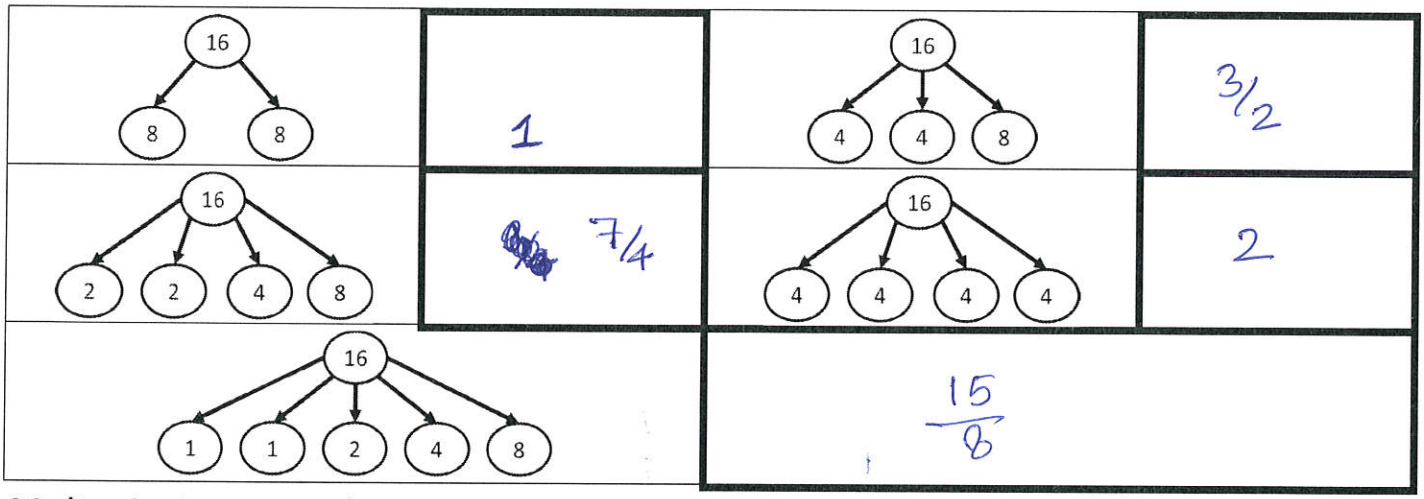
| CS 771A: Intro to Machine Learning, IIT Kanpur | | | | **Midsem Exam** (24 Sep 2022) |
|---|---|---|---|---|
| Name | JAYA GUPTA | | | **40 marks** |
| Roll No | 200471 | Dept. | CSE | Page **1** of **4** |

**Instructions**:

1. This question paper contains 2 page (4 sides of paper). Please verify.
2. Write your name, roll number, department in **block letters** with **ink** on **each page**.
3. If you don't do this, your pages may get lost when we unstaple your paper to scan pages
3. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
4. Don't overwrite/scratch answers especially in MCQ – such cases will get straight 0 marks.

**Q1.** For the hangman problem, 5 decision tree splits at a node are given. For each split, write down the information gain (entropy reduction) in the bold border boxes ~~border~~ next to the diagrams as a single fraction or decimal number. Use logarithms with base 2 in the definition of entropy. The numbers written in the nodes indicate how many words reached that node. **(5 marks)**




**Q2. (Intriguing entropy)** For a random variable $X$ with support $\{-1,1\}$ with $\mathbb{P}[X=1]=p$, define its entropy as $H(X) \stackrel{\text{def}}{=} -p\ln p-(1-p)\ln(1-p)$ (use natural logarithms for sake of simplicity). Find **(a)** a value of $p \in [0,1]$ where the entropy $H(X)$ is largest and **(b)** a value $p \in [0,1]$ where the entropy $H(X)$ is smallest. Show brief calculations/arguments for both parts. **(3 + 2 = 5 marks)**

(a) $H(X) = -p\ln p - (1-p)\ln(1-p)$

$\frac{dH}{dp} = -\ln p - 1 + \ln(1-p) + 1 \Rightarrow \ln(1-p) - \ln p$

$\frac{dH}{dp} = 0 \Rightarrow \ln(1-p) = \ln p \Rightarrow 1-p = p \Rightarrow \boxed{p = 1/2}$

$\frac{d^2H}{dp^2} \Rightarrow \frac{-1}{1-p} - \frac{1}{p} \Rightarrow \left|\frac{d^2H}{dp^2}\right| @ p = 1/2 \Rightarrow -2-2 \Rightarrow -4 < 0$

Hence $\boxed{p = 1/2}$ is point of maxima.

(b) from part a, it is seen $\frac{d^2H}{dp^2} < 0$ for $p \in [0,1]$, means $\frac{dH}{dp}$ is decreasing, which implies $H(X)$ is concave function.

