ML HW3

Maitar Asher

04/09/2023

1 Inconsistency of the fairness definitions

(i) All fairness definitions can be satisfied simultaneously when sensitive feature A is independent of Y.

- DP: $\mathbb{P}_0(\hat{Y} = \hat{y}) = \mathbb{P}_1(\hat{Y} = \hat{y}) = \mathbb{P}(\hat{Y} = \hat{y})$
- EO: $\mathbb{P}_0(\hat{Y} = \hat{y}|Y = y) = \mathbb{P}_1(\hat{Y} = \hat{y}|Y = y) = \mathbb{P}(\hat{Y} = \hat{y}|Y = y)$
- PP: $\mathbb{P}_0(Y = y|\hat{Y} = \hat{y}) = \mathbb{P}_1(Y = y|\hat{Y} = \hat{y}) = \mathbb{P}(Y = y|\hat{Y} = \hat{y})$

(ii)

*For this question, I will use frequently the following formula which is a combination of the chain rule and Bayes rules.

$$\mathbb{P}(D|B,C) = \frac{\mathbb{P}(D,B,C)}{\mathbb{P}(B,C)} = \frac{\mathbb{P}(C)\mathbb{P}(D|C)\mathbb{P}(B|D,C)}{\mathbb{P}(C)\mathbb{P}(B|C)} = \frac{\mathbb{P}(D|C)\mathbb{P}(B|D,C)}{\mathbb{P}(B|C)}$$

Suppose A is dependent on Y and that DP and PP both hold.

Then by PP: $\mathbb{P}_0(Y = y|\hat{Y} = \hat{y}) = \mathbb{P}_1(Y = y|\hat{Y} = \hat{y})$.

By * (C is the random variable A, B is $\hat{y}$, D is Y) we have
$\frac{\mathbb{P}_0(\hat{Y}=\hat{y}|Y=y)\mathbb{P}_0(Y=y)}{\mathbb{P}_0(\hat{Y}=\hat{y})} = \frac{\mathbb{P}_1(\hat{Y}=\hat{y}|Y=y)\mathbb{P}_1(Y=y)}{\mathbb{P}_1(\hat{Y}=\hat{y})}$

By DP: $\mathbb{P}_0(\hat{Y} = \hat{y}) = \mathbb{P}_1(\hat{Y} = \hat{y})$ so we can get rid of the denominator.

$\mathbb{P}_0(\hat{Y} = \hat{y}|Y = y)\mathbb{P}_0(Y = y) = \mathbb{P}_1(\hat{Y} = \hat{y}|Y = y)\mathbb{P}_1(Y = y)$

By expanding the first term in each side of the equation we get:

$\frac{\mathbb{P}(A=0,\hat{Y}=\hat{y},Y=y)\mathbb{P}_0(Y=y)}{\mathbb{P}(A=0,Y=y)} = \frac{\mathbb{P}(A=1,\hat{Y}=\hat{y},Y=y)\mathbb{P}_1(Y=y)}{\mathbb{P}(A=1,Y=y)}$

By expanding the term in the denominator we get:

$\frac{\mathbb{P}(A=0,\hat{Y}=\hat{y},Y=y)\mathbb{P}_0(Y=y)}{\mathbb{P}(A=0)\mathbb{P}_0(Y=y)} = \frac{\mathbb{P}(A=1,\hat{Y}=\hat{y},Y=y)\mathbb{P}_1(Y=y)}{\mathbb{P}(A=1)\mathbb{P}_1(Y=y)}$

$\frac{\mathbb{P}(A=0,\hat{Y}=\hat{y},Y=y)}{\mathbb{P}(A=0)} = \frac{\mathbb{P}(A=1,\hat{Y}=\hat{y},Y=y)}{\mathbb{P}(A=1)}$

$\mathbb{P}_0(\hat{Y} = \hat{y}, Y = y) = \mathbb{P}_1(\hat{Y} = \hat{y}, Y = y)$. So we have shown that the joint probability of $\hat{Y}$ and Y is independent of A.

From DP we know that $\hat{Y}$ is independent of A.

Then Y must also be independent of A. Contradiction! And so DP and PP cannot hold at the same time.

(iii)

Suppose A is dependent on Y and $\hat{Y}$ is dependent on Y. And also suppose that both DP and EO hold.

Since A is dependent on Y, knowing something on A changes the probability of Y. So for either $Y = 1$ or $Y = 0$ the probability of Y given some a is not equal to the probability of Y given the other a. Then for the below q equations, there must be at least one p that is not equal to 0.

