Homework 5

STATS205 (Spring 2021–2022)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. **Up to two additional points will awarded for especially well-organized and elegant solutions.**

Due date: Tuesday, June 7, 2022.

Corrections/Clarifications of Monday June 6 in red.

Corrections/Clarifications of Tuesday June 7 in cyan.

1. Single-Hidden Layer Neural Nets, $d = 1$ (*15 points*)

Last week we considered ridge regression and similar methods in what we called the 'Hinge Feature map' setting:

Suppose we are given n data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \mathbb{R}$. Suppose the x-data are strictly sorted $x^{(1)} < x^{(2)} < \ldots < x^{(n)}$.

Consider the feature map $\phi : \mathbb{R} \to \mathbb{R}^m$ where $m = n$, based on so call *hinge features*. Namely, let $\phi(x) = (\phi_j(x))_{j=0}^{n-1}$, where

$$\phi_j(x) = (x - x^{(j)})_+, \qquad j = 1, \ldots, n-1.$$

with $(u)_+ \equiv max(0, u)$, and $\phi_0(x) = 1$.

In contrast, the current problem considers the 'Single Hidden Layer Relu Neural Net' (ReLu-1-NN) setting. In this setting, we are fitting a model with predictions:

$$f(x; \Theta) = \Theta_0 + \sum_{j=1}^{d_1} w_j^2 \text{ReLu}(w_j^1 x - b_j^1)$$

and parameter $\Theta = (\Theta_0, \Theta_b, \Theta_{w_1}, \Theta_{w_2})$ where $\Theta_0 \in \mathbb{R}$ and

$$\Theta = (\Theta_0, (b_j^1)_{j=1}^{d_1}; (w_j^1)_{j=1}^{d_1}; (w_j^2)_{j=1}^{d_1})$$

a. (*3 points*) Matching ReLu-1-NN predictions with Hinge Feature Map.

For a given set of coefficients $\theta \in \mathbb{R}^m$ and x-data $(x^{(i)})$, the Hinge Feature Map produces a certain vector of predictions $y \equiv \Phi\theta$.

Given a specific such Hinge Feature Map vector θ, explain how to define a corresponding ReLu-1-NN parameter vector $\Theta = \Theta(\theta) \in \mathbb{R} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ with specific choices of the dimension d_1 and specific choices of entries of parameter subvectors of Θ so that the two model architectures produce identical predictions $f((x^{(i)}); \Theta(\theta)) = \Phi\theta$ under the correspondence you define between θ and the induced vector $\Theta(\theta)$.

More generally, define a class $\mathcal{T}$ of vectors $\Theta = (\Theta_0, \Theta_b, \Theta_{w_1}, \Theta_{w_2})$ obeying certain constraints, so that for each $\Theta \in \mathcal{T}$, there is a 'natural correspondence' $\theta = \theta(\Theta)$ between the ReLu-1-NN parameters Θ and induced Hinge Feature Map parameters; i.e. for ReLu-1-NN models $\Theta \in \mathcal{T}$ there is a natural $1-1$ correspondence with Hinge Feature Map models $\theta = \theta(\Theta) \in \mathbb{R}^m$, so that we get identical predictions $f((x^{(i)}); \Theta) = \Phi\theta(\Theta)$.

Your answers $\Theta(\theta)$ and $\theta(\Theta)$ should have the property that $\theta(\Theta(\theta)) = \theta$, for all $\theta \in \mathbb{R}^m$.

There perhaps will be more than one way to do this. The 'best' answer would be one freezing the parameters that enter *nonlinearly* so that they don't vary as the value of the vector θ changes; and which allows to vary with θ only the parameters that enter *linearly*.

b. (*6 points*) **Matching ReLu-1-NN Penalization with ℓ_2-penalization in Hinge Feature Map.**

Consider the following special fitting process from the Hinge Feature map setting:

$$\min_{\theta \in \mathbb{R}^m} n^{-1} \cdot \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2. \tag{1}$$

Explain how this is related to the setting of Homework 4.