If we find $\lim_{p\to 0} -p\ln p - (1-p)\ln(1-p) = \lim_{p\to 0} -p\ln p = 0$

Similarly $\lim_{p\to 1} -p\ln p - (1-p)\ln(1-p) = \lim_{p\to 1} -(1-p)\ln(1-p) = 0$

**Q3. (At a loss for names)** Consider the following loss function where $\tau > 0$ and $z \in \mathbb{R}$.

$$\ell_\tau(z) = \begin{cases} 1 - z & z < 1 - \tau \\ -\frac{(1-z)^4}{16\tau^3} + \frac{3(1-z)^2}{8\tau} + \frac{1-z}{2} + \frac{3\tau}{16} & z \in [1-\tau, 1+\tau] \\ 0 & z > 1 + \tau \end{cases}$$

1. Write down expressions for $\frac{d\ell_\tau(z)}{dz}$ and $\frac{d^2\ell_\tau(z)}{dz^2}$. No need to show calculations.
2. Write down an expression for $\nabla_{\mathbf{w}} f(\mathbf{w})$ where $\mathbf{w}, \mathbf{x}^i \in \mathbb{R}^d, y^i \in \{-1, +1\}$. You can use terms such as $\ell'_\tau(\cdot)$ in your expression to denote the first derivative of $\ell_\tau(\cdot)$ to avoid clutter.

$$f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \ell_\tau\left(y^i \cdot \mathbf{w}^\top \mathbf{x}^i\right)$$

3. Write down an expression for what the loss function $\ell_\tau(\cdot)$ would look like as $\tau \to 0^+$.

Give brief calculations for the 2nd part and brief justification for the 3rd part. **(2+2+3+1 = 8 marks)**

1) $\frac{d\ell_\tau(z)}{dz} = \begin{cases} -1 & z < 1-\tau \\ \frac{(1-z)^3}{4\tau^3} - \frac{3(1-z)}{4\tau} - \frac{1}{2} & z \in [1-\tau, 1+\tau] \\ 0 & z > 1+\tau \end{cases}$

$\frac{d^2\ell_\tau(z)}{dz^2} = \begin{cases} 0 & z < 1-\tau \\ -\frac{3(1-z)^2}{4\tau^3} + \frac{3}{4\tau} & z \in [1-\tau, 1+\tau] \\ 0 & z > 1+\tau \end{cases}$

2) $\boxed{\nabla_w f(w) = w + \sum_{i=1}^{n} \ell'_\tau(y^i \cdot w^\top x^i)\, y^i \cdot x^i}$

Consider $(y^i \cdot w^\top x^i) = z$ $\quad \sum_{i=1}^{n} \ell_\tau(z) = g(z)$

By using chain rule. $\nabla g(z) = \sum_{i=1}^{n} \ell'_\tau(z) \cdot \nabla_w(z)$

$= \sum_{i=1}^{n} \ell'_\tau(y^i \cdot w^\top x^i)\, y^i \cdot x^i$

3) for $z \in [1-\tau, 1+\tau]$.

$\ell_\tau(z) = \tau$ for $z = 1-\tau$.

$0$ for $z = 1+\tau$.

$\ell_{\tau\to 0}(z) = \begin{cases} 1-z & z < 1 \\ 0 & z = 1 \\ 0 & z > 1 \end{cases}$

As $\tau \to 0^+$, $z \in [1]$, $\boxed{\ell_\tau(z) \to 0}$

As we can see, $\ell_\tau(z)$ is a continuous func. on $z$, so as $\tau \to 0^+$, $z \in [1-\tau, 1+\tau] \to 1$ ✓ $z$ in range tends to 1. $\ell_\tau(z) = 0, z > 1$, $\ell_\tau(z) = 0, z < 1$ ↓ so for continuity $\ell_\tau(z) = 0$ for $z = 1$

**Q4. (A regularized median)** Given a set of real numbers $a^1, a^2, \ldots, a^n \in \mathbb{R}$ (all are distinct but may be positive, negative or zero), we wish to find its "regularized median" by solving: $\min_x \frac{1}{2}x^2 + \sum_{i=1}^{n}|x - a^i|$. However, to design a solver, it would be helpful if we first rewrite this objective function as shown on the right-hand side by artificially introducing constraints

$$\min_{x,c^i} \frac{1}{2}x^2 + \sum_{i=1}^{n} c_i$$
$$\text{s.t.} \quad x - a^i \le c^i$$
$$x - a^i \ge -c^i$$
$$c^i \ge 0$$

| CS 771A: Intro to Machine Learning, IIT Kanpur | | | | **Midsem Exam** (24 Sep 2022) |
|---|---|---|---|---|
| Name | JAYA GUPTA. | | | **40 marks** |
| Roll No | 200471 | Dept. | CSE | Page **3** of **4** |