$\mathbb{P}(Y = 1|A = 1) = \mathbb{P}(Y = 1|A = 0) + p$

$\mathbb{P}(Y = 0|A = 1) = \mathbb{P}(Y = 0|A = 0) + p'$

When we try to learn the relationship between p and p' by summing the above 2 equations we get: $1 = 1 + p + p'$
$p' = -p$

By EO: $\mathbb{P}(\hat{Y} = \hat{y}|Y = y, A = 0) = \mathbb{P}(\hat{Y} = \hat{y}|Y = y, A = 1)$

By *(C is the random variable A,B is Y):

$\frac{\mathbb{P}(Y=y|\hat{Y}=\hat{y},A=0)\mathbb{P}(\hat{Y}=\hat{y}|A=0)}{\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(Y=y|\hat{Y}=\hat{y},A=1)\mathbb{P}(\hat{Y}=\hat{y}|A=1)}{\mathbb{P}(Y=y|A=1)}$

By DP: $\mathbb{P}(\hat{Y} = \hat{y}|A = 0) = \mathbb{P}(\hat{Y} = \hat{y}|A = 1)$ so we can omit this term in both sides of the equation.

$\frac{\mathbb{P}(Y=y|\hat{Y}=\hat{y},A=0)}{\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(Y=y|\hat{Y}=\hat{y},A=1)}{\mathbb{P}(Y=y|A=1)}$

By expanding the numerator on each side of the equation we get:

$\frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y},A=0)}{\mathbb{P}(\hat{Y}=\hat{y},A=0)\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y},A=1)}{\mathbb{P}(\hat{Y}=\hat{y},A=1)\mathbb{P}(Y=y|A=1)}$

By expanding the first term in the numerator on each side of the equation we get:

$\frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y},A=0)}{\mathbb{P}(\hat{Y}=\hat{y},A=0)\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y},A=1)}{\mathbb{P}(\hat{Y}=\hat{y},A=1)\mathbb{P}(Y=y|A=1)}$

By expanding the first term in the denominator and numerator on each side of the equation we get:

$\frac{\mathbb{P}(A=0)\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=0)}{\mathbb{P}(A=0)\mathbb{P}(\hat{Y}=\hat{y}|A=0)\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(A=1)\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=1)}{\mathbb{P}(A=1)\mathbb{P}(\hat{Y}=\hat{y}|A=1)\mathbb{P}(Y=y|A=1)}$

By DP: $\mathbb{P}(\hat{Y} = \hat{y}|A = 0) = \mathbb{P}(\hat{Y} = \hat{y}|A = 1)$ so we can again omit this term in both sides of the equation.

$$\frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=0)}{\mathbb{P}(Y=y|A=0)} = \frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=1)}{\mathbb{P}(Y=y|A=1)}$$

By expanding the first term in the denominator we get:

$$\frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=0)}{\mathbb{P}(Y=0|A=0)+\mathbb{P}(Y=1|A=0)} = \frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=1)}{\mathbb{P}(Y=0|A=1)+\mathbb{P}(Y=1|A=1)}$$

Recall that we have shown that:

$$\mathbb{P}(Y=1|A=1) = \mathbb{P}(Y=1|A=0) + p$$

$$\mathbb{P}(Y=0|A=1) = \mathbb{P}(Y=0|A=0) + -p$$

Plugging it in we get:

$$\frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=0)}{(\mathbb{P}(Y=0|A=1)+p)+(\mathbb{P}(Y=1|A=1)-p)} = \frac{\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=1)}{(\mathbb{P}(Y=0|A=0)+-p)+(\mathbb{P}(Y=1|A=0)+p)}$$

And so each denominator sum to 1. So we get:

$$\mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=0) = \mathbb{P}(Y=y,\hat{Y}=\hat{y}|A=1)$$

So we have shown that the joint probability of $\hat{Y}$ and Y is independent of A.

From DP we know that $\hat{Y}$ is independent of A.

Then Y must also be independent of A. Contradiction! And so DP and EO cannot hold at the same time.

(iv)

Suppose A is dependent on Y. And also suppose that both EO and PP hold.

By PP: $\mathbb{P}(Y = y|\hat{Y} = \hat{y}, A = 0) = \mathbb{P}(Y = y|\hat{Y} = \hat{y}, A = 1)$

By *(C is random variable A, D is $\hat{Y}$ and B is Y) we get:

$\frac{\mathbb{P}(\hat{Y}=\hat{y}|Y=y,A=0)\mathbb{P}(Y=y|A=0)}{\mathbb{P}(\hat{Y}=\hat{y}|A=0)} = \frac{\mathbb{P}(\hat{Y}=\hat{y}|Y=y,A=1)\mathbb{P}(Y=y|A=1)}{\mathbb{P}(\hat{Y}=\hat{y}|A=1)}$

By EO: $\mathbb{P}(\hat{Y} = \hat{y}|Y = y, A = 0) = \mathbb{P}(\hat{Y} = \hat{y}|Y = y, A = 1)$ so we can omit the first term in the numerator in each side of the equation

$\frac{\mathbb{P}(Y=y|A=0)}{\mathbb{P}(\hat{Y}=\hat{y}|A=0)} = \frac{\mathbb{P}(Y=y|A=1)}{\mathbb{P}(\hat{Y}=\hat{y}|A=1)}$

