Solutions by **Andrew Lys** andrewlys(at)u.e.

1. Shattering Ellipses in $\mathbb{R}^2$

(a) The four points we will shatter are:

$$\{(-2,0),(0,-1),(0,1),(2,0)\}$$

(a) To get the behavior of 1 at a single point and 0 at the others, all we require are unit circles centered at each point.

(b) Two shatter any two points, draw any ellipse where they are the foci. As the eccentricity increases, all points not on the line spanned by the foci will eventually be excluded, and since in the above points, no 3 are collinear, we satisfy these cases.

(c) For any 3 points, the third point may be excluded by a circle with center low/high enough and radius large enough.

(d) For 4 points the circle with radius 2, centered at 0 suffices.

(b)

$$\frac{(x_1-c_1)^2}{a_1^2}+\frac{(x_2-c_2)^2}{a_2^2}\le r$$
$$a_2^2x_1^2+a_1^2x_2^2-2a_2^2c_1x_1-2a_1^2c_2x_2+c_1^2a_2^2+c_2^2a_1^2\le ra_1^2a_2^2$$

It is clear that the x_1^2, x_2^2, x_1, x_2, and constant terms are all linearly independent. Therefore, the feature map is as follows:

$$\phi(E)=(1,x_1,x_2,x_1^2,x_2^2)^\intercal$$

(c) As shown in class the VC dimension of this class, namely

$$\mathcal{H}_\phi=\{\text{sign}(\langle w,\phi(x)\rangle-y):w\in\mathbb{R}^5,y\in\mathbb{R}\}$$

is 6.

From part (a), we showed that the VC dimension of the ellipse class in $\mathbb{R}^2$ is at least 5.

2. Shattering Sparse Linear Predictors

(a) Suppose $d=2^{n-1}$. Let b_i be an element of $\{x:x\in\{-1,+1\}^n$ and there are at least as many positive entries as negative entries$\}$. Clearly the number of b_i's is 2^{n-1}. Then consider the following 2 dimensional array given by:

$$\begin{bmatrix} | & | & \cdots & | \\ b_1 & b_2 & \cdots & b_d \\ | & | & \cdots & | \end{bmatrix}$$

Let x_i be the i'th row of this array. Then, every behavior of signs, where there are at least as many positive signs as negative signs of $x_1,\ldots,x_n$ is encoded in some b_i. Then, we choose $w=e_i$, where $\{e_i\}$ are the standard basis vectors and our desired behavior is given.

If the behavior we desire has at least as many negative signs as positive signs, then we negate the behavior, and choose b_i as above, and return $w=-e_i$. Thus, we actually manage to shatter a set of size $\log(d)+1$.

(b) To show we can shatter $\max\{k,\log(d)\}$ points with $\mathcal{H}_k$, all we need to do is shatter k points, and shatter $\log(d)$ points. First we shatter k points.

Let $x_i=e_i$, for $i=1,\ldots,k$, and e_i is the standard basis vector with a 1 in the i'th coordinate, and zeroes everywhere else. Then, for any

$$(y_1,y_2,\ldots,y_k)\in\{\pm1\}^k$$

We simply let

$$w=(y_1,y_2,\ldots,y_k,0,\ldots,0)$$

then,

$$\text{sign}(\langle w,x_i\rangle)=\text{sign}(y_i)=y_i$$

and we have successfully shattered k points. To shatter $\log(d)$ points, we assume that $d = 2^n$, and construct the same x_i as in part (a). For any $(y_1, \ldots, y_n)$, we select the b_i corresponding to this behavior, and we let

$$w = e_i + \sum_{j \in J : |J| = k-1, i \notin J} \varepsilon_j e_j$$

Where ε_j are small. Effectively, if ε_j are small enough, then the behavior of $\langle w, x_p \rangle$ will be the same as $\langle e_i, x_p \rangle$. In fact, in order to make sure that the signs are the same, all we have to do is set

$$0 < |\varepsilon_j| < \frac{1}{k-1}$$

Then, we have:

$$\langle w, x_i \rangle = y_i + \sum_{j \in J} x_j \varepsilon_j$$

$$y_i - \sum_{j \in J} |\varepsilon_j| \leq \langle w, x_i \rangle \leq y_i + \sum_{j \in J} |\varepsilon_j|$$