1. Write down the Lagrangian of this problem by introducing dual variables for the constraints.
2. Using the Lagrangian, create the dual problem (show brief derivation). Simplify the dual as much as you can otherwise the next part may get more cumbersome for you.
3. Give an expression for deriving the primal solution $x$ in terms of the dual variables
4. Give peudocode for a coordinate ascent/descent method to solve the dual. Use any coordinate selection method you like. Give precise expressions in pseudocode on how you would process a chosen coordinate taking care of constraints. **(2 + 4 + 2 + 4 = 12 marks)**

1). $\min_{x,\{c^i\}} \left\{ \max_{\{\alpha^i\},\{\beta^i\},\{\gamma^i\}} \left\{ \frac{1}{2}x^2 + \sum_{i=1}^n c_i^i + \sum_{i=1}^n \alpha_i(x - a^i - c^i) + \sum_{i=1}^n \beta_i(a^i - x - c^i) + \sum_{i=1}^n \gamma_i(-c^i) \right\} \right\}$

$\alpha_i \geq 0$, $\beta_i \geq 0$, $\gamma_i \geq 0$

2). $\max_{\{\alpha^i\},\{\beta^i\},\{\gamma^i\}} \left\{ \min_{x,\{c^i\}} \frac{1}{2}x^2 + \sum_{i=1}^n c_i^i + \sum_{i=1}^n \alpha_i(x - a^i - c^i) + \sum_{i=1}^n \beta_i(a^i - x - c^i) - \sum_{i=1}^n \gamma_i c^i \right\}$

$\alpha, \beta, \gamma \geq 0$

$f(x, c^i)$

$\frac{df}{dx} = x + \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \beta_i = 0 \Rightarrow \boxed{x = \sum_{i=1}^n (\beta_i - \alpha_i)}$

$\forall i, \frac{df}{dc^i} = 1 - \alpha_i - \beta_i - \gamma_i = 0 \Rightarrow \boxed{\alpha_i + \beta_i + \gamma_i = 1. \ \forall i}$

$\max_{\{\alpha_i\},\{\beta_i\},\{\gamma_i\}} \left\{ \frac{1}{2}x^2 + \sum_{i=1}^n [a^i(\beta_i - \alpha_i) + x(\alpha_i - \beta_i)] \right\}$

$\boxed{\max_{\{\alpha_i\},\{\beta_i\}, \alpha,\beta \geq 0} \left\{ \sum_{i=1}^n a^i(\beta_i - \alpha_i) - \frac{1}{2}\left(\sum_{i=1}^n (\beta_i - \alpha_i)\right)^2 \right\}}$

**OR**

$\max_{\{\alpha_i,\beta_i\}, \alpha,\beta \geq 0} \left\{ \sum_{i=1}^n a^i(\beta_i - \alpha_i) - \frac{x^2}{2} \right\}$

3) From here it can be seen, $\boxed{x = \sum_{i=1}^n (\beta_i - \alpha_i)}$

4) It is a maximization prblm, we use coordinate ascent.

Initialise $\alpha_i, \beta_i$ randomly $\forall i$.

Choose using cyclic coordinate/random selection $(\alpha_1, \alpha_2 \cdots \alpha_n, \beta_1 \cdots, \beta_n)$. $\leq m$

If $m = \alpha_i$:

$\alpha_i^{t+1} \leftarrow [\alpha_i^t + \eta_{\alpha_i}(-a^i + \beta_i^t - \alpha_i^t)]_+$

$\alpha_j^{t+1} \leftarrow \alpha_j^t \quad \forall j \neq i$

$\beta_k^{t+1} \leftarrow \beta_k^t \quad \forall k$

OR

If $m = \beta_i$:

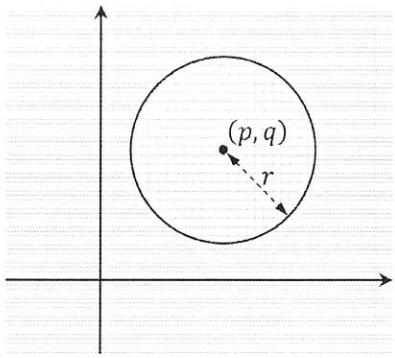
$\beta_i^{t+1} \leftarrow [\beta_i^t + \eta_{\beta_i}(a^i - \beta_i^t + \alpha_i^t)]_+$

$\beta_j^{t+1} \leftarrow \beta_j^t \quad \forall j \neq i$

$\alpha_k^{t+1} \leftarrow \alpha_k^t \quad \forall k$.

