TTIC 31020: Introduction to Machine Learning

Problem Set 2

Hung Le Tran

15 Jan 2024

Problem 2.1 (Problem 1)

(a) Since $\rho(x, x_i)$ is shift invariant, $\int_x K(x, x_i)\mathrm{d}x$ is the same constant for all x_i. Let us denote that $C \coloneqq \int_x K(x, x_i)\mathrm{d}x$.

In the binary classification context, then, let $m_+ \coloneqq |\{i \mid y_i = 1\}|, m_- \coloneqq |\{i \mid y_i = -1\}| = m - m_+$. Then

$$\begin{aligned} 1 &= \int_x \hat{f}(x \mid Y = 1) \\ &= \int_x Z_1 \sum_{y_i = y} K(x, x_i) \\ &= Z_1 \int_x \sum_{y_i = y} K(x, x_i) \\ &= Z_1 m_+ C \\ \Rightarrow Z_1 &= \frac{1}{m_+ C} \end{aligned}$$

Similarly, $Z_{-1} = \frac{1}{m_- C}$. Also, $\hat{p}(y = 1) = \frac{m_+}{m}, \hat{p}(y = -1) = \frac{m_-}{m}$.

It follows that

$$\begin{aligned} \hat{f}(y = +1 \mid x) &= \frac{\hat{f}(x \mid y = +1)\hat{p}(y = +1)}{\hat{f}(x \mid y = +1)\hat{p}(y = +1) + \hat{f}(x \mid y = -1)\hat{p}(y = -1)} \\ &= \frac{\left(Z_1 \sum_{y_i = 1} K(x, x_i)\right) \frac{m_+}{m}}{\left(Z_1 \sum_{y_i = 1} K(x, x_i)\right) \frac{m_+}{m} + \left(Z_{-1} \sum_{y_i = -1} K(x, x_i)\right) \frac{m_-}{m}} \\ &= \frac{\sum_{y_i = 1} K(x, x_i)}{\sum_{y_i = 1} K(x, x_i) + \sum_{y_i = -1} K(x, x_i)} \end{aligned}$$

Similarly,

$$\hat{f}(y=-1 \mid x)=\frac{\sum_{y_i=-1} K(x,x_i)}{\sum_{y_i=1} K(x,x_i)+\sum_{y_i=-1} K(x,x_i)}$$

The Parzen Predictor, as the Bayes Optimal Predictor for $\hat{\mathcal{D}}$, predicts 1 when the first conditional density is greater than the second, and -1 otherwise. Therefore,

$$\begin{aligned} h(x) &= \text{sign}\left(\sum_{y_i=1} K(x,x_i) - \sum_{y_i=-1} K(x,x_i)\right) \\ &= \text{sign}\left(\sum_{i=1}^{m} y_i K(x,x_i)\right) \end{aligned}$$

as required.

(b) In the limit $\sigma \to \infty$,

$$K(x,x') = e^{-\rho(x,x')^2/\sigma^2} \xrightarrow{\sigma\to\infty} 1$$

therefore

$$h(x) = \text{sign}\left(\sum_{i=1}^{m} y_i\right)$$

taking the average of all y_i.

(c) In the limit $\sigma \to 0$, if $0 \le \rho(x,x_i) \le \rho(x,x_j) - \varepsilon$ then $K(x,x_i) \gg K(x,x_j)$ as $\sigma \to 0$. It follows that as $\sigma \to 0$,

$$\begin{aligned} h(x) &= \text{sign}\left(\sum_{i=1}^{m} y_i K(x,x_i)\right) \\ &\overset{\sigma\to 0}{=} \text{sign}\left(\sum_{x_i \text{ closest to } x} y_i K(x,x_i)\right) \\ &= \text{sign}\left(\sum_{x_i \text{ closest to } x} y_i\right) \end{aligned}$$

(d) The Parzen predictor also predicts the label for x using the labels of the $x_i's$ that are closets to x, in this case, summing up the labels of y_i's. The sign of the sum will therefore take the sign of the sign of the majority of the labels; the Parzen predictor (as $\sigma \to 0$) breaks ties by following the majority of closest points.

