Subject:

Date: No:

Linear Predictors – Exercises

1. Show how to cast the ERM problem of linear regression with respect to the absolute value loss function, $l(h,(x,y)) = |h(x)-y|$, as a linear program; namely, show how to write the problem

$$\min_{w} \sum_{i=1}^{m} |\langle w, x_i \rangle - y_i|$$

as a linear program.

Hint: Start with proving that for any $c \in \mathbb{R}$

$$|c| = \min_{a \geq 0} a \quad \text{s.t. } c \leq a \text{ and } c \geq -a$$

We define $s = (s_1, \ldots, s_m)$. Minimizing the empirical risk is equivalent to minimizing the linear objective $\sum_{i=1}^{m} s_i$ under the following constraints:

$$(\forall i \in [m]) \quad w^T x_i - s_i \leq y_i \quad , \quad -w^T x_i - s_i \leq -y_i$$

Let $A \in \mathbb{R}^{2m \times (m+d)}$ be the matrix $A = [X - I_m \; ; -X - I_m]$, where $X_{i,:} = x$ for every $i \in [m]$. Let $v \in \mathbb{R}^{d+m}$ be the vector of variables $(w_1, \ldots, w_d, s_1, \ldots, s_m)$. Define $b \in \mathbb{R}^{2m}$ to be the vector $b = (y_1, \ldots, y_m, -y_1, \ldots, -y_m)^T$. Finally, let $c \in \mathbb{R}^{d+m}$ be the vector $c = (0_d ; 1_m)$. It follows that the optimization problem of minimizing the empirical risk can be expressed as the following LP:

$$\min \; c^T v \quad \text{s.t. } Av \leq b$$

3. Show that theorem 9.1 is tight in the following sense: For any positive integer m, there exist a vector $w^* \in \mathbb{R}^d$ (for some appropriate d) and a sequence of examples $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ such that the following hold:

- $R = \max_i \|x_i\| \leq 1$

Subject:

Date: No:

- $\|w^*\|^2 = m$, and for all $i \le m$, $y_i \langle x_i, w^* \rangle \ge 1$. Note that, using the notation in theorem 9.1, we therefore get

$$B = \min\{\|w\| : \forall i \in [m], y_i \langle w, x_i \rangle \ge 1\} \le \sqrt{m}$$

Thus, $(BR)^2 \le m$.

- When running the Perceptron on this sequence of examples it makes m updates before converting.

Hint: Choose $d=m$ and for every i choose $x_i = e_i$.

Let $d=m$, and for every $i \in [m]$, let $x_i = e_i$. Let us agree that $sign(0) = -1$. For $i = 1, \ldots, d$, let $y_i = 1$ be the label of x_i. Denote by $w^{(i)}$ the weight vector which is maintained by the Perceptron. A simple inductive argument shows that for every $i \in [d]$, $w_i = \sum_{j<i} e_j$. It follows that for every $i \in [d]$, $\langle w^{(i)}, x_i \rangle = 0$. Hence, all the instances $x_1, \ldots, x_d$ are misclassified (and then we obtain the vector $w = (1, \ldots, 1)$ which is consistent with $(x_1, \ldots, x_m)$. The vector $w^* = (1, \ldots, 1)$ satisfies the requirements listed in the question.

4. Given any number m, find an example of a sequence of labeled examples $((x_1, y_1) \ldots (x_m, y_m)) \in (\mathbb{R}^3 \times \{-1, 1\})^m$ on which the upper bound of theorem 9.1 equals m and the perceptron algorithm is bound to make m mistakes.

Hint: Set each x_i to be a third dimensional vector of the form (a, b, y_i), where $a^2 + b^2 = R^2 - 1$. Let w^* be the vector $(0, 0, 1)$. Now, go over the proof of the Perceptron's upper bound, see where we used inequalities ($\le$) rather than equalities ($=$), and figure out scenarios where the inequality actually holds with equality.

Subject:

Date: No:

Consider all positive examples of the form $(\alpha, \beta, 1)$, where $\alpha^2 + \beta^2 + 1 \leq R^2$. Observe that $w^* = (0,0,1)$ satisfies $y\langle w^*, x\rangle \geq 1$ for all such (x, y). We show a sequence of $\mathbb{R}^2$ examples on which the Perceptron makes $\mathbb{R}^2$ mistakes. The idea of the construction is to start with the examples $(\alpha_1, 0, 1)$ where $\alpha_1 = \sqrt{\mathbb{R}^2 - 1}$. Now, on round t let the new example be such that the following conditions hold:

a) $\alpha^2 + \beta^2 + 1 = \mathbb{R}^2$

b) $\langle w_t, (\alpha, \beta, 1)\rangle = 0$

As long as we can satisfy both conditions, the Perceptron will continues to err.

