AI3001: Advanced Topics in Machine Learning
Homework 1
August 26, 2022

Instructions:

- *The maximum number of points is 60. Each point is worth 0.25 marks.*
- *Problems 1, 2 and 3 are compulsory. You need to solve any one of the problems between 4 and 5. If you solve both, the maximum of the two will be considered in your score. Problem 6 does not carry any points.*
- *The last date of submission is end of day **Saturday, 9th Sept 2022**. You can drop either your handwritten homework in my office (C block, room:112/D) by sliding it under the door or hand it over after the class. If you are writing homework in latex/MS word, send the pdf document by email. Please do not scan and send handwritten pages.*
- *Late submissions will incur a penalty of 4 points per day (irrespective of the time of the day).*

Problem 1 (10 pts)**.** Show that a straightforward extension of the MAJORITY algorithm makes at-most $((m+1)\log_2(N))$ mistakes when the best expert makes $m \geq 0$ mistakes.

Problem 2 (10 pts)**.** In this question we will explore the class of instances where NAIVE algorithm makes lesser mistakes than the MAJORITY. Note that for binary i.e. 1-bit prediction with N experts, the mistake bounds for NAIVE algorithm and MAJORITY algorithm are $N-1$ and $\log(N)$ respectively.

Consider a k-bit prediction problem where the decision space is $\{0,1\}^k$ i.e. each expert predicts a k length binary sequence. Also, assume that there exists a perfect expert.

1. What is the mistake bound of the NAIVE algorithm for k-bit prediction problem. (1pt)
2. Consider the following algorithm (call it ALG).

$$p_t \in \arg\max_{s_k \in S_k} |\{j \in [N] : f_{j,t} = s_k\}|$$

Here, S_k denote the set of all possible k length binary sequences. Ties are broken arbitrarily.

   (a) Show that ALG reduces to the MAJORITY algorithm for binary prediction problem i.e. when $k = 1$. (1pt)

   (b) What is the worst case mistake bound of ALG for k-bit expert advice problem. (3pt)

   (c) There exist k_0 above which the mistake bound calculated in question 1 above is lower than the one for ALG calculated in Q2.b. Prove or disprove. (5pts)

Problem 3 (20 pts)**.** In this problem, we will give an alternate proof for the regret upper bound of EXPWTS. Consider the same setting studied in the class (refer to the lecture notes (Lecture 5) for details).

1. Let $X \in [0,1]$ be a random variable. Show that

$$\log(\mathbb{E}[e^{-\eta X}]) \leq (e^{-\eta} - 1)\mathbb{E}[X]. \quad \text{(5pts)}$$

   Hint: Use convexity of exponential function and inequality $\log(1+x) \leq x$.

2. Using the above bound in the proof of exponential weights algorithm, show that

$$\sum_{t=1}^{T} \ell(p_t, y_t) \leq \frac{\eta}{1-e^{-\eta}} \sum_{t=1}^{T} \ell(f_{i,t}, y_t) + \frac{\log(N)}{1-e^{-\eta}} \quad \forall i = 1, 2, \cdots N. \quad \text{(10pts)}$$

   .

3. Show that the above bound gives the regret upper bound of $\sqrt{\frac{T\log(N)}{2}}$. (5pts)

Problem 4 (5*4 = 20 pts)**.** In this question we will revise the concepts from convex optimization. Prove (or disprove) the below statements.

1. KL Divergence is jointly convex in both its inputs i.e. when the tuple (P,Q) is considered as an input. [1]
2. Square loss is 2-strongly convex and $1/2$-exp concave.
3. LogSumExp (LSE) function is convex but not strictly convex.[2]
4. Let the function f be α_1-strongly convex and g be α_2-strongly convex then f+g is $\alpha_1 + \alpha_2$-strongly convex.
5. Let a differentiable function is α-exp concave for some $\alpha > 0$ then show that the function is convex.

Problem 5 (Programming question (20pts))**.** Download the historical data for Apple Inc. and Tesla Inc. for the past 5 years from the shared drive. Assume that we begin with initial wealth of 1000$ on the first day of trading (23rd Aug 2017). Plot the following datapoints for every 6 months starting from the first day of trading.

1. wealth generated by following 11 buy-and-hold strategies with parameters $(i/10, (10-i)/10)$ for $i = 0, 1, 2 \cdots 10$. (2 pts),
2. wealth generated by discretized version of Cover's universal portfolio algorithm with following parameter setting.
   (a) There are $m+1$ CRP strategies with parameters $(i/m, (m-i)/5)$ for strategies $i = 0, 1, 2, \cdots, m$ and the initial wealth is distributed among all m CRPs uniformly. Choose $m = 51, 101$ and 1001. (4pt)

[1] Given two probability distributions P and Q supported over a finite support $\mathcal{X}$, the KL divergence is defined as

$$KL(P|Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

You have to show that $KL(\alpha P_1 + (1-\alpha)P_2 | \alpha Q_1 + 1 - \alpha)Q_2) \leq \alpha KL(P_1|Q_q) + (1-\alpha)KL(P_2|Q_2)$.

[2] Log-sum-exp is defined as

$$LSE(a_1, a_2, \cdots a_n) = \log(\sum_{i=1}^{n} e^{a_i})$$

(b) Repeat the above experiments with Dirichlet distribution (Refer [3] for details of Dirichlet distribution) with parameters $(1/2, 1/2, \cdots, 1/2)$ as initial distribution. (4 pt)

3. What are your insights/conclusions? Be brief. (2pts)

Problem 6 (Challenge question)**.** The mind reader game (link in footnote [4]) implements the exponential weights algorithm as a meta algorithm to aggregate the predictions from 26 different experts algorithms. These expert algorithms are designed to identify patterns and implement user reaction to different past outcomes to make their predictions (refer to the detailed report here[5]). The goal is to come up with a binary input sequence to beat the algorithm. The matlab file with the code is also available online.

[3] https://en.wikipedia.org/wiki/Dirichlet_distribution

[4] https://web.media.mit.edu/~guysatat/MindReader/index.html

[5] https://web.media.mit.edu/~guysatat/MindReader/GuySatat_MindReader_FinalProject.pdf