Foundations of Data Science and Machine Learning – *Homework 3*

Isaac Martin

Last compiled March 23, 2023

EXERCISE 1. Given labeled data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$, consider the support vector machine with misclassification allowed but penalized by $\lambda > 0$:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ b\in\mathbb{R} \\ \mu\in\mathbb{R}^n}} \lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{j=1}^n \mu_j \quad \text{s.t.} \quad y_j(\langle \mathbf{w}, \mathbf{x}_j\rangle - b) \geq 1 - \mu_j,\ \mu_j \geq 0\ \forall j.$$

(a) Derive the dual problem by using the Lagrangian.

(b) Read about "Slater's condition" on Wikipedia. Does strong duality hold here?

Proof: For each j we have two constraints, one of the form $y_j(\langle \mathbf{w}, \mathbf{x}_j\rangle - b) - 1 + \mu_j \geq 0$ and another of the form $\mu_j \geq 0$. This means our Lagrangian is

$$\mathrm{cl}(\mathbf{w}, b, \mu, \alpha, \beta) = \lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{j=1}^n \mu_j \ - \ \sum_{j=1}^n \alpha_j(y_j(\langle \mathbf{w}, \mathbf{x}_j\rangle - b) - 1 + \mu_j) + \beta_j\mu_j.$$

The dual problem is then

$$\max_{\substack{\alpha\geq 0 \\ \beta\geq 0}} \min_{\mathbf{w},b,\mu} \lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{j=1}^n \mu_j \ - \ \sum_{j=1}^n \alpha_j(y_j(\langle \mathbf{w}, \mathbf{x}_j\rangle - b) - 1 + \mu_j) + \beta_j\mu_j$$

which becomes

$$\max_{\substack{\alpha\geq 0 \\ \beta\geq 0}} \|\alpha\|_1 + \min_{\mathbf{w},b,\mu} \lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{j=1}^n \mu_j \ - \ \sum_{j=1}^n \left(1/n - \alpha_j - \beta_j\right)\mu_j - \alpha_j y_j(\langle \mathbf{w}, \mathbf{x}_j\rangle - b)$$

after pulling out the -1 term and combining all μ_j terms. Note that technically these should be infinums and supremums since $\mathbf{w}$, b and μ are valued in the reals. If $1/n - \alpha_j - \beta_j \neq 0$, then the inside minima above will always be $-\infty$ as we can choose μ_j to be arbitrarily large or small. Similarly, unless $\sum_j \alpha_j y_j = 0$ we can choose b such that the minima is $-\infty$. If this is not the case, then we must have that $\beta_j = 1/n - \alpha_j$ and $\sum_j \alpha_j y_j = 0$. In this case we get the simpler optimization problem

$$\max_{\alpha} \min_{\mathbf{w}} \|\alpha\|_1 + \lambda\|\mathbf{w}\|_2^2 \ - \ \sum_{j=1}^n \alpha_j y_j\langle \mathbf{w}, \mathbf{x}_j\rangle$$

where $0 \leq \alpha_j \leq 1/n$ (so that $\beta_j \geq 0$ too) and $\sum_{j=1}^n \alpha_j y_j = 0$. Setting the gradient with respect to $\mathbf{w}$ to zero, we obtain

$$0 = \nabla\left[\|\alpha\|_1 + \lambda\|\mathbf{w}\|_2^2 \ - \ \sum_{j=1}^n \alpha_j y_j\langle \mathbf{w}, \mathbf{x}_j\rangle\right] = 2\lambda\mathbf{w} - \sum_{j=1}^n \alpha_j y_j x_j \implies \mathbf{w}* = \frac{1}{2\lambda}\sum_{j=1}^n \alpha_j y_j x_j.$$

After some simplification, substitution into our dual problem yields

$$\begin{aligned}&\max_{\alpha}\|\alpha\|_1 + \lambda\left\langle \frac{1}{2\lambda}\sum_{j=1}^n \alpha_j y_j x_j, \frac{1}{2\lambda}\sum_{j=1}^n \alpha_j y_j x_j\right\rangle - \sum_{j=1}^n \alpha_j y_j \left\langle \frac{1}{2\lambda}\sum_{k=1}^n \alpha_k y_k x_k, x_j\right\rangle \\ &= \max_{\alpha}\|\alpha\|_1 - \frac{1}{4\lambda}\sum_{j,k=1}^n \alpha_j\alpha_k y_j y_k\langle x_j, x_k\rangle,\end{aligned}$$

again subject to the constrains that $0 \le \alpha_j \le \frac{1}{n}$ and $\sum \alpha_j y_j = 0$. □