Problem 2.2 (Problem 2)

(a) We're given

$$\mathbb{P}_{\mathcal{D}}(Y=+1 \mid x) = \begin{cases} 0.8 & \text{if } x \ge 0 \\ 0.2 & \text{if } x < 0 \end{cases}$$

and it follows that the bayes Optimal Predictor is

$$h_{Bayes}(\mathcal{D})(x) = \text{sign}\left(\eta_{\mathcal{D}}(x) - 0.5\right) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}$$

The Bayes Error is

$$\begin{aligned} L_{\mathcal{D}}(h_{Bayes}(\mathcal{D})) &= \mathbb{P}_{(x,y)\sim\mathcal{D}}[h_{Bayes}(\mathcal{D})(x) \neq y] \\ &= \frac{1}{2}0.2 + \frac{1}{2}0.2 = 0.2 \end{aligned}$$

(b) As $m \to \infty$, the nearest neighbor of $x_{neighbor}$ of most x in S would have the same sign as x. When $x \geq 0$, $x_{neighbor}$ has the label distribution

$$\mathbb{P}(y_{neighbor} = +1 \mid x \geq 0) = 0.8, \mathbb{P}(y_{neighbor} = -1 \mid x \geq 0) = 0.2$$

and thus corresponding labeling probabilities for the label of x itself. Therefore

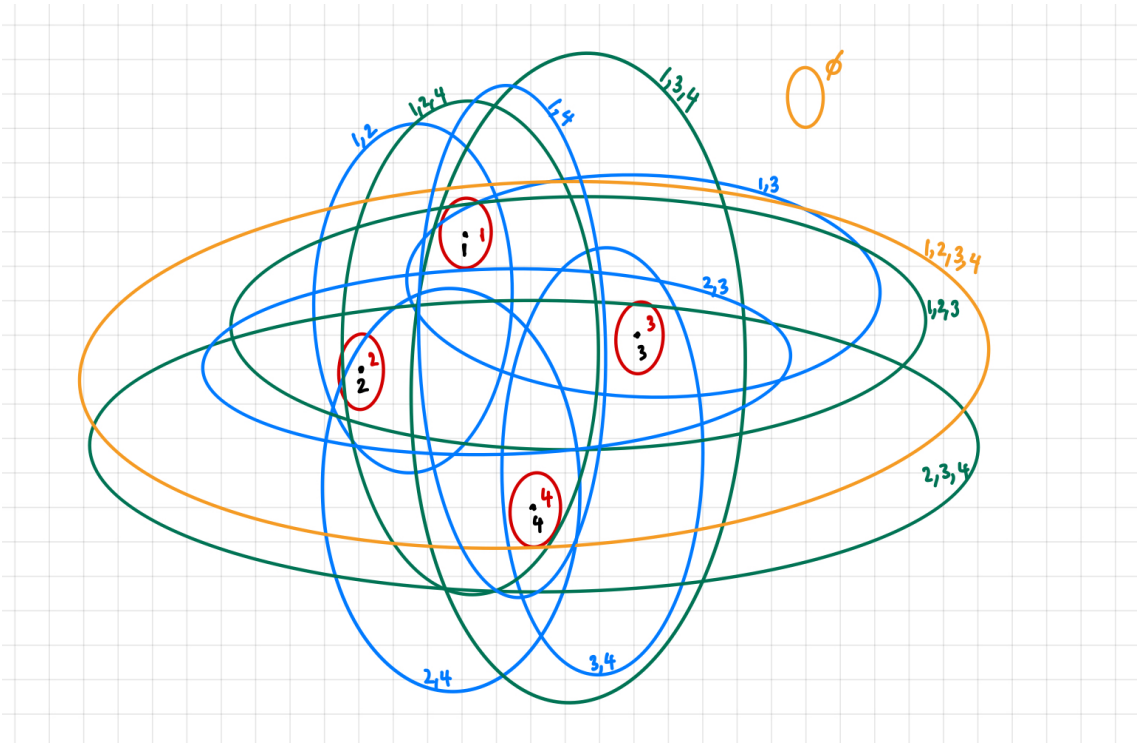
$$\begin{aligned} L_{x\geq 0,(x,y)\sim\mathcal{D}}(h_m) &= \mathbb{P}(y_{neighbor} = +1, y = -1 \mid x \geq 0) + \mathbb{P}(y_{neighbor} = -1, y = +1 \mid x \geq 0) \\ &= 0.8 \times 0.2 + 0.2 \times 0.8 = 0.32 \end{aligned}$$