$\mathbb{P}(Y = y|A = 0)\mathbb{P}(\hat{Y} = \hat{y}|A = 1) = \mathbb{P}(Y = y|A = 1)\mathbb{P}(\hat{Y} = \hat{y}|A = 0)$

$\frac{\mathbb{P}(\hat{Y}=\hat{y}|A=1)}{\mathbb{P}(\hat{Y}=\hat{y}|A=0)} = \frac{\mathbb{P}(Y=y|A=1)}{\mathbb{P}(Y=y|A=0)}$

By expanding the conditional probability on the left side of the equation using $\mathbb{P}(A|B) = \sum_C \mathbb{P}(A|B, C)P(C|B)$ we get:

$\frac{\mathbb{P}(\hat{Y}=\hat{y}|A=1,Y=1)\mathbb{P}(Y=1|A=1)+\mathbb{P}(\hat{Y}=\hat{y}|A=1,Y=0)\mathbb{P}(Y=0|A=1)}{\mathbb{P}(\hat{Y}=\hat{y}|A=0,Y=1)\mathbb{P}(Y=1|A=0)+\mathbb{P}(\hat{Y}=\hat{y}|A=0,Y=0)\mathbb{P}(Y=0|A=0)} = \frac{\mathbb{P}(Y=y|A=1)}{\mathbb{P}(Y=y|A=0)}$

Because we assume EO we can factorize it out

$\frac{\mathbb{P}(\hat{Y}=\hat{y}|A=1,Y=y)(\mathbb{P}(Y=1|A=1)+\mathbb{P}(Y=0|A=1))}{\mathbb{P}(\hat{Y}=\hat{y}|A=0,Y=y)(\mathbb{P}(Y=1|A=0)+\mathbb{P}(Y=0|A=0))} = \frac{\mathbb{P}(Y=y|A=1)}{\mathbb{P}(Y=y|A=0)}$

The probability in the parentheses sum to 1 se we get:

$\frac{1}{1} = \frac{\mathbb{P}(Y=y|A=1)}{\mathbb{P}(Y=y|A=0)}$

$\mathbb{P}(Y = y|A = 1) = \mathbb{P}(Y = y|A = 0)$ contradiction! The above shows that Y and A are independent, which contradicts our assumption. Thus we have proved that if A is dependent on Y, EO and PP cannot both hold.

2 Combining multiple classifiers

(i)

$$D_{T+1}(i) = \frac{D_T(i)\exp(-\alpha_T y_i f_T(x_i))}{Z_T} = \frac{(\frac{D_{T-1}(i)\exp(-\alpha_{T-1} y_i f_{T-1}(x_i))}{Z_{T-1}})\exp(-\alpha_T y_i f_T(x_i))}{Z_T}$$

$$= \frac{\exp((-\alpha_T y_i f_T(x_i)) + (-\alpha_{T-1} y_i f_{T-1}(x_i)))}{Z_T Z_{T-1}}$$

We can continue expanding $D_T - i$ until we get $D_T - i = D_2$. Note that D_2 is computed in the first iteration of the for loop ($t = 1$)

$D_2 = \frac{D_1(i)\exp(-\alpha_1 y_i f_1(x_i))}{Z_1}$

Since we know $D_1(i) = \frac{1}{m}$ for all $i = 1, ..., m$

$D_2 = \frac{1}{m}\frac{\exp(-\alpha_1 y_i f_1(x_i))}{Z_1}$

$$D_{T+1}(i) = \frac{1}{m}\frac{1}{\prod_{t=1}^{T} Z_t}\exp(\sum_{t=1}^{T} -\alpha_t y_i f_t(x_i))$$

$\exp(\sum_{t=1}^{T} -\alpha_t y_i f_t(x_i)) = \exp(-y_i \sum_{t=1}^{T} \alpha_t f_t(x_i)) = \exp(-y_i g(x_i)$

Because $g(x_i) = \sum_{t=1}^{T} \alpha_t f_t(x_i)$ by def. So:

$$D_{T+1}(i) = \frac{1}{m}\frac{1}{\prod_{t=1}^{T} Z_t}\exp(-y_i g(x_i))$$

(ii) Note m is the number of samples

$$err(g) = \frac{\sum_{i=1}^{m} \mathbb{1}[(sign(g(x_i)) \neq y_i]}{m} = \frac{\sum_{i=1}^{m} \mathbb{1}[g(x_i)y_i < 0]}{m}$$

Because if classifier g made a mistake it means the sign of the true label y_i and predicted label ($g(x_i)$) differs (one would be 1 and the other -1), so the product of those two is -1 so smaller than 0.

