ML Homework 3 Solutions

Ahuva Bechhofer

April 11, 2023

1 Inconsistency of the fairness definitions

1.1 Part 0 - i

For all three definitions to be satisfied at the same time, A cannot cannot have a relationship to Y. So for example predicting the final grade of a student in a class based on the color of their shoe laces.

1.2 ii

Proof by contra positive:
DP and PP imply that $Y \perp\!\!\!\perp A|\hat{Y}, \hat{Y} \perp\!\!\!\perp A \rightarrow Y \perp\!\!\!\perp A$
$P[AnY] = P[AnY|\hat{Y} = 0] * P[\hat{Y} = 0] + P[AnY|\hat{Y} = 1] * P[\hat{Y} = 1]$
$\rightarrow$ because of the conditional independance we can write
$P[AnY|\hat{Y}] = P[A|\hat{Y}] * P[Y|\hat{Y}]$ therefore we can rewrite $P[AnY]$ as:
$P[AnY] = P[A|\hat{Y} = 0] * P[Y|\hat{Y} = 0] * P[\hat{Y} = 0] + P[A|\hat{Y} = 1] * P[Y|\hat{Y} = 1] * P[\hat{Y} = 1]$
Using the fact that $A \perp\!\!\!\perp \hat{Y}$ we can rewrite the above as:
$P[AnY] = P[A] * P[Y|\hat{Y} = 0] * P[\hat{Y} = 0] + P[A] * P[Y|\hat{Y} = 1] * P[\hat{Y} = 1]$
If we pull out P[A], we get:
$P[AnY] = P[A] * (P[Y|\hat{Y} = 0] * P[\hat{Y} = 0] + P[Y|\hat{Y} = 1] * P[\hat{Y} = 1]) \rightarrow$
$P[AnY] = P[A]P[Y]$ since $P[Y] = [Y|\hat{Y} = 0]*P[\hat{Y} = 0]+P[Y|\hat{Y} = 1]*P[\hat{Y} = 1]$

Therefor we showed Y and A are independant. and since we showed that Y and A are can only be independant if both rules hold and therefore if one rule doesnt hold, A is dependant on Y.

1.3 iii

DP and EO imply that $Y \perp\!\!\!\perp A, \hat{Y} \perp\!\!\!\perp A|Y$
we start by writing $P_0[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 0nY = 1]P[Y = 1|A = 0] + P[\hat{Y} = 1|A = 0nY = 0]P[Y = 0|A = 0]$ Note we could also do the same with $\hat{Y} = 0$ but it doesnt make a difference as they are equal so I am showing it for one of the cases.
Using EO: $P[\hat{Y} = 1|A = 0nY = 1] = P[\hat{Y} = 1|A = 1nY = 1]$ therefore we can