Observe that, by induction, $w^{(t-1)} = (a, b, t-1)$ for some scalars a, b. Observe also that $\|w_{t-1}\|^2 = (t-1)\mathbb{R}^2$ (this follows from the proof of the Perceptron's mistake bound, where inequalities hold with equality). That is $a^2 + b^2 + (t-1)^2 = (t-1)\mathbb{R}^2$. W.l.o.g. let us rotate $w^{(t-1)}$ w.r.t. the x axis so that it is of the form $(a, 0, t-1)$ and we have $a = \sqrt{(t-1)\mathbb{R}^2 - (t-1)^2}$. Choose $\alpha = -\frac{t-1}{a}$. Then for every β $\langle (a,0,t-1), (\alpha, \beta, 1)\rangle = 0$. We just need to verify that $\alpha^2 + 1 \leq R^2$, because if this is true then we can choose $\beta = \sqrt{R^2 - \alpha^2 - 1}$. Indeed,

$$\alpha^2 + 1 = \frac{(t-1)^2}{a^2} + 1 = \frac{(t-1)^2}{(t-1)\mathbb{R}^2 - (t-1)^2} + 1 = \frac{(t-1)\mathbb{R}^2}{(t-1)\mathbb{R}^2 - (t-1)^2}$$

$$= R^2 \frac{1}{\mathbb{R}^2 - (t-1)} \leq R^2$$

where the last inequality assumes $\mathbb{R}^2 \geq t$.

6. In this problem, we will get bounds on the VC-dimension of

Subject:

Date: No:

the class of (closed) balls in $\mathbb{R}^d$, that is

$$B_d = \{B_{v,r} : v \in \mathbb{R}^d, r > 0\}$$

where $B_{v,r}(x) = \begin{cases} 1 & \text{if } \|x - v\| \leq r \\ 0 & \text{otherwise} \end{cases}$

1. Consider the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$ defined by $\phi(x) = (x, \|x\|^2)$. Show that if $x_1, \ldots, x_m$ are shattered by B_d then $\phi(x_1), \ldots, \phi(x_m)$ are shattered by the class of the halfspaces in $\mathbb{R}^{d+1}$ (in this question we assume that $\text{sign}(0) = 1$.) What does this tell us about $VCdim(B_d)$?

2. Find a set of $d+1$ points in $\mathbb{R}^d$ that is shattered by B_d. Conclude that $d+1 \leq VCdim(B_d) \leq d+2$

1. Assume that $A = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$ is shattered by B_d. Then $\forall y = (y_1, \ldots, y_d) \in \{-1, 1\}^d$ there exists $B_{\mu,r} \in B$ s.t. for every i $B_{\mu,r}(x_i) = y_i$.

Hence, for the above μ and r, the following identity holds for every $i \in [m]$: $\text{sign}((2\mu_i - 1)^T (x_i ; \|x_i\|^2) - \|\mu\|^2 + r^2) = y_i$, where $;$ denotes vector concatenation. For each $i \in [m]$, let $\phi(x_i) = (x_i ; \|x_i\|^2)$. Define the halfspace $h \in \mathcal{Z}_{d+1}$ which corresponds to $w = (2\mu : -1)$, and $b = \|\mu\|^2 - r^2$. Equation above implies that for every $i \in [m]$ $h(x_i) = y_i$

All in all, if $A = \{x_1, \ldots, x_m\}$ is shattered by B, then $\phi(A) := \{\phi(x_1), \ldots, \phi(x_m)\}$ is shattered by $\mathcal{Z}$. We conclude that $d+2 = VCdim(\mathcal{Z}_{d+1}) \geq VC(B_d)$.

2. Consider the set C consisting of the unit vectors $e_1, \ldots, e_d$, and the origin 0. Let $A \subseteq C$. We show there exists a ball such that all the vectors in A are labeled positively, while

Subject:

Date: No:

the vectors in $C \setminus A$ are labeled negatively. We define the center $\mu = \sum_{e \in A} e$. Note that for every unit vector in A its distance of the origin to the center is $\sqrt{|A|-1}$. Also, for every unit vector outside A, its distance to the center is $\sqrt{|A|+1}$. Finally, the distance of the origin to the center is $\sqrt{A}$. Hence, if $0 \in A$, we will set $r = \sqrt{|A|-1}$, and if $0 \notin A$, we will set $r = \sqrt{|A|}$. We conclude that the set C is shattered by B_d. We showed $VCdim(B_d) \geq d+1$.