Because the 0-1 loss is upper bounded by exponential loss. By def it means: $\mathbb{1}[x < 0] \leq \exp(-x)$. We know:

$$err(g) = \frac{\sum_{i=1}^{m} \mathbb{1}[g(x_i)y_i < 0]}{m} \leq \frac{\sum_{i=1}^{m} \exp(-g(x_i)y_i)}{m}$$

We proved in (i) that: $D_{T+1}(i) = \frac{1}{m} \frac{1}{\prod_{t=1}^{T} Z_t} \exp(-y_i g(x_i))$

Then, $\exp(-y_i g(x_i)) = \frac{D_{T+1}(i)}{\frac{1}{m} \frac{1}{\prod_{t=1}^{T} Z_t}} = D_{T+1}(i) m \prod_{t=1}^{T} Z_t$

$$\frac{\sum_{i=1}^{m} \exp(-g(x_i)y_i)}{m} = \frac{\sum_{i=1}^{m} D_{T+1}(i) m \prod_{t=1}^{T} Z_t}{m} = \frac{m \prod_{t=1}^{T} Z_t \sum_{i=1}^{m} D_{T+1}(i)}{m} = \prod_{t=1}^{T} Z_t$$

$\sum_{i=1}^{m} D_{T+1}(i) = 1$ because we normalize the distribution weight in every step.

Thus:

$$err(g) \leq \prod_{t=1}^{T} Z_t$$

(iii)

$$Z_t = \sum_{i=1}^{m} D_{t+1}(i) = \sum_{i=1}^{m} D_t(i) \exp(-\alpha_t y_i f_t(x_i))$$

We can separate the equation to correct and incorrect classified cases where we know that in the correct classified cases $y_i f_t(x_i) = 1$ and in the incorrect classified cases $y_i f_t(x_i) = -1$

$$Z_t = \sum_{i=1}^{m} \mathbb{1}[y_i = f_t(x_i)] D_t(i) \exp(-\alpha_t y_i f_t(x_i)) + \sum_{i=1}^{m} \mathbb{1}[y_i \neq f_t(x_i)] D_t(i) \exp(-\alpha_t y_i f_t(x_i))$$

$$= \sum_{i=1}^{m} \mathbb{1}[y_i = f_t(x_i)] D_t(i) \exp(-\alpha_t) + \sum_{i=1}^{m} \mathbb{1}[y_i \neq f_t(x_i)] D_t(i) \exp(\alpha_t)$$

We are given that $\epsilon_t = \sum_{i=1}^{m} \mathbb{1}[y_i \neq f_t(x_i)] D_t(i)$

So, $1 - \epsilon_t = \sum_{i=1}^{m} \mathbb{1}[y_i = f_t(x_i)] D_t(i)$

$$Z_t = (1 - \epsilon_t) \exp(-\alpha_t) + (\epsilon_t) \exp(\alpha_t) = (1 - \epsilon_t) \frac{1}{\exp(\alpha_t)} + (\epsilon_t) \exp(\alpha_t)$$

$$= \exp(\alpha_t) \left(\frac{(1 - \epsilon_t)}{\exp(2\alpha_t)} + \epsilon_t \right)$$

By def $\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$

$\exp(2\alpha_t) = \exp(2 \cdot \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})) = \exp(ln(\frac{1-\epsilon_t}{\epsilon_t})) = \frac{1-\epsilon_t}{\epsilon_t}$ (recall $\exp(ln(x)) = x$)

$\exp(\alpha_t) = \exp(\frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})) = (\exp(ln(\frac{1-\epsilon_t}{\epsilon_t})))^{\frac{1}{2}} = (\frac{1-\epsilon_t}{\epsilon_t})^{\frac{1}{2}}$ (recall $a^{m \cdot n} = (a^m)^n$)

$$Z_t = \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \left(\frac{(1 - \epsilon_t)\epsilon_t}{(1 - \epsilon_t)} + \epsilon_t \right) = \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \cdot (2\epsilon_t) = 2\sqrt{\frac{(1 - \epsilon_t) \cdot (\epsilon_t)^2}{\epsilon_t}} = 2\sqrt{(1 - \epsilon_t) \cdot (\epsilon_t)}$$

(iv)

So we have shown that $err(g) \leq \prod_{t=1}^{T} 2\sqrt{(1-\epsilon_t)\cdot(\epsilon_t)}$

Suppose there is $\gamma_t > 0$ s.t $\epsilon_t = \frac{1}{2} - \gamma_t$

Then $\prod_{t=1}^{T} 2\sqrt{(1-\frac{1}{2}+\gamma_t)\cdot(\frac{1}{2}-\gamma_t)} = \prod_{t=1}^{T} 2\sqrt{(\frac{1}{2}+\gamma_t)\cdot(\frac{1}{2}-\gamma_t)} = \prod_{t=1}^{T} \sqrt{4(\frac{1}{4}-\gamma_t^2)}$

