COMS 4771 HW3 (Spring 2023)

Due: Fri Apr 07, 2023 at 11:59pm

This homework is to be done **alone**. No late homeworks are allowed. To receive credit, a typesetted copy of the homework pdf must be uploaded to Gradescope by the due date. You must show your work to receive full credit. Discussing possible approaches for solutions for homework questions is encouraged on the course discussion board and with your peers, but you must write your own individual solutions and **not** share your written work/code. You must cite all resources (including online material, books, articles, help taken from/given to specific individuals, etc.) you used to complete your work.

1 Inconsistency of the fairness definitions

Recall the notation and definitions of group-based fairness conditions:
Notation:

Denote $X \in \mathbb{R}^d,\ A \in \{0,1\}$ and $Y \in \{0,1\}$ to be three random variables: non-sensitive features of an instance, the instance's sensitive feature and the target label of the instance respectively, such that $(X, A, Y) \sim \mathcal{D}$. Denote a classifier $f : \mathbb{R}^d \to \{0,1\}$ and denote $\hat{Y} := f(X)$.

For simplicity, we also use the following abbreviations:

$$\mathbb{P} := \mathbb{P}_{(X,A,Y)\sim D} \qquad \text{and} \qquad \mathbb{P}_a := \mathbb{P}_{(X,a,Y)\sim D}$$

Group based fairness definitions:

- Demographic Parity (DP)

$$\mathbb{P}_0[\hat{Y} = \hat{y}] = \mathbb{P}_1[\hat{Y} = \hat{y}] \qquad \forall \hat{y} \in \{0,1\}$$

(equal positive rate across the sensitive attribute)

- Equalized Odds (EO)

$$\mathbb{P}_0[\hat{Y} = \hat{y} \mid Y = y] = \mathbb{P}_1[\hat{Y} = \hat{y} \mid Y = y] \qquad \forall \hat{y},\ y \in \{0,1\}$$

(equal true positive- and true negative-rates across the sensitive attribute)

- Predictive Parity (PP)

$$\mathbb{P}_0[Y = y \mid \hat{Y} = \hat{y}] = \mathbb{P}_1[Y = y \mid \hat{Y} = \hat{y}] \qquad \forall \hat{y},\ y \in \{0,1\}$$

(equal positive predictive- and negative predictive-value across the sensitive attribute)

Unfortunately, achieving all three fairness conditions simultaneously is not possible. An impossibility theorem for group-based fairness is stated as follows.

- If A is dependent on Y, then Demographic Parity and Predictive Parity cannot hold at the same time.
- If A is dependent on Y and $\hat{Y}$ is dependent on Y, then Demographic Parity and Equalized Odds cannot hold at the same time.
- If A is dependent on Y, then Equalized Odds holds and Predictive Parity cannot hold at the same time.

These three results collectively show that it is impossible to simultaneously satisfy the fairness definitions except in some trivial cases.

(i) State a scenario where all three fairness definitions are satisfied simultaneously.

(ii) Prove the first statement.

(iii) Prove the second statement.

(iv) Prove the third statement.

Hint: First observe that

$$\mathbb{P}_0[Y = y|\hat{Y} = \hat{y}] = \mathbb{P}_1[Y = y|\hat{Y} = \hat{y}] \ \forall \hat{y},\ y \in \{0, 1\}$$

is equivalent to:

$$\mathbb{P}_0[Y = 1|\hat{Y} = \hat{y}] = \mathbb{P}_1[Y = 1|\hat{Y} = \hat{y}] \ \forall \hat{y} \in \{0, 1\}.$$

A necessary condition for PP is the equality of positive predictive value (PPV):

$$\mathbb{P}_0[Y = 1|\hat{Y} = 1] = \mathbb{P}_1[Y = 1|\hat{Y} = 1]$$

To prove the third statement, it is enough to prove a stronger statement: if A is dependent on Y, Equalized Odds and equality of Positive Predictive Value cannot hold at the same time.

Next, try to express the relationship between FPR$_a$ ($= \mathbb{P}_a[\hat{Y} = 1|Y = 0]$) and FNR$_a$ ($= \mathbb{P}_a[\hat{Y} = 0|Y = 1]$) using p_a ($= \mathbb{P}[Y = 1 \mid A = a]$) and PPV$_a$ ($= \mathbb{P}_a[Y = 1|\hat{Y} = 1]$), $\forall a \in \{0, 1\}$ and finish the proof.