rewrite the previous equation as:
$P_0[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1nY = 1]P[Y = 1|A = 0] + P[\hat{Y} = 1|A = 1nY = 0]P[Y = 0|A = 0]$
Using the fact that $Y \perp\!\!\!\perp A, P[Y = 1|A = 0] = P[Y = 1|A = 1] + \alpha$
and $P[Y = 0|A = 0] = P[Y = 0|A = 1] + \beta$
We know $P[Y = 1|A = 0]+P[Y = 0|A = 0] = 1 \rightarrow P[Y = 1|A = 1]+\alpha+P[Y = 0|A = 1] + \beta = 1 \rightarrow \beta = -\alpha$
Now we can rewrite the equation as $P_0[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1nY = 1](P[Y = 1|A = 1] + \alpha) + P[\hat{Y} = 1|A = 1nY = 0](P[Y = 0|A = 1] - \alpha)$
distributing this we get: $P[\hat{Y} = 1|A = 1nY = 1] * P[Y = 1|A = 1] + P[\hat{Y} = 1|A = 1nY = 1] * \alpha) + P[\hat{Y} = 1|A = 1nY = 0] * P[Y = 0|A = 1] - P[\hat{Y} = 1|A = 1nY = 0]\alpha$
since we know $\hat{Y} \perp\!\!\!\perp Y, P[\hat{Y} = 1|A = 1nY = 1] = P[\hat{Y} = 1|A = 0nY = 1] + \gamma$
we can again rewrite the equation as: $P[\hat{Y} = 1|A = 1nY = 1] * P[Y = 1|A = 1] + (P[\hat{Y} = 1|A = 0nY = 1] + \gamma) * \alpha) + P[\hat{Y} = 1|A = 1nY = 0] * P[Y = 0|A = 1] - P[\hat{Y} = 1|A = 1nY = 0]\alpha$ if we distribute this we get:
$P[\hat{Y} = 1|A = 1nY = 1] * P[Y = 1|A = 1] + P[\hat{Y} = 1|A = 0nY = 1] * \alpha + \gamma * \alpha + P[\hat{Y} = 1|A = 1nY = 0] * P[Y = 0|A = 1] - P[\hat{Y} = 1|A = 1nY = 0]\alpha$
This simplifies to $P[\hat{Y} = 1|A = 1nY = 1] * P[Y = 1|A = 1] + \gamma * \alpha + P[\hat{Y} = 1|A = 1nY = 0] * P[Y = 0|A = 1]$
Using the fact that $P_1[\hat{Y}] = P[\hat{Y}|A = 1], P[\hat{Y} = 1|A = 1] = P[\hat{Y} = 1|A = 1nY = 1] * P[Y = 1|A = 1] + P[\hat{Y} = 1|A = 1nY = 0] * P[Y = 0|A = 1]$
$P_0[\hat{Y} = 1] = P_1[\hat{Y} = 1] + \gamma * \alpha$ where neither equal 0
Therefore $P_0[\hat{Y}] \neq P_1[\hat{Y}] \rightarrow$ DP can't be true!

1.4 iv

We start by stating what we are given and some helpful facts: $FPR_a(= P_a[\hat{Y} = 1|Y = 0]), FNR_a(= P_a[\hat{Y} = 0|Y = 1])$ and that if A is dependant on Y it is not (EO and PP).

$P_a[\hat{Y} = 1|Y = 0]$ can be expanded with bayes rule to $\frac{P_a[Y=0|\hat{Y}=1]*P_a[\hat{Y}=1]}{P_a[Y=0]}$ $= \frac{(1-P_a[Y=1|\hat{Y}=1])*P_a[\hat{Y}=1]}{P_a[Y=0]}$

$$= \frac{(1 - PPV_a) * P_a[\hat{Y} = 1]}{P_a[Y = 0]} \tag{1}$$

$P_a[\hat{Y} = 1|Y = 1]$ can be expanded with bayes rule to

$$\frac{P_a[Y = 1|\hat{Y} = 1] * P_a[\hat{Y} = 1]}{P_a[Y = 11]} \tag{2}$$

If we multiply the first equation by the reciprical of the second equation we get:

$$\frac{FPR_a}{1-FNR_a} = \frac{(1-PPV_a) * P_a[Y=1]}{PPV_a * P_a[Y=0]} \tag{3}$$

if we solve for the following $(1-FNR_a) = \frac{PPV_a P_0[Y=0]}{FPR_a(1-PPV_a)P_0[Y=1]}$ So if PP holds:
$PPV_1 = PPV_0$
and if EO holds: $FPR_1 = FPR_0$
And as stated at the begining if A is dependant on Y: $P_0[Y=1] \neq P_1[Y=0]$

therefore

$$(1-FNR_1) = \frac{PPV_1 P_1[Y=0]}{FPR_1(1-PPV_1)P_1[Y=1]} \neq (1-FNR_0) = \frac{PPV_0 P_0[Y=0]}{FPR_0(1-PPV_0)P_0[Y=1]} \tag{4}$$

Therefore they cannot hold all at the same time.