EXERCISE 2. The *weighted least squares problem* is a generalization of usual least squares:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\|\mathbf{WAx} - \mathbf{Wb}\|_2^2,$$

where $\mathbf{W}$ is a fixed diagonal matrix of positive weights

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix}.$$

Suppose $\mathbf{A}$ is an $n \times d$ full-rank matrix, and $n \ge d$. Suppose $\mathbf{b} = \mathbf{Ax}^* + \varepsilon$ for a fixed $\mathbf{x}^* \in \mathbb{R}^d$, and suppose the noise vector $\varepsilon = (\varepsilon_i)_{i=1}^n$ is a random vector whose components $\varepsilon_1, ..., \varepsilon_n$ are independent, mean-zero Gaussians with variances $\sigma_1^2, ..., \sigma_n^2$.

(a) Under this model, derive a natural choice of weights in the weighted least squares problem by considering the likelihood function.

(b) What is a closed-form expression for the solution to the weighted least squares problem? What if we add regularization to the problem and consider $\min_{x\in\mathbb{R}^d}\|\mathbf{WAx} - \mathbf{Wb}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ for $\lambda > 0$?

Proof: Recall that the likelihood function in this case is proportional to

$$\mathcal{L}(\mathbf{w} \mid (\mathbf{x}_i, y_i)) \propto \prod_{i=1}^n \frac{1}{\sigma}\exp\left(-\frac{(\mathbf{w}^T\mathbf{x}_i - y_i)^2}{2\sigma_i^2}\right).$$

Maximizing likelihood over $\mathbf{w} \in \mathbb{R}^n$ is then equivalent to minimizing $\exp\left(-\frac{(\mathbf{w}^T\mathbf{x}_i - y_i)^2}{2\sigma_i^2}\right)$. □

EXERCISE 3. Consider a two-layer neural network model of the form

$$f(\mathbf{W}, \mathbf{a}, x) = \sum_{r=1}^m a_r\sigma(\langle \mathbf{w}_r, \mathbf{x}\rangle)$$

with ReLU activatino function $\sigma(x) = \max(0, x)$. Fixing the second layer weights $\mathbf{a}$ and only training the first layer weights $\mathbf{W} = (\mathbf{w}_r)_{r=1}^m \in \mathbb{R}^{m\times d}$ via least squares loss, the optimization problem is

$$\min_{\mathbf{W}\in\mathbb{R}^{m\times d}} L(\mathbf{W}) = \min_{\mathbf{W}\in\mathbb{R}^{m\times d}} \frac{1}{n}\sum_{j=1}^n (f(\mathbf{W}, \mathbf{A}, \mathbf{x}_j) - y_j)^2.$$

(a) Derive an expression for the partial gradient $\frac{\partial L(\mathbf{W})}{\partial \mathbf{w}_r} \in \mathbb{R}^d$ where at $x = 0$ we set $\frac{d\sigma}{dx}\Big|_{x=0} = 1$.

EXERCISE 5. (Apologies that this is longer than a single page, but the problem warrants a description.) An **Ising Graph** consists of a finite set $G = [1..g] \subset \mathbb{N}$ together with parameters $h \in \mathbb{R}^G$ called the local biases and $J \in \mathbb{R}^{G \times G}$ called the interaction terms, where $J_{i,j} = J_{j,i}$ and $J_{i,i} = 0$. It comes equipped with a Hamiltonian function

$$H : S^G \longrightarrow \mathbb{R}, \qquad H(s) = \sum_{i \in G} h_i s_i \; + \; \frac{1}{2} \sum_{i,j \in G \times G} J_{i,j} s_i s_j,$$

where $S = \{-1, 1\}$. The space S^G is the set of all functions $G \longrightarrow S$ and is called the *spin space* of G, whereas an individual component $s(i)$ is called the *spin* of s at vertex s. This collection of data is a simple graph in the sense that G is a vertex set and J an edge matrix which does not allow self connections.