$= \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2}$

Given the proof that $1 + x \leq \exp(x)$ (see README.txt for citation) we know that $\prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \prod_{t=1}^{T} (\exp(-4\gamma_t^2))^{0.5} = \prod_{t=1}^{T} (\exp(-2\gamma_t^2))$
$= \exp(-2\sum_{t=1}^{T} \gamma_t^2)$

Thus $err(g) \leq \prod_{t=1}^{T} 2\sqrt{(1-\epsilon_t)\cdot(\epsilon_t)} = \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leq \exp(-2\sum_{t=1}^{T} \gamma_t^2)$

3 1-Norm Support Vector Machine

(i)

Minimize $\|\vec{w}\|_1 = \sum_{j=1}^{n} |\vec{w}_j|$

subject to $y_i(\vec{w} \cdot \vec{x}_i + w_0) \geq 1 \; i = 1, ..., m$

Note that $\forall j = 1, ..., n \; |\vec{w}_j| = \max(\vec{w}_j, -\vec{w}_j)$

Minimizing $\max(\vec{w}_j, -\vec{w}_j)$ would be the same as minimizing some real number z_j where $z_j \geq \vec{w}_j$ and $z_j \geq -\vec{w}_j$

So this optimization problem is equivalent to the linear problem:

$$\text{Minimize } \|\vec{w}\|_1 = \sum_{j=1}^{n} \max(\vec{w}_j, -\vec{w}_j) \propto \sum_{j=1}^{n} z_j$$

with the following constraints:

- $z_j \geq \vec{w}_j \longrightarrow n$ constraints
- $z_j \geq -\vec{w}_j \longrightarrow n$ constraints
- $y_i(\vec{w} \cdot \vec{x}_i + w_0) \geq 1 \longrightarrow m$ constraints

Thus the total number of constraints is $2n + m$

(ii)

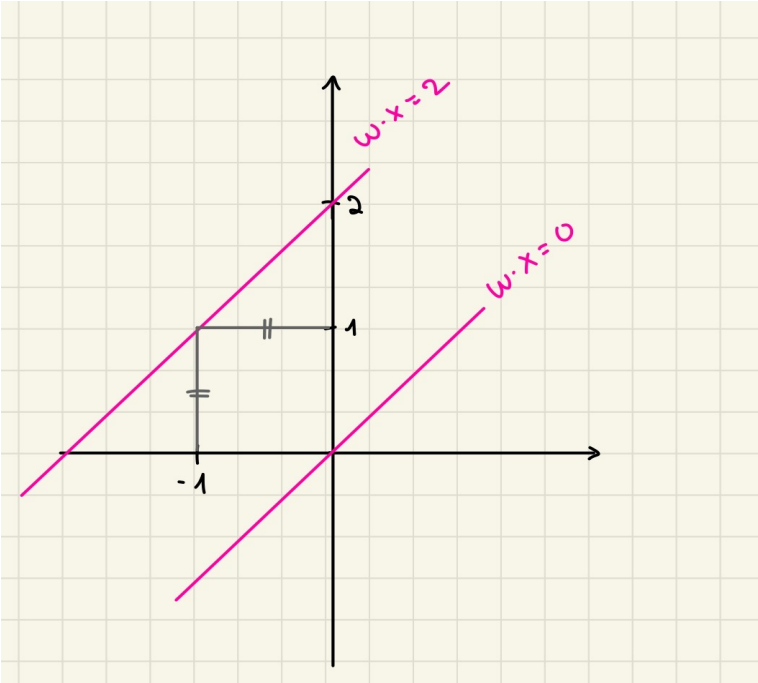
The (ℓ_∞) norm is the absolute max value of a vector's elements. Say $\vec{x} \in \mathbb{R}^n$, then $\|\vec{x}\|_\infty = \max_i |\vec{x}_i|$

The sign vector minimizes the (ℓ_∞) distance from the origin to the plane $\vec{w} \cdot \vec{x} = 2$ because it will intersect at a point at which the components of the vector are equal to each other but possibly only their sign differs. Suppose we're working in two dimensions, and the sign vector that intersects with the place is at ($|x_1| = |x_2|$) where x_1 is some value in the X_1 axis and x_2 is some value in the X_2 axis. If we take the (ℓ_∞) distance at different intersections point we are guaranteed to end up with an absolute value greater than $|x_1|$ or $|x_2|$ because the two components will no longer be equal. Since the sign vector only passes through points in which the absolute value of a vector component is equal, then the sign vector will always minimize the (ℓ_∞) distance from the origin to any plane.

Now we move $\vec{x} \cdot \vec{w} + w_0 = +-1$ in the space while assuming $w_0 = 0$ "without loss of generality" and in a way that $\vec{x} \cdot \vec{w} = -1$ will go through the origin.