and similarly, $L_{x<0,(x,y)\sim\mathcal{D}}(h_m) = 0.32$.

It follows that $L_{(x,y)\sim\mathcal{D}}(h_m) = 0.32 \times \frac{1}{2} + 0.32 \times \frac{1}{2} = 0.32$

Problem 2.3 (Problem 3)

(a)



(b) Using the feature map $\varphi : \mathbb{R}^2 \to \mathbb{R}^5, (x_1, x_2) \mapsto (x_1^2, x_1, x_2^2, x_2, 1)$ then

$$\mathbb{1}\left[\frac{(x_1 - c_1)^2}{a_1^2} + \frac{(x_2 - c_2)^2}{a_2^2} \leq r\right] = \mathbb{1}\left[\langle \varphi(x_1, x_2), \left(\frac{1}{a_1^2}, \frac{-2c_1}{a_1^2}, \frac{1}{a_2^2}, \frac{-2c_2}{a_2^2}, \frac{c_1^2}{a_1^2} + \frac{c_2^2}{a_2^2} - r\right)\rangle \leq 0\right]$$

Thus

$$\mathcal{H} = \left\{\mathbb{1}\left[\langle \varphi(x_1, x_2), \left(\frac{1}{a_1^2}, \frac{-2c_1}{a_1^2}, \frac{1}{a_2^2}, \frac{-2c_2}{a_2^2}, \frac{c_1^2}{a_1^2} + \frac{c_2^2}{a_2^2} - r\right)\rangle \leq 0\right] : c_1, c_2, a_1, a_2, r \in \mathbb{R}; a_1 \neq a_2\right\}$$

is a hypothesis class of linear predictors, and therefore a subset of H_φ^{linear}, the hypothesis class of all linear predictors in $\varphi(x_1, x_2)$:

$$\mathcal{H}_\varphi^{linear} = \{\mathbb{1}\left[\langle \varphi(x_1, x_2), w\rangle \leq 0\right] : w \in \mathbb{R}^5\}$$

(c) Since we have represented in (b) that $\mathcal{H}$ is a subset of $\mathcal{H}_\varphi^{linear}$,

$$VCDim(\mathcal{H}) \leq VCDim(\mathcal{H}_\varphi^{linear}) = 5$$

We have shattered 4 points in (a), which demonstrates that

$$VCDim(\mathcal{H}) \geq 4$$

There is a difference of $5 - 4 = 1$ between the VCDim bounds. The gap can be explained by that in the linear representation of $\mathcal{H}$, $w[0] = \frac{1}{a_1^2}, w[2] = \frac{1}{a_2^2} > 0$ are forced to be positive; in other words, the predictors with negative values in either $w[0]$ or $w[2]$ are included in $\mathcal{H}_\varphi^{linear}$ but are not included in $\mathcal{H}$, which potentially decreases the VCDim.