Consider the Relu-1-NN setting, with would-be penalized optimization:

$$\min_{\Theta \in \mathcal{T}} n^{-1} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2 + \lambda Q(\Theta). \tag{2}$$

Here $\mathcal{T}$ is the parameter set defined in the last subproblem and $Q(\Theta)$ is an as-yet-undefined penalty function, which you are to define as part of the answer to this problem. For example it might only be a function of a subvector of Θ.

Explain how to define the penalty $Q(\Theta)$ so that, optimizing (2) is 'naturally isomorphic' to optimizing (1).

Is the optimization (2) representable as a quadratic optimization over a linear subspace?

c. (*6 points*) **Matching ReLu-1-NN Weight Decay with Hinge Feature Map.**

In the Relu-1-NN setting when one says they are using 'a weight decay penalty'. they generally mean they are optimizing an expression like:

$$n^{-1} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2 + \lambda \left(\|\Theta_{w_1}\|_2^2 + \|\Theta_{w_2}\|_2^2 \right)$$

Consider the would-be penalized optimization:

$$\min_{\Theta \in \mathcal{T}_1} n^{-1} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2 + \lambda \left(\|\Theta_{w_1}\|_2^2 + \|\Theta_{w_2}\|_2^2 \right). \tag{3}$$

Here $\mathcal{T}_1$ is a parameter set you are to define as part of the answer to this problem. In Lecture 13 we pointed to an article where it was claimed (somewhat vaguely) that this was somehow equivalent to a reparametrized version of ℓ_1 penalization in the Hinge Feature Map; seemingly something like:

$$\min_{\theta \in \mathbb{R}^m} n^{-1} \cdot \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_1. \tag{4}$$

The purpose of this problem is to make clear that we can't take this quite literally; and to see what it might really mean.

The lecture's claim might be elucidated as follows. Define a constraint set $\mathcal{T}_1$ of vectors $\Theta = (\Theta_0, \Theta_b, \Theta_{w_1}, \Theta_{w_2})$ in $\mathbb{R} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$, and let Θ_{opt} denote the solution of (3) (supposed unique).

Helped by observations in lecture, identify an 'optimum set' $\mathcal{T}_{opt} \subset \mathcal{T}_1$ such that the solution Θ_{opt} obeys further specialized constraints expressed by membership in a class $\mathcal{T}_{opt}$, $\Theta_{opt} \in \mathcal{T}_{opt}$; i.e. solving the optimization over $\mathcal{T}_1$ always has this special structure.

Define a 'natural correspondence' between such a solution $\Theta_{opt} \in \mathcal{T}_{opt}$ and *both* a Θ_{opt}-dependent Hinge Feature map $\Phi_{\Theta_{opt}}$ and a Θ_{opt}-dependent parameter vector $\theta_{opt}(\Theta_{opt})$ so that the solution of (3) is naturally corresponding to the solution of the special optimization problem:

$$\min_{\theta \in \mathbb{R}^m} \; n^{-1} \cdot \|y - \Phi_{\Theta_{opt}} \theta\|_2^2 + \lambda \|\theta\|_1. \tag{5}$$

I.e. $\theta_{opt}(\Theta_{opt})$ solves (5) and the predictions match:

$$f((x^{(i)}); \Theta_{opt}) = \Phi_{\Theta_{opt}} \theta_{opt}(\Theta_{opt}).$$

Note that in this problem, specifics of the Hinge Feature Map $\Phi_{\Theta_{opt}}$ depend on Θ_{opt}. Namely, the article's claim is (effectively) that there is indeed a correspondence between weight decay in ReLu-1-NN and ℓ_1- penalization in Hinge Feature Maps; but this correspondence demands that we specify an *adaptively-chosen* feature map.

Give a detailed description of the Hinge Feature map $\Phi_{\Theta_{opt}}$.

Is the optimization (3) representable as quadratic optimization over a linear space?