Now we look at (ℓ_1)SVM; (ℓ_∞)SVM will minimze (ℓ_1) distance. This distance is the dot product between sign($\vec{w}$) and $\vec{w}$. This is proportional to the cos of the angle between $\vec{x}$ and its sign vector. Hence, it will try to maximize the angle between them. When we are trying to maximize the angle between $\vec{w}$ and sign($\vec{w}$), we are pushing $\vec{w}$ away from the sign vector toward this axis(in the constraints of each quadrant) because the sign is minimizing (ℓ_∞) between origin and $\vec{w} \cdot \vec{x} = 2$. When we maximize the angle between sign($\vec{w}$) and $\vec{w}$ vectors we are maximizing the (ℓ_∞) between the two planes. $\longrightarrow$ the $\vec{w}$ found in (ℓ_1)SVM will maximize the distance between the planes.

SEE graphic in next page.


w·x=2
2
1
-1
w·x=0

(iii)

We can take the same approach here as we did in (i).

Our first objective function:
Minimize $\|\vec{w}\|_1 = \sum_{j=1}^{n} \max(w_j, -w_j) \propto \sum_{j=1}^{n} z_j$

with the following constraints(for $j = 1, ..., n$):

- $\vec{\lambda_1} : z_j \geq w_j \longrightarrow n$ constraints
- $\vec{\lambda_2} : z_j \geq -w_j \longrightarrow n$ constraints

Our second objective function (for $i = 1, ..., m$):
Minimize $\sum_{i=1}^{m}[1-y_i(\vec{w}\cdot\vec{x}_i+w_0)]_+ = \sum_{i=1}^{m} \max(0, 1-y_i(\vec{w}\cdot\vec{x}_i+w_0)) \propto \sum_{i=1}^{m} t_i$

with the following constraints:

- $\vec{\lambda_3} : t_i \geq 0 \longrightarrow m$ constraints
- $\vec{\lambda_4} : t_i \geq 1 - y_i(\vec{w} \cdot \vec{x}_i + w_0) \longrightarrow m$ constraints

So our optimal problem, using Lagrange, becomes:

$$\min_{(\vec{w},w_0,\vec{z},\vec{t})} \max_{(\vec{\lambda_i}\geq 0,\forall i=1,..,4)*} L(\vec{w}, w_0, \vec{z}, \vec{t}, \vec{\lambda_i}) = \sum_{j=1}^{n} z_j + \sum_{i=1}^{m} t_i + \sum_{i=1}^{4} (\vec{\lambda_i} g_i(\vec{w}, w_0, \vec{z}, \vec{t}))$$

$$= \sum_{j=1}^{n} z_j + \sum_{i=1}^{m} t_i + \sum_{j=1}^{n} \lambda_{1j}(w_j - z_j) - \sum_{j=1}^{n} \lambda_{2j}(w_j + z_j) - \sum_{i=1}^{m} (\lambda_{3i} t_i) + \sum_{i=1}^{m} (\lambda_{4i}(1 - y_i(\vec{w}\cdot\vec{x}_i + w_0) - t_i))$$

*All vector component of each λ_i are greater or equal to 0

**Note that $\vec{w}, \vec{t}, \vec{z}, \vec{\lambda_1}, \vec{\lambda_2}, \vec{\lambda_3}, \vec{\lambda_4}$ are all vectors and the notation of w_j for instance is the j's element in $\vec{w}$

$$\min_{(\vec{w},w_0,\vec{z},\vec{t})} \max_{(\vec{\lambda_i}\geq 0,\forall i=1,..,4)*} L(\vec{w}, w_0, \vec{z}, \vec{t}, \vec{\lambda_i}) =$$

$$\sum_{j=1}^{n} z_j + \sum_{i=1}^{m} t_i + \sum_{j=1}^{n} \lambda_{1j}(w_j - z_j) - \sum_{j=1}^{n} \lambda_{2j}(w_j + z_j) - \sum_{i=1}^{m} (\lambda_{3i} t_i)$$

$$+ \sum_{i=1}^{m} (\lambda_{4i}) - \sum_{i=1}^{m} (\lambda_{4i} y_i (\vec{w} \cdot \vec{x}_i)) - \sum_{i=1}^{m} (\lambda_{4i} y_i w_0) - \sum_{i=1}^{m} (\lambda_{4i} t_i)$$

$$= \sum_{j=1}^{n} z_j + \sum_{i=1}^{m} t_i + \sum_{j=1}^{n} (\lambda_{1j} w_j) - \sum_{j=1}^{n} (\lambda_{1j} z_j) - \sum_{j=1}^{n} (\lambda_{2j} w_j) - \sum_{j=1}^{n} (\lambda_{2j} z_j) - \sum_{i=1}^{m} (\lambda_{3i} t_i)$$

$$+ \sum_{i=1}^{m} (\lambda_{4i}) - \sum_{j=1}^{n} \left(w_j \left(\sum_{i=1}^{m} (\lambda_{4i} y_i ((\vec{x}_i)_j)) \right) \right) - w_0 \sum_{i=1}^{m} (\lambda_{4i} y_i) - \sum_{i=1}^{m} (\lambda_{4i} t_i)$$