In physics, one is typically given an Ising Graph and wishes to understand its dynamics. In particular, one is interested in finding the local minima of the Hamiltonian function. This is known as the *Ising problem.* In the *reverse Ising problem*, one is instead given a collection of spins $X \subset S^G$ and wishes to find h and J such that the spin states in X are local minima.

Example 1.1. Set $G = 1, 2, 3$, $N = 1, 2$, and $M = 3$. Decompose S^G As $S^N \times S^M$ and let $X = \{(1, 1, 1), (1, -1, -1), (-1, 1, -1), (-1, -1, -1)\}$. Then for $h_3 = 1$, $J_{1,3} = -1$ and $J_{2,3} = -1$ each spin in X is a "local minima", in the sense that

$$\begin{aligned} H(1,1,1) &< H(1,1,-1), \qquad & H(1,-1,-1) &< H(1,-1,1), \\ H(-1,1,-1) &< H(-1,1,1), \qquad & H(-1,-1,-1) &< H(-1,-1,1). \end{aligned}$$

Notice that in the previous example we have abused the notion of a "local minima" slightly; we say that a pair $(s, t) \in S^N \times S^M$ is a local minimum if $H(s, t) < H(s, t')$ for all $t \neq t' \in S^M$. To avoid confusion, we will instead say that (s, t) is the minimum of the level $L_s := \{s\} \times S^M \subseteq S^N \times S^M$.

Notice also that the set X is precisely the truth table for AND with -1 in the place of 0:

s_1	s_2	s_3
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

We think of those spins in S^N as *input spins* and those in S^M as output spins. If we had some way to magically fix a spin $s \in S^N$ while allowing the Ising dynamics to affect the vertices in M, then we could build circuits out of this model. Indeed, if $f : S^N \longrightarrow S^M$ is some function and $X = \{(s, f(s)) \in S^N \times S^M \mid s \in S^N\}$ is the graph of f then finding an Ising circuit which models f is akin to solving the following optimization problem:

$$\text{find } \; h, J \quad \text{such that} \quad H(s, f(s)) < H(s, t) \quad \text{for all } s \in S^N \text{ and } t \neq f(s).$$

This is a linear programming problem with $2^{|N| \cdot (|M|-1)}$ constraints, but those constraints are often inconsistent and render the problem impossible. However, it becomes solvable if we add auxiliary spins which we ignore in the output.

Example 1.2. The XOR circuit defined by the function $f(s_1, s_2) = -s_1 \cdot s_2$ is infeasible. However, if we instead write $G = \{1, 2, 3, 4\}$, $N = 1, 2$, $M = \{3\}$ and $A = \{4\}$ and decompose the spin space $S^G = S^N \times S^M \times S^A$, then the function

$$\begin{aligned}(1,1) &\mapsto (-1,1)\\ (1,-1) &\mapsto (1,1)\\ (-1,1) &\mapsto (1,-1)\\ (-1,-1) &\mapsto (-1,1)\end{aligned}$$

is both feasible and recovers the XOR circuit when the spin s_4 is ignored.

Thus, the reverse Ising problem can be solved by adding some number of auxiliary spins A to the circuit. Unfortunately, this both increases the number of constraints exponentially (there are now $2^{|N|(|M|+|A|-1)}$ many) and introduces nonlinearity due to the added challenge of choosing an appropriate auxiliary spin for each input.

In this project, we wish to investigate the $N_1 \times N_2$ Ising multiply circuits $\mathrm{MUL}_{N_1 \times N_2}$ for $N_1, N_2 \leq 8$. Preliminary theoretical results suggest that embedding S^G into the higher dimensional spin space S^V where $V = G \cup \{(i,j) \in S^{G \times G} \mid i < j\}$ and clustering with respect to Hamming distance might yield a way to make informed guesses of feasible auxiliary values. We will test our proposed technique against available data for $N_1, N_2 \in \{2, 3\}$ and if successful, proceed to investigate $\mathrm{MUL}_{N_1 \times N_2}$ for higher values of N_1 and N_2.