2 Combining multiple classifiers

2.1 i

Proof by induction:
Base case: T=1

$$D_1(i) = \frac{1}{m}$$

$$\begin{aligned} D_{1+1}(i) &= \frac{\frac{1}{m}\exp\left(-\alpha_1 y_i f_1(x_i)\right)}{\sum_j \frac{1}{m}\exp(\alpha_1 y_j f_1(x_j) = D_2 = z_1} \\ &= \frac{1}{m}\frac{1}{z_1}\cdot\exp\left(-\alpha y_i f_1(x_i)\right) \\ &= \frac{1}{m}\frac{1}{z_1}\cdot\exp\left(-y_i g(x_i)\right) \end{aligned}$$

base case proved!

showing if it holds for T it holds for T+1

$$D_{(T+1)+1} = \frac{D_{T+1}}{Z_T}\cdot\exp\left(-y_i\alpha_{T+1}f_{T+1}(x_i)\right)$$

$$= \frac{\frac{1}{m}\cdot\frac{1}{\prod_t Z_t}\exp\left(\sum_{j=1}^{T} -\alpha_j y_i f_j(x_i)\right)}{Z_T} * \exp\left(-y_i\alpha_{T+1}f_{T+1}(x_i)\right)$$

we can pull out y_i so we get:

$$D_{(T+1)+1} = \frac{\frac{1}{m}\cdot\frac{1}{\prod_t Z_t}\exp\left(-y_i g(x_i)\right)}{Z_T} * \exp\left(-y_i g(x_i)\right)$$

$$D_{(T+1)+1} = \frac{D_{T+1+1}}{Z_{T+1}}$$

2.2 ii

$Z_t : \mathrm{err}(g) \leq \prod_t Z_t = \prod_t \sum_i D_{t+1}(i) = \prod_t \sum_i D_t(i) \exp\left(-\alpha_t y_i f_t\left(x_i\right)\right)$
$\mathrm{err}(g) = \frac{1}{m} \sum_i \left[y_i \neq \operatorname{sign}\left(g\left(x_i\right)\right)\right]$

in the $\neq$ case we know $y_i\left(g\left(x_2\right)\right) \leq 0$

$$\mathrm{err}(g) = \frac{1}{m}\left[1\left[y_i\left(g\left(x_i\right)\right) < 0\right]\right.$$

because we know

$$\rightarrow 1[x<0] \leqslant \exp(-x)$$

$$\mathrm{err}(g) = \frac{1}{m} \sum_i 1\left[y_i \cdot g\left(x_i\right) \leq 0\right] \leq \frac{1}{m} \sum_i \exp\left[-y_i g\left(x_i\right)\right]$$

$$D_{t+1}(i) = \frac{1}{m} \frac{1}{\prod_t Z_t} \exp\left(-y_i g(x_i)\right)$$

$$\sum_i D_{T+1(i)} \cdot \prod_t Z_t = \sum_i \frac{1}{m} \exp\left(-y_i g(x_i)\right.$$

$$\mathrm{err}(g) \leq \sum_i D_{T+1}(i)\left(\prod_t Z_t\right) \text{ we can pull out } \prod_t Z_t \text{ from the sum}$$

$$\prod_t Z_t \underbrace{\sum_i D_{T+1}(i)}_{1} \leq \prod_t Z_t$$

$$\text{because: } \sum_j D_{T+1}(j) = \sum_j \frac{D_{T+1}(j)}{\sum_i D_{T+1}(i)}$$

$$\sum_j D_{T+1}(j) = \frac{1}{\sum_i D_{T+1}(i)} \cdot \sum_j D_{T+1}(j)$$

$$\sum_j D_{T+1}(j) = 1 \quad \therefore \mathrm{err}(g) \leq \prod_t Z_t$$

2.3 iii

iii) show $z_t = 2\sqrt{\epsilon_t\left(1-\epsilon_t\right)}, \epsilon_j = \sum_{i=1}^{m} D_t(i) \cdot 1\left[y_i \neq f_j\left(x_j\right)\right]$ for much $f_i \in \mathcal{F}$

