Introduction to Machine Learning – COS 324: Midterm Exam

March 9, 2020

We discussed in class linear regression with the data matrix X being binary (either 0 or 1). In the problem below, you encounter linear regression where the entries of X are 0, 1, or 2.

You are given a dataset represented as a matrix $X \in \{0, 1, 2\}^{n \times d}$. That is, each entry of the matrix is either 0, 1, or 2. Each row of the matrix represents an example. The i^{th} example, denoted as $\mathbf{x}_i$, corresponds to the i^{th} row in the matrix and is thus a d dimensional vector over $\{0, 1, 2\}$. Each example is associated with a label $y_i \in \{-1, +1\}$. The labels are represented as a column vector denoted $\mathbf{y} \in \{-1, +1\}^n$.

You are provided with a vector of weights $\mathbf{w}$. You are barred from looking at individual entries of $\mathbf{w}$. Each element of $\mathbf{w}$ is in $[-2, -1] \cup [1, 2]$, namely, $\forall j : 1 \leq |w_j| \leq 2$. Let $z_i = y_i(\mathbf{w} \cdot \mathbf{x}_i)$ be the (signed) margin for example i. We denote

$$\overline{m} = \max_{i=1}^{n} z_i \quad \text{and} \quad \underline{m} = \min_{i=1}^{n} z_i \quad .$$

1. Construct an example such that $z_i = y_i(\mathbf{w} \cdot \mathbf{x}_i) = 0$. Namely, pick values for y_i and $\mathbf{x}_i$ so that $z_i = 0$.

   <u>Answer</u>: It suffices to choose $\mathbf{x}_i = \mathbf{0}$ (the zero vector). A more elegant solution is to take any of the vectors in the null space of $\mathbf{w}$, which amounts to any vector orthogonal to $\mathbf{w}$.

2. Bound $\overline{m}$ and $\underline{m}$, based on d. In words, given a particular value of d, what is the largest possible value of $\overline{m}$ and the smallest possible value of $\underline{m}$?

   <u>Answer</u>: Taking $\mathbf{x}_i = \mathbf{2} = (2, 2, \ldots, 2)$ and $\mathbf{w} = \mathbf{2}$ gives that $\mathbf{w} \cdot \mathbf{x} = 4d$. If we choose $y_i = 1$ we get $z_i = 4d$. Since we cannot further increase the norm of neither $\mathbf{x}_i$ (the $\|\cdot\|_1$ norm) nor $\mathbf{w}$ (the $\|\cdot\|_\infty$ norm) we have $\overline{m} = 4d$. Analogously, choosing $\mathbf{x}_j = \mathbf{2}$, $\mathbf{w} = \mathbf{2}$, but $y_j = -1$ yields $z_j = -4d$. For the same reasons it is the minimal value that can be attained and thus $\underline{m} = -4d$.

3. You are told that the dataset is linearly separable by $\mathbf{w}$: $\forall i : z_i = y_i(\mathbf{w} \cdot \mathbf{x}_i) > 0$. What are the bounds on $\overline{m}$ and $\underline{m}$ in light of this new information?

   Answer: The fact that the dataset is separable does not change the value of $\overline{m}$ which is positive. However, since $z_i > 0$ for all i, $\underline{m} = 0$ as there can exist $\mathbf{w}$ and $\mathbf{x}_i$ for which the signed margin is arbitrarily close to 0. In fact, merely two of the coordinates of $\mathbf{x}$ and $\mathbf{w}$ need not be zero. Assume that $\mathbf{x}_i = (1, 1, 0, \ldots, 0)$ and $y_i = 1$. If $\mathbf{w}$ ends up being $(1 + \epsilon, -1, w_3, \ldots, w_n)$ then $\mathbf{w} \cdot \mathbf{x}_i = 1 + \epsilon - 1 = \epsilon$. As ϵ can be arbitrarily small then the minimum (or formally the infimum) is 0. An answer of ϵ is also fine.

