Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Κυρτή Βελτιστοποίηση με Εφαρμογές στη Μηχανική Μάθηση

Διδάσκοντες: Κ. Χρυσαφίνος, Δ. Φωτάκης

2η Σειρά Ασκήσεων, Ημ/νια Παράδοσης: 15/6/2021

Άσκηση 1. Θεωρούμε τον Halving Learner (ή τον Consistent Learner), στο 1ο set διαφανειών της 1ης διάλεξης (διαφ. 12 και 10), οι οποίοι προσπαθούν να ελαχιστοποιήσουν το πλήθος των λαθών (επί τη ευκαιρία, είναι ενδιαφέρον να παρατηρήσετε τη σχέση μεταξύ αυτών των αλγορίθμων και του Online Loss Minimization framework). Να δείξετε ότι αν τα δείγματα έρχονται από μια οποιαδήποτε (άγνωστη, αλλά συγκεκριμένη) κατανομή $\mathcal{D}$ και το σύνολο έγκυρων υποθέσεων V_t δεν μεταβάλλεται για $\Omega(\log(1/\delta)/\varepsilon)$ συνεχόμενα δείγματα, τότε με πιθανότητα τουλάχιστον $1-\delta$, κάθε έγκυρη υπόθεση $h \in V_t$ επιτυγχάνει loss $L_{(\mathcal{D},h^*)}(h) \le \varepsilon$ (δηλαδή έχουμε επιτύχει το guarantee του PAC Learning – εδώ υποθέτουμε realizability και h^* είναι μια υπόθεση με μηδενικό λάθος). Ποια είναι η δειγματική πολυπλοκότητα που επιτυγχάνουν οι δύο αλγόριθμοι για μια πεπερασμένη κλάση υποθέσεων $\mathcal{H}$;

Άσκηση 2. Να λύσετε την άσκηση **3.5.2**, σελ. 50, του βιβλίου Understanding Machine Learning.

Άσκηση 3. Να λύσετε την άσκηση **3.5.3**, σελ. 51, του βιβλίου Understanding Machine Learning.

Άσκηση 4. (α) Να υπολογίσετε το VC dimension των παρακάτω κλάσεων συναρτήσεων:

1. $\mathcal{H}_1 = \{f_{\theta_1,\theta_2,\theta_3,\theta_4} : [0,1] \to \{0,1\} \,|\, 0 < \theta_1 < \theta_2 < \theta_3 < \theta_4 < 1\}$,
   όπου $f_{\theta_1,\theta_2,\theta_3,\theta_4}(x) = \begin{cases} 1 & \text{αν } \theta_1 \le x \le \theta_2 \text{ ή } \theta_3 \le x \le \theta_4 \\ 0 & \text{διαφορετικά} \end{cases}$
2. $\mathcal{H}_2$ που περιλαμβάνει όλες τις συναρτήσεις $f : \mathbb{R}^3 \to \{0,1\}$ που κατηγοριοποιούν με βάση κάποιο ορθογώνιο παραλληλεπίπεδο παράλληλο προς τους άξονες (θεωρήστε ότι τα σημεία εντός του παραλληλεπιπέδου αποτιμώνται στο 1 και τα σημεία εκτός αποτιμώνται στο 0).
3. $\mathcal{H}_3 = \{f_{\theta_1,\theta_2} : \mathbb{R}^2 \to \{0,1\} \,|\, 0 < \theta_1 < \theta_2\}$, όπου $f_{\theta_1,\theta_2}(x,y) = \begin{cases} 1 & \text{αν } \theta_1 x \le y \le \theta_2 x \\ 0 & \text{διαφορετικά} \end{cases}$
4. $\mathcal{H}^k = \{f : \mathcal{X} \to \{0,1\} \,|\, |\{x \in \mathcal{X} : f(x) = 1\}| \le k\}$ (δηλαδή $\mathcal{H}^k$ είναι η κλάση υποθέσεων που αναθέτουν label 1 σε k το πολύ δείγματα του $\mathcal{X}$), όπου $\mathcal{X}$ ένα πεπερασμένο domain. Εξαρτάται το VC dimension της $\mathcal{H}^k$ από τον πληθικό αριθμό του $\mathcal{X}$, και αν ναι, πώς;

(β) Να δείξετε ότι:

1. Για κάθε κλάση υποθέσεων $\mathcal{H}$ και για κάθε υπόθεση $f : \mathcal{X} \to \{0,1\}$, VCdimension$(\mathcal{H} \cup \{f\}) \le$ VCdimension$(\mathcal{H}) + 1$.
2. Για κάθε δύο κλάσεις υποθέσεων $\mathcal{H}' \subseteq \mathcal{H}$, ισχύει ότι VCdimension$(\mathcal{H}') \le$ VCdimension$(\mathcal{H})$.