2 Combining multiple classifiers

The concept of "wisdom-of-the-crowd" posits that collective knowledge of a group as expressed through their aggregated actions or opinions is superior to the decision of any one individual in the group. Here we will study a version of the "wisdom-of-the-crowd" for binary classifiers: how can one *combine* prediction outputs from multiple possibly low-quality binary classifiers to achieve an aggregate high-quality final output? Consider the following iterative procedure to combine classifier results.

Input:

- S – a set of training samples: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, where each $y_i \in \{-1, +1\}$
- T – number of iterations (also, number of classifiers to combine)
- $\mathcal{F}$ – a set of (possibly low-quality) classifiers. Each $f \in \mathcal{F}$, is of the form $f : X \to \{-1, +1\}$

Output:

- F – a set of selected classifiers $\{f_1, \dots, f_T\}$, where each $f_i \in \mathcal{F}$.
- A – a set of combination weights $\{\alpha_1, \dots, \alpha_T\}$

Iterative Combination Procedure:

```
- Initialize distribution weights $D_1(i) = \frac{1}{m}$    [for $i = 1, \dots, m$]
- **for** $t = 1, \dots, T$ do
-         // $\epsilon_j$ is weighted error of j-th classifier w.r.t. $D_t$
-     Define $\epsilon_j := \sum_{i=1}^{m} D_t(i) \cdot \mathbf{1}[y_i \neq f_j(x_i)]$    [for each $f_j \in \mathcal{F}$]
-         // select the classifier with the smallest (weighted) error
-     $f_t = \arg\min_{f_j \in \mathcal{F}} \ \epsilon_j$
-     $\epsilon_t = \min_{f_j \in \mathcal{F}} \ \epsilon_j$
-         // recompute weights w.r.t. performance of $f_t$
-     Compute classifier weight $\alpha_t = \frac{1}{2} \ln \left( \frac{1-\epsilon_t}{\epsilon_t} \right)$
-     Compute distribution weight $D_{t+1}(i) = D_t(i) \exp(-\alpha_t y_i f_t(x_i))$
-     Normalize distribution weights $D_{t+1}(i) = \frac{D_{t+1}(i)}{\sum_i D_{t+1}(i)}$
- **endfor**
- return weights $\alpha_t$, and classifiers $f_t$ for $t = 1, \dots, T$.
```

Final Combined Prediction:

- For any test input x, define the aggregation function as: $g(x) := \sum_t \alpha_t f_t(x)$, and return the prediction as $\text{sign}(g(x))$.

We'll prove the following statement: If for each iteration t there is some $\gamma_t > 0$ such that $\epsilon_t = \frac{1}{2} - \gamma_t$ (that is, assuming that at each iteration the error of the classifier f_t is just γ_t better than random guessing), then error of the aggregate classifier

$$\text{err}(g) := \frac{1}{m} \sum_i \mathbf{1}[y_i \neq \text{sign}(g(x_i))] \leq \exp(-2 \sum_{t=1}^{T} \gamma_t^2).$$

That is, the error of the aggregate classifier g decreases exponentially fast with the number of combinations T!

(i) Let $Z_t := \sum_i D_{t+1}(i)$ (i.e., Z_t denotes the normalization constant for the weighted distribution D_{t+1}). Show that

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

(ii) Show that error of the aggregate classifier g is upper bounded by the product of Z_t: $\text{err}(g) \leq \prod_t Z_t$.

(hint: use the fact that 0-1 loss is upper bounded by exponential loss)

(iii) Show that $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$.

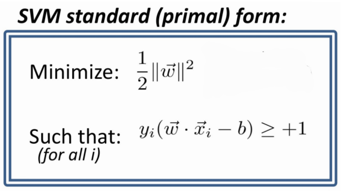
(hint: noting $Z_t = \sum_i D_t(i)\exp(-\alpha_t y_i f_t(x_i))$, separate the expression for correctly and incorrectly classified cases and express it in terms of ϵ_t)

(iv) By combining results from (ii) and (iii), we have that $\text{err}(g) \leq \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)}$, now show that:

$$\prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} = \prod_t \sqrt{1-4\gamma_t^2} \leq \exp(-2\sum_t \gamma_t^2).$$

Thus establishing that $\text{err}(g) \leq \exp(-2\sum_t \gamma_t^2)$.