$$\min_{(\vec{w},w_0,\vec{z},\vec{t})} \max_{(\vec{\lambda_i}\geq 0,\forall i=1,..,4)*} L(\vec{w},w_0,\vec{z},\vec{t},\vec{\lambda_i}) =$$

$$\sum_{j=1}^{n} z_j(1-\lambda_{1j}-\lambda_{2j})+\sum_{i=1}^{m} t_i(1-\lambda_{3i}-\lambda_{4i})+\sum_{j=1}^{n} w_j \left(\lambda_{1j} - \lambda_{2j} - \sum_{i=1}^{m}(\lambda_{4i}y_i((\vec{x}_i)_j))\right)$$

$$-w_0 \sum_{i=1}^{m}(\lambda_{4i}y_i) + \sum_{i=1}^{m}(\lambda_{4i})$$

We can redefine the constraints as followed because if the parameters we're trying to learn $(\vec{w},w_0,\vec{z},\vec{t})$ can go to $-\infty$ on their own without lagrange var pushing to the other direction, then the expression multiplied by the parameters has to be 0.

- $(1-\lambda_{1j}-\lambda_{2j})=0$, $1=\lambda_{1j}+\lambda_{2j}$, so $\lambda_{1j},\lambda_{2j}\leq 1$
- $(1-\lambda_{3i}-\lambda_{4i})=0$, $1=\lambda_{3i}+\lambda_{4i}$, so $\lambda_{3i},\lambda_{4i}\leq 1$
- $(\lambda_{1j}-\lambda_{2j}-\sum_{i=1}^{m}(\lambda_{4i}y_i((\vec{x}_i)_j)))=0$

  since we know that $\lambda_{1j},\lambda_{2j}\leq 1$, so $\lambda_{1j}-\lambda_{2j}\leq 1$

  so $|\sum_{i=1}^{m}(\lambda_{4i}y_i((\vec{x}_i)_j))|\leq 1$
- $\sum_{i=1}^{m}(\lambda_{4i}y_i)=0$

so this problem can be expressed as:

$$\max_{\vec{\lambda_4}\in\mathbb{R}^m} \sum_{i=1}^{m}(\lambda_{4i})$$

s.t:

$|\sum_{i=1}^{m}(\lambda_{4i}y_i((\vec{x}_i)_j))|\leq 1$ for $j=1,...,n$

$\sum_{i=1}^{m}(\lambda_{4i}y_i)=0$

$0\leq\lambda_{4i}\leq 1$

Where $(\vec{\lambda_4})$ here is $\vec{\pi}$

(iv)

For a scenario where the optimal weight w is sparse, using the 1-norm SVM would be more appropriate than the 2-norm SVM. This is because the 1-norm SVM enforces sparsity by driving weights to zero, whereas the 2-norm SVM only drives some weights close to zero.

4 Estimating Model Parameters for Regression

(i)

Our optimization problem is to find

$$\max_{\vec{\beta}\in\mathbb{R}^d} Q(\vec{\beta}) \text{ s.t } \|\vec{\beta}\| \leq 1$$

Similarly, to find optimal $\vec{\beta}$ that maximizes our objective function, we can find the min of the negation of the objective function:

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) \text{ s.t } \|\vec{\beta}\| \leq 1$$

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} \ln f_{\vec{\beta}}(\vec{x_i}, y_i) = -\frac{1}{n}\sum_{i=1}^{n} \ln \mathbb{P}_{\vec{\beta}}(\vec{x_i}, y_i) = -\frac{1}{n}\sum_{i=1}^{n} \ln \mathbb{P}_{\vec{\beta}}(\vec{x_i})\mathbb{P}_{\vec{\beta}}(y_i|\vec{x_i})$$

Where $\vec{x_i} \in \mathbb{R}^d$ and n is the number of samples.

Because we know that $\vec{x_i} \sim N(0,1)$ and that it does not depend on $\vec{\beta}$ which is the parameter we're trying to maximize we can ignore this term in our optimization $\mathbb{P}_{\vec{\beta}}(\vec{x_i}) = \mathbb{P}(\vec{x_i})$

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) \propto -\sum_{i=1}^{n} \ln \mathbb{P}_{\vec{\beta}}(y_i|\vec{x_i})$$

We know that $\mathbb{P}_{\vec{\beta}}(y_i|\vec{x_i}) \sim N(\vec{x_i}^T\vec{\beta}, \|\vec{x_i}\|^2)$. So $\mathbb{P}_{\vec{\beta}}(y_i|\vec{x_i}) = \frac{1}{\sqrt{2\pi\|\vec{x_i}\|^2}} \exp(-\frac{(y_i-(\vec{x_i}^T\vec{\beta}))^2}{2\|\vec{x_i}\|^2})$

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) \propto -\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\|\vec{x_i}\|^2}} \exp\left(-\frac{(y_i-(\vec{x_i}^T\vec{\beta}))^2}{2\|\vec{x_i}\|^2}\right)\right) =$$

$$-\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\|\vec{x_i}\|^2}}\right) - \sum_{i=1}^{n} \ln\left(\exp\left(-\frac{(y_i-(\vec{x_i}^T\vec{\beta}))^2}{2\|\vec{x_i}\|^2}\right)\right)$$

The first term in the above equation does not depend on the parameter β that we're trying to learn so we ignore it. For the second term ln cancels the exp and the denominator is a positive constant so we can omit it as well.

