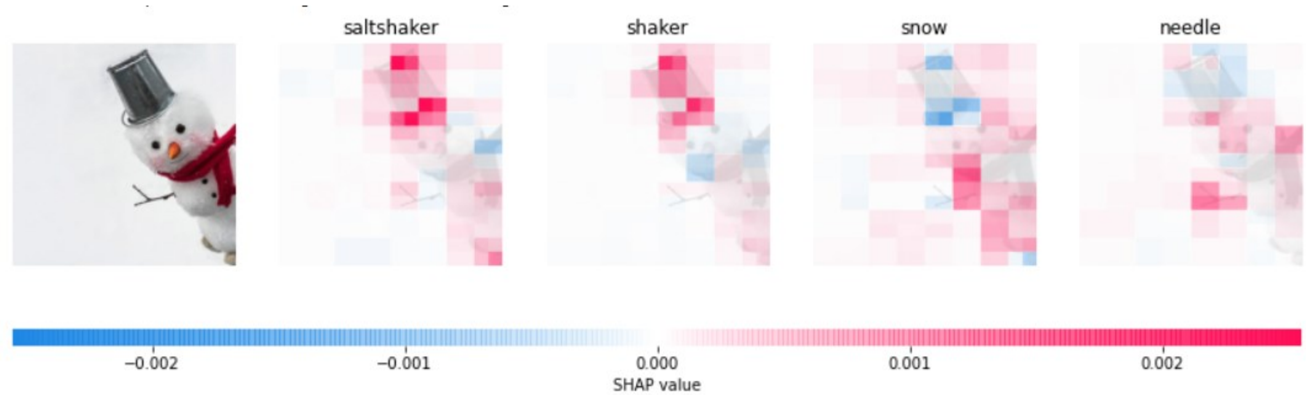
Mahdi Abdollah Chalaki   ID: 1795946   #HW 2

Q1: (a) As it could be seen from the corresponding image, the faces are the most important part. The pixels on Obama's face in the class related to the president is red. Also, Bush's face on ex-president class seems to be the most important part as the pixels of that area are red.

(b) Probably, the curves of the mask under the person's eyes and her eyebrows form a shap that looks like a Domino mask. Also, shape of the lower part of her face has not changed after wearing the mask. The color of this mask might be another reason.




```
Prediction: saltshaker, salt_shaker
```

... another reason.

(C) As the body of the snowman is white and its hat is silver, the model misidentified it as a saltshaker. The pixels of the hat are red, so they played an important role and the hat is like a saltshaker head.

(d) The model suffers from gender bias, and has detected the Chancellor as as first lady. It is probably due to the fact that the number of the women in power is by far smaller than men, and the model has been trained on a dataset that was imballanced in this manner.

(e) when we want to train on a dataset, not only we should pay attention to the class balance, but we should also consider the diversity of example s in each group. For example, in the picture of our intructor which is labeled as eccentric and geeky, the glasses are important features. It means that in the data set, these classes were full of pictures of people with glasses, so it became a feature to determine the labels.
Another example could be the number of each gender in different job labels, so the model won't misclassify based on the gender of each person.

Q2: (a) i- True

The learnt prediction function is: $\text{sign}(w^T \phi(x_i))$

$$= \text{sign}\left(\sum y_i \alpha_i \phi(x_i) \phi(x)\right)$$

so if α_i for i-th instance is zero, its contribution to the sum is 0

ii- False

Slack variables larger than zero means that there are some points on the other side of the +1 (or -1) plane. So, there is no guarantee that we don't have a point that will even penetrate the decison boundary and in that case, it would be misclassified.

iii- True

In this case, we will chose the closest points of the groups to be the support vectors. The dataset in linearly separable, so this distance exists and there won't be other points in between.

iV- False

As we might have some misclassified points, it should be $\min_i \left| \frac{y_i w^T \varphi(x_i)}{\|w\|} \right|$

Proof: $x \to$ nearest point

$\bar{x} \to$ on the separation line

$$\left.\begin{array}{l} \Rightarrow x = \bar{x} + \lambda w \\ w^T \bar{x} = 0 \end{array}\right\} \Rightarrow w^T x = 0 + \lambda w^T w \to \lambda = \frac{w^T x}{w^T w} \Rightarrow |\lambda| = \frac{|w^T x|}{\|w\|^2}$$

$$|\lambda| . \|w\| = \frac{|w^T x|}{\|w\|} \xrightarrow[x \to \varphi(x)]{\min} \min_i \frac{|w^T \varphi(x)|}{\|w\|} = \min_i \left| y_i \frac{w^T \varphi(x)}{\|w\|} \right|$$

$\frac{}{\|w\|} \quad x \to \varphi(x) \quad i \quad \frac{}{\|w\|} = \min_i \left| \frac{}{\|w\|} \right|$

v- False

In this case, the margins change and some points might get on the other side of these lines. So, we will have some new ξs and this will change the value of our function $\frac{w^T w}{4} + \sum_{i=1}^{n} C \xi_i$. So, the w might change and so will the distance we calculated in the previous part.

