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

Isaac Martin

Last compiled March 17, 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 $0 \leq \alpha_j \leq \frac{1}{n}$ and $\sum \alpha_j y_j = 0$. □

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

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{W}\mathbf{A}\mathbf{x} - \mathbf{W}\mathbf{b}\|_2^2,$$

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 $n \geq d$. Suppose $\mathbf{b} = \mathbf{A}\mathbf{x}^* + \varepsilon$ 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 $\sigma_1^2, ..., \sigma_n^2$.