3 1-Norm Support Vector Machine



SVM standard (primal) form:

Minimize: $\frac{1}{2}\|\vec{w}\|^2$

Such that: $y_i(\vec{w}\cdot\vec{x}_i - b) \geq +1$
(for all i)

(i) Recall the standard support vector machine formulation:

$$\begin{aligned} &\text{minimize} && \|\mathbf{w}\|_2^2 \\ &\text{subject to} && y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1, \qquad i = 1,\ldots,m, \end{aligned}$$

where m is the number of points and n is the number of dimensions, is a *quadratic program* because the objective function is quadratic and the constraints are affine. A *linear program* on the other hand uses only affine objective function and constraints, and is generally easier to solve than a quadratic program. By replacing the 2-norm in the objective function with the 1-norm ($\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j|$), we get

$$\begin{aligned} &\text{minimize} && \|\mathbf{w}\|_1 \\ &\text{subject to} && y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1, \qquad i = 1,\ldots,m. \end{aligned}$$

Note that the objective function here is not linear because there are absolute values involved. Show that this problem is equivalent to a linear program with $2n$ variables and $m+2n$ constraints.

(ii) The Chebyshev (ℓ_∞) distance between two points **x** and **y** is defined as $\max_i |x_i - y_i|$. Show that the 1-norm SVM maximizes the Chebyshev distance between the two separating hyperplanes $\mathbf{w}\cdot\mathbf{x} + w_0 = \pm 1$. (*Hint:* Show that the vector $(\text{sign}(w_1), \ldots \text{sign}(w_n))$ minimizes the l_∞ distance from the origin to the plane $\mathbf{w}\cdot\mathbf{x} = 2$.)

(iii) When the input data are not perfectly separable, we can apply a *soft-margin* approach (this is an alternative to the usual *slack-variables* approach discussed in class):

$$\text{minimize} \quad \|\mathbf{w}\|_1 + \sum_{i=1}^m [1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0)]_+ \,, \tag{1}$$

where $[\cdot]_+$ is the hinge loss function given by $\max(0,\cdot)$. Note that we've replaced the constraints with a penalty in the objective function.

Using the fact that strong duality always applies for linear programs, show that (1) can be expressed as

$$\begin{aligned} \text{maximize} \quad & \|\boldsymbol{\pi}\|_1 \\ \text{subject to} \quad & \left| \sum_{j=1}^{m} y_i x_{ij} \pi_i \right| \leq 1 \qquad j = 1, \ldots, n, \\ & \sum_{i=1}^{m} y_i \pi_i = 0, \\ & 0 \leq \pi_i \leq 1 \qquad\qquad i = 1, \ldots, n, \end{aligned}$$

where $\boldsymbol{\pi} \in \mathbb{R}^m$. (*Hint:* First express (1) as a linear program and then find its dual.)

(iv) Suppose we know that the output y depends only on a few input variables (i.e. the optimal **w** is sparse). Would the 1-norm or 2-norm SVM make more sense? Justify your answer.

4 Estimating Model Parameters for Regression

Let P_β be the probability distribution on $\mathbb{R}^d \times \mathbb{R}$ for the random pair (X, Y) (where $X = (X_1, \ldots, X_d)$) such that

$$X_1, \ldots, X_d \sim_{iid} N(0,1), \qquad \text{and} \qquad Y|X = x \;\sim\; N(x^\mathsf{T}\beta, \|x\|^2), \quad x \in \mathbb{R}^d$$

Here, $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ are the parameters of P_β, and $N(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2.

(i) Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be a given sample, and assume $x_i \neq 0$ for all $i = 1, \ldots, n$. Let f_β be the probability density for P_β as defined above. Define $Q : \mathbb{R}^d \to \mathbb{R}$ by

$$Q(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln f_\beta(x_i, y_i), \quad \beta \in \mathbb{R}^d.$$

Write a convex optimization problem over the variables $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ such that its optimal solutions are maximizers of Q over all vector of Euclidean length at most one.

(ii) Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be a given sample, and assume $x_i \neq 0$ for all $i = 1, \ldots, n$. Let f_β be the probability density for P_β as defined above. Define $Q : \mathbb{R}^d \to \mathbb{R}$ by

$$Q(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln f_\beta(x_i, y_i), \quad \beta \in \mathbb{R}^d.$$

Find a system of linear equations $A\beta = b$ over variables $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ such that the solutions are maximizers of Q over all vectors in $\mathbb{R}^d$.