$Z_t = \sum_i D_t(i) \exp\left(-\alpha_t y_i f_t\left(x_i\right)\right) \quad y_i \in (-1, +1)$

if Correct: $D_+(i) \exp\left(-\alpha_t\right) \quad f : x \rightarrow (-1, +1)$

if incorrect: $D_i(t) \exp\left(\alpha_t\right)$

$= \sum_i D_t(i) \exp\left(\alpha_t\right)\ 1\left[y_i = f\left(x_i\right)\right] + \sum_i D_t(i) \exp(-\alpha_t)\ 1\left[y_i \neq f\left(x_i\right)\right]$

$\exp\left(-\alpha_t\right) \underbrace{\sum_i D_t(i) 1\left[y_i = f\left(x_i\right)\right]}_{1-\epsilon_t} + \exp\left(\alpha_t\right) \underbrace{\sum_i D_t(i) 1\left[y_i \neq f\left(x_i\right)\right]}_{\epsilon_t}$

$\exp\left(-\alpha_t\right)\left[1-\epsilon_t\right] + \exp\left(\alpha_t\right)\left[\epsilon_t\right]$

$\exp\left(-\frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)\right)\left[1-\epsilon_t\right] + \exp\left(\frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)\right) \epsilon_t$

$a \ln x = \ln\left(x^a\right)$

$\exp\left(\ln\left(\left(\frac{1-\epsilon_t}{\sigma_t}\right)^{-1/2}\right)\left[1-\epsilon_j\right] + \exp\left(\ln\left(\left(\frac{1-\epsilon_t}{\epsilon_t}\right)^{1/2}\right) \epsilon_t\right.\right.$

$= \frac{\left[1-\epsilon_t\right]}{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} + \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$

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

$= \left(1-\epsilon_t\right)^{1/2} \epsilon_t^{1/2} + \epsilon_t^{1/2}\left(1-\epsilon_t\right)^{1/2}$

$= \sqrt{1-\epsilon_t} \cdot \sqrt{\epsilon_t} + \sqrt{\epsilon_t}\sqrt{1-\epsilon_t}$

$= 2\sqrt{\epsilon_t} \cdot \sqrt{1-\epsilon_t} = 2\sqrt{\epsilon_t\left(1-\epsilon_t\right)}$

2.4 iv

iv) $\text{err}(g) \le \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)}$

$$\prod_t 2\sqrt{\epsilon_t(1-c_t)}$$

$$= \epsilon_t = \frac{1}{2} - \gamma_t \to \left(\frac{1}{2} - \gamma_t\right)\left(1 - \frac{1}{2} + \gamma_t\right)$$

$$\frac{1}{2} - \frac{1}{4} + \frac{1}{2}\gamma_t - \gamma_t + \frac{1}{2}\gamma_2 - \gamma_t^2 = \frac{1}{4} - \gamma_t^2$$

$$\prod_t \cdot \sqrt{4 \cdot \left(\frac{1}{4} - \gamma_t^2\right)}$$

$$= \prod_t \cdot \sqrt{1 - 4\gamma_t^2}$$

$$1 - x \le e^{-x}$$
$$(1-x)^{1/2} \le \left(e^{-x}\right)^{1/4} = e^{-x/2}$$

$$\le \prod_t \exp\left(-2\gamma_t^2\right)$$

$$e^a \cdot e^b = a^{a+b}$$

$$e^{-2\gamma_1^2} \cdot e^{-2\gamma_2^2} \dots$$

$$e^{-2\gamma_1^2 - 2\gamma_2^2} \dots = e^{-2\left(\gamma_1^2 + \gamma_2^2 - \cdots\right)}$$

$$= \exp\left(-2\sum_t \gamma_t^2\right)$$

3 1-Norm Support Vector Machine

3.1 i

from $y_i(w \cdot x_i + w_0) \ge 1 \to m$
from def:

$$\|x\|_1 = \sum_{j=1}^{n} |x_j|$$

To get rid of the absolute value we can add the following constraints

$$\begin{array}{l} x_j \leqslant y_j \\ -x_j \leqslant y_j \\ j = 1...n \end{array} \Rightarrow min \sum_{j=1}^{n} y_i \quad \begin{array}{r} s.t \\ x_i \leqslant y_j \\ -x_j \leqslant y_j \end{array}$$