Άσκηση 5. Να αποδείξετε ένα (όσο το δυνατόν καλύτερο) άνω φράγμα για το regret του αλγόριθμου Online Gradient Descent (όπως παρουσιάζεται στις διαφάνειες της 4ης διάλεξης) λαμβάνοντας υπόψη την προβολή στο κυρτό σύνολο των εφικτών λύσεων S και χρησιμοποιώντας χρονικά μεταβαλλόμενο βήμα $\eta_t = \frac{B}{G\sqrt{t}}$.

Άσκηση 6. Στην μέθοδο Support Vector Machines (SVM), χρησιμοποιούμε ως συνάρτηση κόστους (loss function) την $\ell_{(\vec{x},y)}(\vec{w}) = \max\{0, 1 - y\,\vec{w}\cdot\vec{x}\}$ και υπολογίζουμε τον classifier $\vec{w}$ που ελαχιστοποιεί την παρακάτω συνάρτηση regularized loss:

$$f_{(\vec{x}_1,y_1),\ldots,(\vec{x}_n,y_n)}(\vec{w}) = \frac{\lambda}{n}\sum_{i=1}^{n}\ell_{(\vec{x}_i,y_i)}(\vec{w}) + \frac{\|\vec{w}\|^2}{2}\,,$$

όπου λ είναι το regularization factor και $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ τα δείγματα. Να εξειδικεύσετε τον αλγόριθμο Stochastic Gradient Descent για την μέθοδο Support Vector Machines. Τι βήμα η (ή η_t) θα χρησιμοποιήσετε και τι regret θα επιτύχετε με αυτό;

Άσκηση 7. Θεωρούμε τον παρακάτω αλγόριθμο A για Online Linear Optimization, για τον οποίο θα αναλύσουμε το regret που επιτυγχάνει.

- **Input**: n actions $\{1, \ldots, n\}$, time horizon T, $\vec{w}_1 = (1, \ldots, 1)$, $\vec{x}_1 = (1/n, \ldots, 1/n)$
- for $t = 1$ to T do:
  - Select action $i_t \in \{1, \ldots, n\}$ with probability $\vec{x}_t(i_t)$
  - Get loss $\vec{\ell}_t \in [0, 1]^n$ for all actions and incur loss $\vec{\ell}_t(i_t)$
  - Update weights $\vec{w}_{t+1}(i) = \vec{w}_t(i)e^{-\eta\vec{\ell}_t(i)}$, for all $i \in \{1, \ldots, n\}$
  - Update probabilities $\vec{x}_{t+1}(i) = \frac{\vec{w}_{t+1}(i)}{\sum_{i=1}^{n} w_{t+1}(i)}$, for all $i \in \{1, \ldots, n\}$

(α) Θεωρούμε τη συνάρτηση δυναμικού $\Phi(t) = \sum_{i=1}^{n} w_t(i)$. Αρχικά είναι $\Phi(1) = n$. Να δείξετε ότι $\Phi(T) \geq e^{-\eta\sum_{t=1}^{T}\vec{\ell}_t(i^*)}$, όπου $i^* = \arg\min_i \sum_{t=1}^{T}\vec{\ell}_t(i)$ η βέλτιστη επιλογή.
(β) Να δείξετε ακόμη ότι για κάθε $t \geq 1$,

$$\Phi(t+1) \leq \Phi(t)e^{-\eta\vec{x}_t\cdot\vec{\ell}_t + \eta^2\vec{x}_t\vec{\ell}_t^2}\,,$$

όπου $\vec{\ell}_t^2(i) = (\vec{\ell}_t(i))^2$, για όλα τα $i \in \{1, \ldots, n\}$. Χρησιμοποιώντας αυτή τη σχέση, να βρείτε ένα άνω φράγμα για το $\Phi(T)$ ως συνάρτηση του $\Phi(1) = n$.
(γ) Χρησιμοποιώντας τα άνω και κάτω φράγματα για την $\Phi(T)$ που υπολογίσατε στα (α) και (β) και το γεγονός ότι $\vec{\ell}_t \in [0, 1]^n$, για κάθε $t \in \{1, \ldots, T\}$, να δείξετε ότι:

$$\text{Regret}_A(T) \leq \eta T + \ln(n)/\eta$$

Ποια τιμή του η θα επιλέγατε και ποιο είναι το $\text{Regret}_A(T)$ για αυτή;

Υποβολή. Οι εργασίες πρέπει να αναρτηθούν στο `https://courses.corelab.ntua.gr/convex-opt` μέχρι τα μεσάνυχτα της Τρίτης 15 Ιουνίου (παραμονή της ημέρας της εξέτασης).