4. You further notice that $\forall i : 2 \leq \|\mathbf{x}_i\|_1 = \sum_j |X_{ij}| \leq 10$. In words, for each row $\mathbf{x}_i$ the sum of the absolute value of its entries is between 2 and 10.

   (a) What is the minimum and maximum number of non-zero entries in each row $\mathbf{x}_i$?
   Answer: Since $X_{ij} \in \{0, 1, 2\}$ when $\mathbf{x}_i = (2, 0, \ldots, 0)$ (or using the matrix form $X_{ij} = 2$ and $X_{ij} = 0$ for $j > 1$) has only *one* non-zero entry and thus $\|\mathbf{x}_i\|_1 = 2$. For the maximum we can set $X_{ij} = 1$ for $j = 1, \ldots, 10$ and $X_{ij} = 0$ for $j > 10$. Thus the maximum number of non-zero entries is *ten*.

   (b) Given the above condition, tighten the bounds on $\overline{m}$ and $\underline{m}$ when the data is linearly separable by $\mathbf{w}$.
   Answer: For the upper bound $\overline{m}$ we can use Holder's inequality,

   $$\mathbf{w} \cdot \mathbf{x} \leq \|\mathbf{w}\|_\infty \|\mathbf{x}\|_1 = 2 \times 10 = 20 \ .$$

   This implies that $\overline{m} = 20$. A direct derivation is as follows. Since $|w_j| \leq 2$ choose the vector $\mathbf{w} = \mathbf{2} = (2, 2, \ldots, 2)$, $\mathbf{x}_i = (1, 1, \ldots, 1, 0, \ldots, 0)$ (the first 10 coordinates are 1 and the rest are 0), and $y_i = 1$. Then, $z_i = y_i(\mathbf{w} \cdot \mathbf{x}_i) = 20$. We cannot obtain any larger value given the bound above by Holder's inequality. As for the lower bound $\underline{m}$, separability implies that $\underline{m}$ can be arbitrarily close to zero (see also previous answer) and thus $\underline{m} = 0$.

5. You are provided with the following empirical loss function,

   $$\mathcal{L}(\mathbf{w}) = \log \left(\sum_{i=1}^{n} e^{-y_i(\mathbf{w} \cdot \mathbf{x}_i)} \right) \ .$$

   We give below a simple derivation where the bounds are not tight but very close to the tightest bounds.

(a) Find an upper bound on $\mathcal{L}(\mathbf{w})$ as a function of $\underline{m}$, $\overline{m}$, and n. In words, using $\underline{m}$, $\overline{m}$, and n, express the largest possible value for $\mathcal{L}(\mathbf{w})$?

Answer: Since we have an inverse relationship between z_i and $\mathcal{L}(\mathbf{w})$ to obtain the upper bound use $z_i = \underline{m}$ which gives,

$$\mathcal{L}(\mathbf{w}) \leq \log\left(\sum_{i=1}^{n} e^{-\underline{m}}\right) = \log\left(ne^{-\underline{m}}\right) = \log(n) - \underline{m} \ .$$

(b) Find a lower bound on $\mathcal{L}(\mathbf{w})$ as a function of $\underline{m}$, $\overline{m}$, and n. In words, using $\underline{m}$, $\overline{m}$, and n, express the smallest possible value for $\mathcal{L}(\mathbf{w})$?