$\alpha_i \geq 0, \beta_i \geq 0$ Repeat till convergence.

**Q5. (Circular argument)** Given a circle in 2D plane with centre at $\mathbf{c} = (p, q) \in \mathbb{R}^2$ and radius $r > 0$, we wish to build a classifier that gives output $y = -1$ if a point $\mathbf{x} = (x, y) \in \mathbb{R}^2$ is inside the circle i.e., $(x-p)^2 + (y-q)^2 < r^2$ and $y = +1$ otherwise. Do not worry about points on the boundary. Give a feature map $\phi: \mathbb{R}^2 \to \mathbb{R}^D$ for some $D > 0$ and a corresponding classifier $\mathbf{W} \in \mathbb{R}^D$ such that for any $\mathbf{x} \in \mathbb{R}^2$, $\text{sign}(\mathbf{W}^\top \phi(\mathbf{x}))$ is the correct output. Your map $\phi$ must **not depend on** $p, q, r$ but your classifier $\mathbf{W}$ may depend on $p, q, r$. **(2 + 2 = 4 marks)**




we know for a circle, any point inside the circle $C(p,q,r) < 0$ and for pt. outside $C(p,q,r) > 0$.

$C(p,q,r) = (x-p)^2 + (y-q)^2 - r^2$

$\rightarrow x^2 + y^2 - 2px - 2qy + p^2 + q^2 - r^2$

$\phi: R^2 \to R^d$

s.t. $\phi\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x^2 \\ y^2 \\ x \\ y \end{bmatrix}$ $\begin{bmatrix} x \\ y \end{bmatrix} \in R^2$

Here $D = 4$

$\rightarrow$ corresponding classifier is

$W = \begin{bmatrix} 1 \\ 1 \\ -2p \\ -2q \end{bmatrix}$ $b = p^2 + q^2 - r^2$.

we can hide bias inside and add 1 dim. to feature.

$\phi\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = [x^2 \; y^2 \; x \; y \; 1]^T$ $W = [1 \;\; 1 \;\; (-2p) \;\; (-2q) \;\; (p^2+q^2-r^2)]^T$

**Q6.** Melbo has learnt a decision tree to solve a binary classification problem with 95 train points of which it gets 48 correct and 47 wrong. There are 10 real features for every training point and the first feature $x_1$ is an interesting one. $x_1$ takes only 3 values, namely 0, 1, 2. Among the train points that Melbo classified correctly, a 5/12 fraction had $x_1 = 0$, a 1/6 fraction had $x_1 = 1$ and the rest had $x_1 = 2$. Melbo got 2/3rd of the training points that had $x_1 = 0$ wrong. Melbo got 1/5th of the training points that had $x_1 = 1$ wrong and 1/5th of the training points that had $x_1 = 2$ wrong. Find out how many train points had the feature value $x_1 = 0$, $x_1 = 1$ and $x_1 = 2$.

**(Bonus)** Do you notice anything funny about the way Melbo's decision tree gets answers right and wrong? Can you improve its classification accuracy on the training set? You cannot change the decision tree itself, but you can take the decision tree output on a data point and the value of the first feature for that data point $x_1$ and possibly change the output. What is the improved accuracy? For this part, you may assume that the binary labels are $+1$ and $-1$. **(2 x 3 = 6 + 3 marks)**

$T = 95$, $C = 48$, $W = 47$.

$N[x_1 = 0 \mid C]$ $\rightarrow$ means number of training pts having $x_1 = 0$ which melbo classified correctly.

$N[x_1 = 0 \mid C] = \frac{5}{12} \times 48 = 20$ $N[x_1 = 1 \mid C] = \frac{1}{6} \times 48 = 8$ $N[x_1 = 2 \mid C] = 20$

Let number of data points having $x_1 = 0$ be $N_0$, $x_1 = 1$ be $N_1$, $x_1 = 2$ be $N_2$.

$\frac{2}{3} N_0 \rightarrow$ Melbo got wrong $\Rightarrow$ $\frac{N_0}{3}$ melbo got correct $(x_1 = 0) \Rightarrow \frac{N_0}{3} = 20 \Rightarrow \boxed{N_0 = 60}$

$\frac{1}{5} N_1 \rightarrow$ melbo got wrong $\Rightarrow$ $\frac{4}{5} N_1 = 8 \Rightarrow \boxed{N_1 = 10} \rightarrow \boxed{x_1 = 1}$

Num-pts $(x_1 = 0)$

$\frac{1}{5} N_2 \rightarrow$ " " " $\Rightarrow$ $\frac{4}{5} N_2 = 20 \Rightarrow \boxed{N_2 = 25} \rightarrow \boxed{x_1 = 2}$.