therefore we get m+2n constraints and 2n + 1 (for w_0) variables

3.2 ii

$$\vec{w} \cdot \vec{x} = 2 = \sum_{i=1}^{n} w_i * x_i = \sum_{i=1}^{n} \lambda * w_i * x_i = 2$$

we need λ as a scalar to be able to sum to 2 to ensure our dot product sums to 2 in all cases of different vector values, and we wart our w vector to all be the same value and sign corrected since we want to get to 2 in as few iterations as possible/ as fast as possible without affecting the l_∞ distance.

$$\begin{aligned} \textstyle\sum_{i=1}^{n} \lambda * w_i * x_i = 2 \\ \lambda = \tfrac{2}{w_i \cdot x_i} \end{aligned}$$

$$= \frac{2}{\sum_i w_i \times x_i}$$

$$\lambda = \frac{2}{\sum_i |w_i|} = \frac{2}{\|w\|_1}.$$

Therefore we see by the above equation that if we maximize lambda we minimize w

3.3 iii

express (1) as linear program

$$\min \sum_{i=1}^{n} z_i + \sum_{i=1}^{m} \underbrace{[1 - y_i\,(w \cdot x_i + w_0)]}_{t_i}{}_{+} \geq 0, \alpha_i = 1 - y_i(w \cdot x_i + w_0)$$

$$\text{s.t} \quad \begin{aligned} w_i &\leq z_i \\ -w_i &\leq z_i \end{aligned}$$

$$\min \sum_{i=1}^{n} z_i + \sum_{i=1}^{m} t_i$$

s.t.

$$\begin{array}{ll} w_i - z_i \leq 0 & -w - z_i \leq 0 \\ -t_i \leq 0 & \alpha_i - t_i \leq 0 \end{array}$$

$$p^* = \min_{z_i, t_i, \omega,} \max_{\lambda, \rho, \gamma, \beta} L(\overbrace{z_i, t_i, \omega,}^{S_i} \overbrace{\lambda, \rho, \gamma, \beta}^{\sigma_i})$$

$$d^* = \max_{\sigma_i} \min_{s_i} L(S_i, \sigma_i)$$

$$L\,(S_i, \sigma_i) = \sum_{i=0}^{n} z_i + \sum_{i=0}^{m} t_i + \sum_{i=1}^{n} \lambda_i\,(w_i - z_i) + \sum_{i=1}^{n} \rho_i\,(-w_i - z_i)$$

$$+ \sum_{i=1}^{m} \gamma_i\,(-t_i) + \sum_{i=1}^{m} \beta_i\,(\alpha_i - t_i)$$

$$= \sum_{i=0}^{n} z_i + \sum_{i=0}^{m} t_i + \sum_{i=1}^{n} \lambda_i\,(w_i) - \sum_{i=1}^{n} \lambda_i\,(z_i) - \sum_{i=1}^{n} \rho_i\,(w_i) - \sum_{i=1}^{n} \rho_i\,(z_i$$

$$+ \sum_{i=1}^{m} \gamma_i\,(-t_i) + \sum_{i=1}^{m} \beta_i\,(\alpha_i) - \sum_{i=1}^{m} \beta_i\,(t_i)$$

we can rewrite everything in terms of dot products for simplicity

let A be a vector of all ones

$$= A \cdot z + A \cdot t + \lambda \cdot w - \lambda \cdot z - \rho \cdot w - \rho \cdot z$$

$$+ \gamma \cdot t + \underbrace{\beta \cdot \alpha}_{} - \beta \cdot t$$

$$\beta \cdot (1 - y_i\,(w \cdot x_i + \omega_0))$$

$$= z \cdot (A - \lambda - \rho) + t \cdot (A - \gamma - \beta) + w \cdot (\lambda - \rho - \beta \cdot y \cdot x) - w_0 \cdot (\beta \cdot y) + \beta$$