Is your 'natural correspondence' $\Theta_{opt} \mapsto \theta_{opt}(\Theta_{opt})$ linear or nonlinear?

2. Two-Hidden Layer Neural Nets, $d = 1$ (*15 points*)

a. (*2 points*) Composition of Standard ReLu's

We are interested in two-layer ReLu Neural Nets. Let $x \in \mathbb{R}$ and define the scalar ReLu nonlinearity with scalar parameter $b > 0$, $\phi_b : \mathbb{R} \mapsto \mathbb{R}$, defined via

$$\phi_b(x) = (x - b)_+.$$

Derive an expression

$$\phi_{b_2}(\phi_{b_1}(x)) = \phi_{b_3}(x).$$

where $b_3 = B_+(b_1, b_2)$, $b_i > 0$. I.e. we want an expression for $B_+()$.

~~Justify the expression by systematically considering cases where $b_2 > 0$ and $b_2 < 0$.~~

b. (*2 points*) Composition of Weighted ReLu's

Let $w_1 > 0$ and $w_2 > 0$. Derive an expression

$$w_2 \phi_{b_2}(w_1 \phi_{b_1}(x)) = w_3 \phi_{b_3}(x).$$

where $(w_3, b_3) = C_+(w_1, w_2, b_1, b_2)$, $b_i > 0$.

Justify your expression for C_+.

c. (*2 points*) Composition of ReLu's with Coefficients

Let $w_1, w_2 \in \mathbb{R}$; i.e. the w_i can be either positive or negative. Derive an expression

$$w_2 \phi_{b_2}(w_1 \phi_{b_1}(x)) = w_3 \phi_{b_3}(x).$$

where $(w_3, b_3) = C_\pm(w_1, w_2, b_1, b_2)$, $b_i > 0$.

Justify your expression for $C_\pm$.

d. (*6 points*) **Composition of Monotone NN's**

Let $f_1(x) = \sum_{i=1}^{n_1} c_i^1 \phi_{b_i^1}(x)$ and $f_2(x) = \sum_{i=1}^{n_2} c_i^2 \phi_{b_i^2}(x)$.

Suppose that $c_i^\ell \geq 0$ and that the $b_i^\ell \geq 0$ and are strictly increasing in i for each fixed $\ell = 1, 2$; i.e.$b_i^\ell > b_{i-1}^\ell$, $i = 2, 3, \dots, n_\ell$.

Show that $f_\ell(x)$ are nondecreasing functions, $\ell = 1, 2$; and, defining $x_{0,\ell} \equiv \min\{b_i^\ell : c_i^\ell > 0\}$, then f_ℓ is strictly increasing on $\{x : x > x_{0,\ell}\}$.

For the remainder of the problem, suppose that $c_i^\ell > 0$ for $i = 1, \dots, n_\ell$.

Show that $f_3(x) = f_2(f_1(x))$ is a nondecreasing function, and that it is strictly increasing on $\{x : x > x_{0,3}\}$, where $x_{0,3} \equiv max(x_{0,1}, f_1^{-1}(x_{0,2}))$. Here $f_1^{-1}(y) \equiv \sup\{x : f_1(x) \leq y\}$.

Show that $f_3(x)$ is piecewise linear.

Show that $f_3(x)$ has knots at $(b_i^1)_{i=1}^{n_1}$. You may draw a sketch.

Show that $f_3(x)$ also may have knots at $(f_1^{-1}(b_j^2) : j = 1, \dots, n_2)$. You may draw a sketch. Hint: this is related to boardwork that was done in lecture.

Explain that the knotset $\{b_i^3 : i = 1, \dots n_3\}$ of f_3 is precisely the pointset $\{b_i^1\} \cup \{f_1^{-1}(b_j^2)\}$. What is the maximal number of knots n_3?

