

(i) For any $\mathbf{z}$ whose maximum element is unique, show that

$$\lim_{T\to\infty} (f_1(\mathbf{z}), \ldots, f_K(\mathbf{z})) = \text{argmax}(\mathbf{z})$$

(ii) For a $\mathbf{z}$ whose maximum is not necessarily unique, compute an expression for

$$\lim_{T\to\infty} (f_1(\mathbf{z}), \ldots, f_K(\mathbf{z}))$$

(iii) What happens when $T \to 0$?

(3) Write the gradient update rule for a logistic regression model, when the usual loss of the negative log likelihood is regularized by $\frac{1}{2}\|\mathbf{w}\|^2$.

(1) By definition

$$\text{softmax}(\mathbf{z} + b\mathbf{1}) = \frac{\exp(\mathbf{z}_i + b)}{\sum_k \exp(\mathbf{z}_k + b)} = \frac{\exp(\mathbf{z}_i)\exp(b)}{\sum_k \exp(\mathbf{z}_k)\exp(b)} = \frac{\exp(\mathbf{z}_i)}{\sum_k \exp(\mathbf{z}_k)} = \text{softmax}(\mathbf{z})$$

(2)

(i) Suppose $\text{argmax}(z) = i$, then for every $j \neq i$, $\exp(Tz_j)/\exp(Tz_i) = \exp(T(z_j - z_i))$ and since $z_j - z_i < 0$, $\exp(T(z_j - z_i)) \to 0$ as $T \to \infty$. And so

$$f_i(\mathbf{z}) = \frac{1}{\sum_j \exp(Tz_j)/\exp(Tz_i)} \longrightarrow \frac{1}{\sum_j \delta_{ij}} = 1$$

And so $\exp(Tz_i)/\exp(Tz_j) \to \infty$ and thus

$$f_j(\mathbf{z}) = \frac{1}{\sum_k \exp(Tz_k)/\exp(Tz_j)} = \frac{1}{\exp(Tz_i)/\exp(Tz_k) + \sum_{k\neq i} \exp(Tz_k)/\exp(Tz_j)} \longrightarrow 0$$

thus since convergence in $\mathbb{R}^K$ is equivalent to pointwise convergence,

$$\lim_{T\to\infty} (f_1(\mathbf{z}), \ldots, f_K(\mathbf{z})) = e_i$$

as required.

(ii) Suppose I is the set of indexes for which z_i are maximal. Then for $i \in I$:

$$f_i(\mathbf{z}) = \frac{1}{\sum_{j\in I} \exp(Tz_j)/\exp(Tz_i) + \sum_{j\notin I} \exp(Tz_j)/\exp(Tz_i)} \longrightarrow \frac{1}{\sum_{j\in I} 1} = \frac{1}{|I|}$$

And for $i \notin I$:

$$f_i(\mathbf{z}) = \frac{1}{\sum_{j \in I} \exp(Tz_j)/\exp(Tz_i) + \sum_{j \notin I} \exp(Tz_j)/\exp(Tz_i)} \longrightarrow 0$$

since $\exp(Tz_j)/\exp(Tz_i) \to \infty$. Thus

$$\lim_{T \to \infty} (f_1(\mathbf{z}), \dots, f_K(\mathbf{z})) = \frac{1}{|I|} \sum_{i \in I} e_i, \qquad I = \{1 \le i \le K \mid z_i \text{ is maximal}\}$$

(iii) For any i, j, $\lim_{T \to 0} \exp(Tz_j)/\exp(Tz_i) = 1$. Thus

$$f_i(\mathbf{z}) = \frac{1}{\sum_j \exp(Tz_j)/\exp(Tz_i)} = \frac{1}{K}$$

so the limit is just $\frac{1}{K}\mathbf{1}$.

(**3**) The total loss function just becomes

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

where $\sigma = \sigma(\mathbf{w}^\top \mathbf{x}_i)$. Then the gradient becomes

$$\nabla\, Err(\mathbf{w}) = \nabla \left(\sum_{i=1}^{n} y_i \log \sigma + \sum_{i=1}^{n} (1 - y_i) \log(1 - \sigma) \right) + \mathbf{w} = -\sum_{i=1}^{n} \big(y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)\mathbf{x}_i + \mathbf{w}$$

So the update rule becomes (as we take the ith component of the sum of the gradient):

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \left(\sigma(\mathbf{w}^\top \mathbf{x}_i) - y_i + \frac{1}{n}\mathbf{w} \right)$$

Since $\mathbf{w}$ gives a component of $\frac{1}{n}\mathbf{w}$ to each of the summands.