Constrains:

- $A - \lambda - \rho = 0 \rightarrow 1 = \lambda + \rho \rightarrow \lambda, \rho \leq 1$
- $A - \gamma - \beta = 0 \rightarrow 1 = \gamma + \beta \rightarrow \gamma, \beta \leq 1$

-and any lagrange needs to be ≥ 0

$\lambda - \rho - \beta \cdot y \cdot x = 0 \rightarrow |\beta \cdot y \cdot x| \leq 1$

- $\beta \cdot y = 0$

If our vectors $\vec{z}, \vec{t}, \vec{w}$ go on thier own to negative infinity without the lagrange variables having the ability to push them towards positive infinity then the expression multiplied with the parameter has to be zero.

So our optimization problem becomes:

$$\max_{0 \leq \beta i} \sum_{i}^{m} \beta i$$
$$s.t$$
$$\rho \cdot y = 0$$
$$|\beta \cdot y \cdot x| \leq 1$$
$$0 \leq \beta_i \leq 1$$

3.4 iv

L_1 SVM will make make more sense because it causes the weights to go all the way to zero, while L_2 only makes some of the weights approach zero.
L_1 is better when there is multicolinearity and L_2 is better when only a few inputs control the output.

4 Estimating Model Parameters for Regression

4.1 i

$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} ln(f_\beta(x_i, y_i))$
$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} ln(f_\beta(y_i|x_i) * f_\beta(x_i))$
since the disterbution of x is known and we are looking for beta that minimizes Q, we can leave out the $f_\beta(x_i))$ term.
$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} ln(f_\beta(y_i|x_i))$
since we are given $(y_i|x_i)$ disterbution we can replace the inside of the log with the Gaussian distri- bution with mean $\mu = X^T\beta$ and variance $\sigma^2 = \|x\|^2$
$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} ln(\frac{1}{\sqrt{2\pi\|x_i\|^2}} \exp \frac{-(y_i - x_i^T\beta}{2\|x_i\|^2})$
$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} ln(\frac{1}{\sqrt{2\pi\|x_i\|^2}}) + ln(\exp \frac{-(y_i - x_i^T\beta}{2\|x_i\|^2}))$
Since the first ln term does not contain beta in it, we can remove it from the optimization.
$Q(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{-(y_i - x_i^T\beta}{2\|x_i\|^2})$
To prove convexity, we can show the hessian matrix is positive semi definite and therefore it implies convexity.
$d\beta' = \sum_{i=1}^{n} \frac{-(x_i 2(y_i - x_i^T\beta))}{2\|x_i\|^2}$

$d\beta'' = \sum_{i=1}^{n} \frac{x_i x_i^T}{\|x_i\|^2}$
To prove this is positive semi definite we need to show there is some H in

$Z^T H Z \geq 0$ that makes the inequaliy true in our case $H = x_i x_i^T$
If we replace H with $x_i x_i^T \rightarrow Z^T x_i x_i^T Z \geq 0 \rightarrow (x_i^T Z)^T x_i^T Z \rightarrow \|x_i^T Z\|^2 \geq 0$
since we know $x^T \cdot x = x \cdot x = \|x\|^2$
therefore the second derivative is positive semi definite and thus convex, and the sum of convex functions are convex therefore the whole thing is convex
Now since the objective function is convex and the constraints are convex then the optimization problem is convex.

.

4.2 ii

from part 1 we know $d\beta' = \frac{-2(x_i(y_i - x_i^T \beta)}{\|x_i\|^2} = 0$
We can ignore the denominator since since it is not with respect to beta
$\sum_{i=1}^n (-2x_i y_i + 2x_i x_i^T \beta) = 0 \rightarrow (\underbrace{\sum_{i=1}^n x_i x_i^T}_{A})\beta = \underbrace{\sum_{i=1}^n x_i y_i}_{b} \rightarrow A\beta = b$
A is a dxd matrix and b is a dx1 vectro.