(b)

i- The left one belongs to second order kernel (it has a convex shape)
The middle one belongs to the linear kernel.
The right one belongs to the fifth order kernel.
As it could be seen from the pictures, using a higher order polynomial lead to having more support vectors and it won't perform well on new data points.

ii - The left one $\to C = 200$
The right one $\to C = 2$

Increasing C means more penalty on ξs and hence, less number of support vectors. (which is the figure A case)

A small C leads to overfitness. It captures all datapoints and won't generalize to test points.

iii - The RBF kernel formula is $\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) = \exp(-\gamma \|x-z\|^2)$; $\gamma = \frac{1}{2\sigma^2}$

The parameter γ defines the influence of a datapoint; If it is low, then σ^2 will be high and the influence will rich far distances. So, it is proportional to the inverse of the radius of influence of samples. So we have: The left figure $\rightarrow$ $\gamma = 0.0025$
The middle figure $\rightarrow$ $\gamma = 400$
The right figure $\rightarrow$ $\gamma = 1$

A large gamma leads to overfitness, as the points should be closer together.

(C) i- $y_i f(x_i) = y_i (w^T \varphi(x_i)) = y_i \sum_i \alpha_i y_i e^{-\gamma\|x-x_i\|^2} = y_i \sum_P \alpha_i y_i e^{-\gamma\|x-x_i\|^2} + y_i \sum_D \alpha_i y_i e^{-\gamma\|x-x_i\|^2}$

$$= + \sum_P \alpha_i e^{-\gamma\|x-x_i\|^2} - \sum_D \alpha_i e^{-\gamma\|x-x_i\|^2}$$

We also have: $\begin{cases} \|x-x_i\| \leq S_1 & \text{for } x \in P \rightarrow e^{-\gamma\|x-x_i\|^2} \geq e^{-\gamma S_1^2} \\ \|x-x_i\| \geq S_2 & \text{for } x \in D \rightarrow e^{-\gamma\|x-x_i\|^2} \leq e^{-\gamma S_2^2} \\ -y_i^2 \geq -1 \end{cases}$

$$\Rightarrow y_j f(x_j) - y_j^2 \geq \sum_{j\in P} \alpha_j e^{-\gamma S_1^2} - \sum_{j\in D} \alpha_j e^{-\gamma S_2^2} - 1 \quad \checkmark$$

ii - From the right hand side of the above equation, we have:

$$\sum_{j\in P} \alpha_j e^{-\gamma S_1^2} - \sum_{j\in D} \alpha_j e^{-\gamma S_2^2} - 1 + \left(\sum_{j\in P} \alpha_j e^{-\gamma S_2^2} - \sum_{j\in P} \alpha_j e^{-\gamma S_2^2}\right)$$

$$= \sum_{j\in P} \alpha_j \left(e^{-\gamma S_1^2} - e^{-\gamma S_2^2}\right) + \left(\sum_{j\in P} \alpha_j - \sum_{j\in D} \alpha_j\right) e^{-\gamma S_2^2} - 1$$

① we have $\sum_{i=1}^{m} \alpha_i y_i = 0 \rightarrow \sum_{P} \alpha_j \underset{+1\,(-1)}{\underset{\downarrow}{y_j}} + \sum_{D} \alpha_j \underset{-1\,(1)}{\underset{\downarrow}{y_j}} = 0 \Rightarrow \sum_{P} \alpha_j - \sum_{D} \alpha_j = 0$

② $0 \leq \alpha_j \leq C \longrightarrow \sum_{P} \alpha_j \leq |P| . C \leq mC \rightarrow -\sum_{P} \alpha_j \geq -mC$ ⊛

③ for convex functions we have: $f(x) - f(y) \geq f'(y)(x-y) \longrightarrow \begin{cases} x := s_1^2 & \rightarrow f(x) = e^{-\gamma x} \\ y := s_2^2 & \rightarrow f(y) = e^{-\gamma y} \rightarrow f'(y) = -\gamma e^{-\gamma y} \end{cases}$

$\Rightarrow e^{-\gamma x} - e^{-\gamma y} \geq -\gamma e^{-\gamma y}(x-y) \Rightarrow e^{-\gamma s_1^2} - e^{-\gamma s_2^2} \geq -\gamma e^{-\gamma s_2^2}(s_1^2 - s_2^2)$