Answer: Analogously, due to the inverse relationship between z_i and $\mathcal{L}(\mathbf{w})$ to obtain the upper bound choose $z_i = \overline{m}$ as $-\overline{m}$ would be the smallest margin that can be attained. This gives,

$$\mathcal{L}(\mathbf{w}) \geq \log\left(\sum_{i=1}^{n} e^{-\overline{m}}\right) = \log\left(ne^{-\overline{m}}\right) = \log(n) - \overline{m} \ .$$

Summarizing, we get that, $\log(n) - \overline{m} \leq \mathcal{L}(\mathbf{w}) \leq \log(n) - \underline{m}$. These two bounds are meaningful and $\mathcal{L}(\mathbf{w})$ encompasses a bias term $\log(n)$. We can eliminate the bias term by modifying the loss to be normalized,

$$\mathcal{L}(\mathbf{w}) = \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{-y_i(\mathbf{w}\cdot\mathbf{x}_i)}\right) \ .$$

We then get that, $-\overline{m} \leq \mathcal{L}(\mathbf{w}) \leq -\underline{m}$.

(c) If the data is linearly separable by $\mathbf{w}$, what is the largest value $\mathcal{L}(\mathbf{w})$ can attain?

Answer: Since the data is linearly separable $\underline{m} = 0$ which implies $\mathcal{L}(\mathbf{w}) \leq \log(n)$ and for the normalized version we get $\mathcal{L}(\mathbf{w}) \leq 0$. In words, the loss $\mathcal{L}(\mathbf{w})$ is a softened version of $\max_i(-z_i) = -\min_i z_i$.

Optional Bonus Question Is $\mathcal{L}(\mathbf{w})$ convex in each w_j separately? That is, if you fix all the other entries of $\mathbf{w}$ except for w_j (so $\mathcal{L}(\mathbf{w})$ becomes a function of just w_j) is $\mathcal{L}(w_j)$ convex?

Answer: $\mathcal{L}(\mathbf{w})$ is convex in $\mathbf{w}$ and thus convex in each w_j separately. To prove convexity of $\mathcal{L}(\mathbf{w})$ in the general case we use the fact that it suffices to prove that $\mathcal{L}(\mathbf{u} + t\mathbf{v})$ is convex in t, where $\mathbf{u}$ and $\mathbf{v}$ are *arbitrary* vectors in $\mathbb{R}^d$.

Given $\mathbf{u}$ and $\mathbf{v}$ we can rewrite $\mathcal{L}$ as follows,

$$\mathcal{L}(t) = \log\left(\sum_{i=1}^{n} e^{b_i + a_i t}\right) \quad \text{where } b_i = -y_i(\mathbf{u}\cdot\mathbf{x}_i) \ , \ a_i = -y_i(\mathbf{v}\cdot\mathbf{x}_i) \ .$$

We show the convexity of $\mathcal{L}(t)$, which is a function from $\mathbb{R}$ to $\mathbb{R}$, by examining its second derivative. To do so we need to start with the first derivative. To simplify notation we denote, $Z = \sum_{i=1}^{n} e^{b_i + a_i t}$. Having introduced Z we have,

$$\mathcal{L}'(t) = \frac{\sum_i e^{b_i + a_i t} \, a_i}{Z} \ .$$

Last (kinda) notation we are going to introduce is, $q_i = (e^{b_i + a_i t})/Z$. Note that by definition $\sum_i q_i = 1$. We can now write $\mathcal{L}'(t) = \mathbb{E}_{\mathbf{q}}[\mathbf{a}] = \mathbf{q} \cdot \mathbf{a}$. Next we take the second derivative,

$$\begin{aligned} \mathcal{L}''(t) &= \frac{\sum_i e^{b_i + a_i t} \, a_i^2}{Z^2} - \frac{\sum_i (e^{b_i + a_i t} \, a_i)^2}{Z^2} \\ &= \sum_i q_i a_i^2 - \left(\sum_i q_i a_i\right)^2 \ = \ \mathbb{V}_{\mathbf{q}} \left[\mathbf{a}\right] \ . \end{aligned}$$

Here $\mathbb{V}_{\mathbf{q}}[\mathbf{v}]$ denotes the variance of the vector $\mathbf{v}$ w.r.t. the multinomial distribution $\mathbf{q}$. Since the variance of a random variable is non-negative we have $\mathcal{L}''(t) \geq 0$ as conjectured.