Conclude that f_3 can be represented as

$$f_3(x) = \sum_{i=1}^{n_3} c_i^3 \phi_{b_i^3}(x)$$

Explain the following phrase:

> *A Two-Hidden Layer ReLu Deepnet on $x \in \mathbb{R}$ with positive coefficients can be represented as a piecewise linear spline. Letting n_i denote the number of knots in f_i, it might involve roughly the sum of the number of knots $n_1 + n_2$.*

e. (*3 points*) **Composition of Non-Monotone Piecewise Linear Functions**

Let $f_1(x)$ be piecewise linear interpolation of $\{(x_i, y_i) i = 0, \dots, N\}$ where the data are

$$(x_i, y_i) = \begin{cases} (i/N, +1) & i \text{ odd} \\ (i/N, -1) & i \text{ even} \end{cases}$$

Let $f_2(x)$ be the function $f_2(x) = 2|x| - 1$.

Let $f_3(x) = f_2(f_1(x))$.

Is f_3, viewed as a function on $[0, 1]$, piecewise linear? You may draw a sketch.

If so, how many knots does f_3 have on $[0, 1]$? You may draw a sketch.

Explain the following phrase:

> *A Two-Hidden Layer ReLu Deepnet on $x \in \mathbb{R}$ with any coefficients whatsoever can still be represented as a piecewise linear spline. If some coefficients of f_1 or f_2 are nonpositive, the spline might involve more knots than would have been required in the positive-coefficient case. Letting n_i denote the number of knots in f_i, it might involve roughly as many as $(n_1 + 1) \cdot (n_2 + 1)$ knots.*

3. Two-Hidden Layer Neural Nets, $d = 2$ (*20 points*)

Let $f_2(x) = \sum_{i=1}^{3} c_i \textsc{ReLu}(u_i' x - 1)$ where $u_i = (\cos(\theta_i), \sin(\theta_i))'$, $\theta_i = 2\pi \cdot ((i-1)/3)$ and $c_i = 1$. Here u_i is a column vector and u_i' denotes transpose, i.e. row vector. $x \in \mathbb{R}^2$ is a column vector with two entries and $u_i' x \equiv \sum_{j=1}^{2} u_i(j) x(j)$, where $u_i(j)$ denotes the j-th entry of the vector u_i, etc.

a. (*3 points*) **Plotting f_2**

Make a heatmap or contour plot of f_2 on the domain $\{(x, y) : |x| < 3, |y| < 3\}$.

Define $f_1 : \mathbb{R}^2 \mapsto \mathbb{R}^2$ via $f_1(x) = [\textsc{ReLu}(x_1 - 1), \textsc{ReLu}(x_2 + 1)]^T$; $f_1(x) \in \mathbb{R}^2$ is a column vector with two entries.

b. (*3 points*) **Plotting f_1**

Make a heatmap or contour plot of the coordinates $(f_1)_1$ and $(f_1)_2$ on the domain $\{(x, y) : |x| < 3, |y| < 3\}$.

c. (*3 points*) **Plotting the composition**

Make a heatmap or contour plot of $f_3(x) = f_2(f_1(x))$ on the domain $\{(x, y) : |x| < 3, |y| < 3\}$.

d. (*3 points*) **The Cell Decomposition induced by f_1**

As indicated in Lecture 15, there is a cell decomposition of $\mathbb{R}^2$ defined by f_1, on which f_1 is piecewise affine. Work out what this is in detail.

e. (*3 points*) **The Cell Decomposition induced by f_2**

Similarly, there is a cell decomposition of $\mathbb{R}^2$ defined by f_2, on each cell of which f_2 is piecewise affine. Work out what this is in detail.

f. (*5 points*) **The Affine Representation**

The composition $f_3(x) = f_2(f_1(x))$ is piecewise affine on $\mathbb{R}^2$. Specify one region A of $\mathbb{R}^2$ on which $f_3 = f_2(f_1)$ is affine, and specify the affine transformation involved, i.e. give the coefficients

$$f_3(x) = a + b_1 x_1 + b_2 x_2, \qquad (x_1, x_2) \in A.$$