Introduction to Machine Learning – COS 324, Fall 2017

Midterm exam

Name:

NetID:

Please state and sign the Princeton honor code:

1. Online Learning.

You consult with n experts in order to make accurate predictions. You were told that the best expert will make *at most* 1 mistake.

(i) Describe the possible values of the weight vectors that can be attained by the Halving algorithm after $T \geq n$ rounds.

(ii) Construct a sequence of experts' predictions and actual outcomes that force the Consistent algorithm to make T mistakes for any $T > 0$. You may choose any n.

(iii) Recall that $\forall T > 0$ and $i \in [n]$ the Weighted Majority algorithm is guaranteed to make at most $2(1+\eta)L_i^T + \frac{2\log(n)}{\eta}$ mistakes where $0 < \eta \leq \frac{1}{2}$. Derive a mistake bound for the setting of this question.

SOLUTION

(i) An expert's weight is 1 if it hasn't made any mistake in the T rounds, and it is 0 otherwise.

(ii) Choose $n = T$. Assuming the Consistent alogritm starts with the first expert ($w_1 = 1$, and $w_i = 0$ for $0 \leq i \leq n$). The algorithm always predicts $\hat{y}_t = 1$ while the outcome is always $y_t = -1$. The algorithm makes a mistake at each round, and it takes T rounds to prune all inconsistent expert. So the algorithm makes T mistakes in T rounds.

(iii) We know that $L_{i*} = \min_{1\leq i \leq n} L_i^T \leq 1$.

$$L^T \leq 2(1+\eta)L_{i*}^T + \frac{2\log(n)}{\eta} = 2L_{i*}^T + 2(\eta L_{i*}^T + \frac{\log(n)}{\eta})$$

The tightes upperbound could have be achieved by $\eta^* = \sqrt{\frac{\log(n)}{L_{i*}}}$ (for non-zero L_{i*}). However, we need to be careful whether wuch eta^* lies in $(0, 1/2]$.

1) if $\sqrt{\frac{\log(n)}{L_{i*}}} \leq 1/2$, (or $L_{i*} = 0$)

$$L^T \leq 2L_{i*}^T + 4\sqrt{L_{i*}^T \log(n)} \leq 2 + 4\sqrt{\log(n)}$$

2) if $\sqrt{\frac{\log(n)}{L_{i*}}} \geq 1/2$, (or $L_{i*} = 0$)

$$L^T \leq 2L_{i*}^T + 2(\frac{1}{2}L_{i*}^T + 2\log(n)) \leq 3 + 4\sqrt{\log(n)}$$

Derive a mistake bound for the setting of this question.

2. Convex Analysis.

Consider the function $f(x, y) = x^4 + y^4$ on the domain

$$\mathcal{K} = \{(x, y) : -10 \leq x \leq 10, -10 \leq y \leq 10\} .$$

(i) Prove or disprove: f is convex.
(ii) Prove or disprove: $\mathcal{K}$ is a convex set.
(iii) Compute the gradient of f. What is the smallest L such that f is L-Lipschitz?
(iv) Is f β-smooth for some $\beta < \infty$? If so, find the smallest such β.
(v) Is f α-strongly convex for some $\alpha > 0$? If so, find the largest such α.

SOLUTION

(i) f is convex. A twice-differentiable function is convex if and only if its hessian is positive semi-definite and its domain is convex. The convexity of f's domain $\mathcal{K}$ is proved in (ii). Now we Check the hessian of f

$$\nabla^2 f(x, y) = \begin{bmatrix} 12x^2 & 0 \\ 0 & 12y^2 \end{bmatrix}$$

Since x^2 and y^2 are non-negative on the domain $\mathcal{K}$, the hessian is positive semi-definite on $\mathcal{K}$. Thus f is a convex function

(ii) $\mathcal{K}$ is a convex set. For any two points $(x_1, y_1), (x_2, y_2) \in \mathcal{K}$ and any $0 \leq t \leq 1$, we know that

$$-10 \leq -10t - 10(1-t) \leq tx_1 + (1-t)x_2 \leq 10t + 10(1-t) \leq 10$$

$$-10 \leq -10t - 10(1-t) \leq ty1 + (1-t)y2 \leq 10t + 10(1-t) \leq 10$$

indicating that $t(x_1, y_1) + (1-t)(x_2, y_2) \in \mathcal{K}$. Thus $\mathcal{K}$ is convex.

(iii) The gradient of f is

$$\nabla f(x, y) = \begin{bmatrix} 4x^3 \\ 4y^3 \end{bmatrix}$$

$$L = \max_{(x,y)\in\mathcal{K}} ||\nabla f(x, y)||_2 = \max_{(x,y)\in\mathcal{K}} 4\sqrt{x^6 + y^6} = 4\sqrt{10^6 \times 2} = 4000\sqrt{2}$$

(iv) Yes, For any $(x, y) \in \mathcal{K}$

$$\nabla^2 f(x, y) = \begin{bmatrix} 12x^2 & 0 \\ 0 & 12y^2 \end{bmatrix} \preceq 12 \times 10^2 I$$

Thus the smallest β is 1200.

(v) No. Since at point $(0,0)$,

$$\nabla^2 f(0,0) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \geq 0I$$

there is no such $\alpha > 0$.

3. PAC LEARNING.

You need to build an apple selection machine, which measures an apple's weight in grams and the radius of the smallest enclosing sphere in millimeters. The weight of any apple you'd encounter is between 100 and 600 grams, inclusive. The radius is between 200 and 1200 millimeters, inclusive. The measuring device does *not* have fractional units. A good apple is designated as having a positive label, and a bad one as negative. You are requested to find a classification rule of the form: "If the weight is measured to be $\geq a$ *or* the radius is measured to be $\geq b$, then the apple is good. Otherwise, it is considered bad."

(i) Describe formally the hypothesis class you need to employ. How many different predictors are in the class?

(ii) You were told that there exists a rule specified by $(a^\star, b^\star)$ that would perfectly classify all apples. What is the smallest number of examples you would need in order to find a predictor which is correct on 95% of the apples? Your procedure for finding such a classifier may completely fail with probability of at most 1%.

($\star$ bonus) Can you still find a 95% accurate classifier if you were told instead that there exists a classifier $(a^\star, b^\star)$ with a 2% error rate? If yes, what is the number of examples that will be required to find one with the same failure probability of 1%?

SOLUTION

(i)

$$\mathcal{H} = \{h_{a,b} \mid h_{a,b}(w,r) = \begin{cases} 1, & \text{if } w \geq a \text{ and } r \geq b \\ -1, & \text{otherwise} \end{cases}, 100 \leq a \leq 600, 200 \leq b \leq 1200, a, b \in \mathbb{N}\}$$

$$|\mathcal{H}| = 501 \times 1001 = 501501$$

(ii) $\varepsilon = 0.05, \delta = 0.01,$

$$m \geq \frac{\log(|\mathcal{H}|) + \log(1/\delta)}{\varepsilon} = \frac{\log(501501) + \log(100)}{0.05}$$

(bonus) $\varepsilon = 0.05 - 0.02 = 0.03, \delta = 0.01,$ using agnostic PAC bound

$$m \geq \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2} = \frac{2\log(2 \times 501501 \times 100)}{0.03^2}$$