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) \propto -\sum_{i=1}^{n}\left(-\frac{(y_i-(\vec{x_i}^T\vec{\beta}))^2}{2\|\vec{x_i}\|^2}\right) \propto \sum_{t=1}^{n}(y_i-(\vec{x_i}^T\vec{\beta}))^2$$

So for our optimization problem, we have:

$$\min_{\vec{\beta}\in\mathbb{R}^d} -Q(\vec{\beta}) \propto \sum_{t=1}^{n}(y_i-(\vec{x_i}^T\vec{\beta}))^2 \text{ s.t } \|\vec{\beta}\| \leq 1$$

For the optimization problem to be convex we need to show that our objective function is a convex function and that the feasible set induced by the constraint is a convex set.

For $\|\vec{\beta}\| \leq 1$ we have a form of $L2$ ball (unit ball), which is a convex set.

For the objective function, I will explore its second derivative and show it is positive-semi definite.

$$\frac{\partial}{\partial \vec{\beta}} \left(\sum_{i=1}^{n} (y_i - (\vec{x_i}^T \vec{\beta}))^2 \right) = \sum_{i=1}^{n} 2(y_i - (\vec{x_i}^T \vec{\beta}))(-\vec{x_i}) = \sum_{i=1}^{n} \left(-2y_i \vec{x_i} + (\vec{x_i}^T \vec{\beta}) \vec{x_i} \right) = \sum_{i=1}^{n} \left(-2y_i \vec{x_i} + \vec{x_i} (\vec{x_i}^T \vec{\beta}) \right)$$

$$= \sum_{i=1}^{n} \left(-2y_i \vec{x_i} + (\vec{x_i} \vec{x_i}^T) \vec{\beta}) \right)$$

$$\frac{\partial}{\partial \vec{\beta}} \frac{\partial}{\partial \vec{\beta}} \left(\sum_{t=i}^{n} (y_i - (\vec{x_i}^T \vec{\beta}))^2 \right) = \sum_{i=1}^{n} (\vec{x_i} \vec{x_i}^T)$$

$\sum_{i=1}^{n} (\vec{x_i} \vec{x_i}^T)$ is a dXd matrix. $\forall \vec{v} \in \mathbb{R}^d \sum_{i=1}^{n} \vec{v}^T (\vec{x_i} \vec{x_i}^T) \vec{v} = \sum_{i=1}^{n} (\vec{v}^T \vec{x_i})(\vec{x_i}^T \vec{v})$

$= \sum_{i=1}^{n} (\vec{v}^T \vec{x_i})^2$

Thus

$$\forall \vec{v} \in \mathbb{R}^d \sum_{i=1}^{n} \vec{v}^T (\vec{x_i} \vec{x_i}^T) \vec{v} = \sum_{i=1}^{n} (\vec{v}^T \vec{x_i})^2 \geq 0$$

And so our second derivative is a positive-semi definite matrix and our objective function is convex.

(i)

As we have shown, to find maximizers of $\vec{\beta}$ over all vectors in $\mathbb{R}^d$ we can find the minimizers of the negation of the function $Q(\vec{\beta})$.
Since we have also shown that our optimization problem is convex, to find the minimum we can simply set the first derivative to 0, and the solution would be the min.

$$\frac{\partial}{\partial \vec{\beta}} \left(\sum_{i=1}^{n} (y_i - (\vec{x_i}^T \vec{\beta}))^2 \right) = \sum_{i=1}^{n} \left(-2y_i \vec{x_i} + (\vec{x_i} \vec{x_i}^T) \vec{\beta} \right) = 0$$

$$-1 \left(\sum_{i=1}^{n} (2y_i \vec{x_i}) \right) + \sum_{i=1}^{n} \left((\vec{x_i} \vec{x_i}^T) \vec{\beta} \right) = 0$$

$$\sum_{i=1}^{n} \left(\vec{x_i} \vec{x_i}^T \right) \vec{\beta} = \sum_{i=1}^{n} (2y_i \vec{x_i})$$

And so

$$A = \sum_{i=1}^{n} \left(\vec{x_i} \vec{x_i}^T \right)$$

$$b = \sum_{i=1}^{n} (2y_i \vec{x_i})$$