$$y_i - 1 < y_i - \sum_{j \in J} |\varepsilon_j|$$

$$y_i + 1 > y_i + \sum_{j \in J} |\varepsilon_j|$$

Since $y_i = \pm 1$,

$$y_i - 1 < \langle w, x_i \rangle < y_i + 1$$

If $y_i = 1$, the we still have that $\langle w, x_i \rangle$ is positive, and if $y_i = -1$, we still have that $\langle w, x_i \rangle$ is negative. Thus

$$\text{sign}(\langle w, x_i \rangle) = \text{sign}(y_i) = y_i$$

and we have successfully shattered $\log(d)$ points.

3. Realizable Online-to-Bach and Leave-One-Out Cross-Validation

(a)

$$|\{i : \mathcal{F}(S_{-i})(x_i) \neq y_i\}| = \sum_{i=1}^{m} \mathbb{1}\{\mathcal{F}(S_{-i})(x_i) \neq y_i\}$$

$$\begin{aligned} E_{S \sim \mathcal{D}^m}[\mathbb{1}\{\mathcal{F}(S_{-i})(x_i)\}] &= E_{S \sim \mathcal{D}^{m-1}, (x,y) \sim \mathcal{D}}[\mathbb{1}\{\mathcal{F}(S)(x) \neq y\}] \\ &= E_{S \sim \mathcal{D}^{m-1}}[E_{(x,y) \sim \mathcal{D}}[\mathbb{1}\{\mathcal{F}(S)(x) \neq y | S\}]] \\ &= E_{S \sim \mathcal{D}^{m-1}}[L_{\mathcal{D}}(\mathcal{F}(S))] \end{aligned}$$

(b) By definition, $\tilde{\mathcal{A}}$ runs an iteration only if there is some $(x_i, y_i) \in S$ such that $h(x_i) \neq y_i$. Assume WLOG, that the method of $\tilde{\mathcal{A}}$ checking for this is simply running through the lowest to highest index, and it stops as soon as it hits the first mistake, meaning $\tilde{\mathcal{A}}$ makes exactly 1 mistake per iteration. Thus, if S is realizable by $\mathcal{H}$, then we make at most M iterations.

Since $\mathcal{D}$ is realizable by $\mathcal{H}$, there exists h s.t.

$$P_{\mathcal{D}}[h(x) \neq y] = 0$$

Thus, given S, with probability one, $h(x_i) = y_i$ for every $(x_i, y_i) \in S$, so S is realizable by $\mathcal{H}$.

Therefore, $T \leq M$.

(c) Let $|S| = m + 1$. Let K be the set of indices that get selected to train $\tilde{\mathcal{A}}(S)$. $|K| \leq M$, since we make at most M mistakes.

Note then that

$$\tilde{\mathcal{A}}(S) = \mathcal{A}\{(x_i, y_i) : i \in K\}$$

Let $j \notin K$. Further assume that, during the training process, even if $h(x_j) \neq y_j$, and it isn't picked as a result of the deterministic rule, its absence does not impact the deterministic rule's selection. For a concrete example, suppose that the deterministic rule picks the mistake with the lowest index.

Then,

$$\tilde{\mathcal{A}}(S_{-j}) = \mathcal{A}\{(x_i, y_i) : i \in K\} = \tilde{A}(S)$$

Therefore,

$$\tilde{\mathcal{A}}(S_{-j})(x_j) = \tilde{A}(S)(x_j) = y_j$$

Thus, we have:

$$\sum_{i=1}^{m+1} \mathbb{1}\{\mathcal{A}(S_{-i})(x_i) \neq y_i\} \leq |K| \leq M$$

So

$$LOOCV_S(\tilde{\mathcal{A}}) \leq \frac{M}{m+1}$$

(d)

$$\begin{aligned} E_{S\sim\mathcal{D}^m}[L_D(\tilde{\mathcal{A}}(S))] &= E_{S\sim\mathcal{D}^{m+1}}[LOOCV_S(\tilde{A})] \\ &\leq \frac{M}{m+1} < \varepsilon \\ &\iff \frac{M}{\varepsilon} < m+1 \end{aligned}$$