$\Rightarrow y_i(f(x_i) - y_i) \geq \sum_{P} \alpha_j (e^{-\gamma s_1^2} - e^{-\gamma s_2^2}) - 1 \geq \sum_{P} \alpha_j (-\gamma e^{-\gamma s_2^2}(s_1^2 - s_2^2)) - 1$

we also have $\gamma s_2^2 \geq 0 \Rightarrow -e^{-\gamma s_2^2} \geq -1 \rightarrow y_i(f(x_i) - y_i) \geq \sum_{P} \alpha_j (-\gamma(s_1^2 - s_2^2)) - 1$

$\underline{⊛} \rightarrow y_i(f(x_i) - y_i) \geq -mC\gamma(s_1^2 - s_2^2) - 1 = Cm\gamma(s_2^2 - s_1^2) - 1$ ✓

iii - For $y_i = 1 \rightarrow \begin{cases} f(x_i) > 0 & \Leftrightarrow \text{Correctly classified} \\ \underbrace{f(x_i) > -1}_{\text{I}} & \Leftrightarrow \text{not misclassified out side the margin} \end{cases}$

For $y_i = -1 \Rightarrow \begin{cases} f(x_i) < 0 & \Leftrightarrow \text{Correctly classified} \\ \underbrace{f(x_i) < 1}_{\text{II}} & \Leftrightarrow \text{not misclassified outside the margin} \end{cases}$

$\left.\begin{array}{l} \text{I} \rightarrow y_i f(x_i) > -1 \\ \text{II} \rightarrow y_i f(x_i) > -1 \end{array}\right\} \Rightarrow$ The condition for not being misclassified outside of the margin is $y_i f(x_i) > -1$

$\Rightarrow y_i f(x_i) - y_i^2 > -1 - 1 = -2$

$\Rightarrow Cm\gamma(s_2^2 - s_1^2) - 1 > -2 \xrightarrow{s_1 > s_2} C\gamma < \frac{-1}{m(s_2^2 - s_1^2)}$

$\rightarrow C\gamma < \frac{1}{m(s_1^2 - s_2^2)}$

Q4: a) i) Here, we define a loss function $J(x;w)$ such that maximizing this function leads to a weak performance by the classifier. However, we also don't want to change the input hugely; so we should constraint the perturbation.

By using a first-order approximation, we can generalize this loss function:

$$J(x;w) = J(x_0 + r; w) \simeq J(x_0; w) + \nabla_x J(x_0; w)^T r$$

To formulate our optimization problem, we should find this r (which is $\leq \eta$) so it will maximize $J(x;w)$

$$\underset{r}{\text{maximize}} \quad J(x_0; w) + \nabla_x J(x_0; w)^T r \qquad \text{s.t} \quad \|r\|_\infty \leq \eta$$

ii) The above equation can be regarded as:

$$\underset{r}{\text{minimize}} \quad -J(x_0; w) - \nabla_x J(x_0; w)^T r \qquad \text{s.t} \quad \|r\|_\infty \leq \eta$$

we define: $w_0 = J(x_0; w)$ and $w^T r = -\nabla_x J(x_0; w)^T r$

$$\Rightarrow \underset{r}{\text{minimize}} \quad w^T r - w_0 \qquad \text{s.t.} \quad \|r\|_\infty \leq \eta$$

Now, we use the Holder's inequality: $\langle w, r \rangle = |w^T r| \leq \|w\|_p \|r\|_q$, $\frac{1}{p} + \frac{1}{q} = 1$ $\xrightarrow{p_1,\ q=\infty}$ $|w^T r| \leq \|w\|_1 \|r\|_\infty$

$$\Rightarrow -\|w\|_1 \|r\|_\infty \leq w^T r \leq \|w\|_1 \|r\|_\infty$$

As we have $\|r\|_\infty \leq \eta \rightarrow w^T r \geq -\eta \|w\|_1$ if we choose $r = -\eta \,\text{sign}(w)$, we will reach the lower bound $(-\eta \|w\|_1)$ &

$$w^T r = -\eta\, w^T \text{sign}(w) = -\eta \sum_i w_i \,\text{sign}(w_i) = -\eta \sum_i |w_i| = -\eta \|w\|_1$$

⇒ So, the answer to our problem is $r = -\eta \,\mathrm{sign}(-\nabla_x J(x_0 ; w)) = \eta \,\mathrm{sign}(\nabla_x J(x_0 ; w))$

⇒ $x = x_0 + r = x_0 + \eta \,\mathrm{sign}(\nabla_x J(x_0 ; w))$

b) $z = w_1 x + b \qquad y = w_2 z = w_2 w_1 x + w_2 b \qquad J = \sum_i (\hat{y}_i - y_i)^2$

FGSM attack → $x = x_0 + \eta \,\mathrm{sign}\left(\nabla_x \left(\sum (\hat{y}_i - w_2 w_1 x_i - w_2 b)^2\right)\right)$

$$= x_0 + \eta \,\mathrm{sign}\left(2(-w_1 w_2)(\hat{y}_i - w_2 w_1 x_i - w_2 b)\right)$$