Problem 2.4 (Problem 4)

(a) We write out explicitly the definition for $\mathcal{H}_1$:

$$\mathcal{H}_1 = \{y = \text{sign}\left(\langle w, x\rangle\right) : w \in \mathbb{R}^d, \|w\|_0 = 1\}$$

$\|w\|_0 = 1$ means that w can have 1 non-zero coordinate, say, i-th coordinate. It follows that

$$\mathcal{H}_1 = \{y = \text{sign}\left(\alpha x[i]\right) : i \in [d], \alpha \in \mathbb{R}\backslash\{0\}\}$$

And therefore for a set of points to be classified into the same label, all the points must share at least 1 index i.

We can therefore construct a set of $\log_2 d$ points, A, in the following manner. Initialize all points to have (-1) as their default coordinate for all coordinates. Within $[\log_2 d] (:= \{1, \ldots, \log_2 d\})$, there are $2^{\log_2 d} = d$ subsets of indices. We then iterate through the subsets in some, but fixed, order:

```
subsets = all_subsets(A)
for i in range(d):
    subset = subsets[i]
    for point in subset:
        point[i] = 1
```

i.e. assigning 1 to be the value of the i-th index of points in the i-th subset.

Then, with this A, for any $(y_1, \ldots, y_{\log_2 d}) \in \{\pm 1\}^{\log_2 d}$, let

$$B = \{j : y_j = +1\}$$

is a subset of $[\log_2 d]$. Let its position, in the fixed order above (0-index), be i_B.

Then the predictor $h(x) = \text{sign}\,(x[i_B])$ would give us $h(x_i) = y_i \;\forall\; 1 \leq i \leq |A|$, since $x_k[i_B] > 0 \Leftrightarrow x_k \in B \Leftrightarrow y_k = +1$.

We have thus shattered $\log_2 d$ points using $\mathcal{H}_1$.

(b) We write out $\mathcal{H}_k$:

$$\mathcal{H}_k = \{y = \text{sign}\,(\langle w, x\rangle) : w \in \mathbb{R}^d, \|w\|_0 = k\}$$

If $k < \log d$ then we have completed this in (a), by using the same method as $\mathcal{H}_1$ for any index of w and setting the remaining $(k-1)$ non-zero coordinates of w to be infinitesimally small so that their contribution to $\text{sign}\,(\langle w, x\rangle)$ is negligible, so that same reasoning works.

When $k \geq \log d$, then $\max\{k, \log d\} = k$. Note that $k \leq d$, since $w \in \mathbb{R}^d$.

Our job now is to shatter k points with $\mathcal{H}_k$. Define x_j to have all (-1) except in the j-th coordinate, and $2d$ for the j-th coordinate, for $1 \leq j \leq k$.

Then for any $(y_1, \ldots, y_k) \in \{\pm 1\}^k$, let

$$B = \{j : y_j = +1\}$$

Then we construct $\hat{w}$ by assigning 1 in coordinates that are in B, and $\varepsilon \ll 1$ or 0 everywhere else (fill ε in $(k - |B|)$ coordinates to get a total of k non-zero coordinates for $\hat{w}$ (so that $\|\hat{w}\|_0 = k$), then assign 0 to remaining coordinates; it does not matter which coordinates we choose for ε and which for 0).

Then $\hat{h} = \text{sign}\,(\langle \hat{w}, x\rangle)$ would give predictions:

$$\langle \hat{w}, x_j\rangle \geq 2d \times 1 + (-1) \times (|B| - 1) + (-1) \times \varepsilon \times (k - |B|) > 0 \text{ (since } |B| \leq d), \text{ for } j \in B$$
$$\Rightarrow \hat{h}(x_j) = 1, \text{ for } j \in B$$
$$\langle \hat{w}, x_j\rangle \leq 2d \times \varepsilon + (-1) \times |B| < 0, \text{ for } j \notin B$$
$$\Rightarrow \hat{h}(x_j) = -1, \text{ for } j \notin B$$

as required.

We have thus shattered $\max\{k, \log d\}$ points with $\mathcal{H}_k$.