Computational Learning Theory

Sommersemester 2018

Mario Holldack, M. Sc.
Prof. Dr. Georg Schnitger
Hannes Seiwert, M. Sc.

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Institut für Informatik
AG Theoretische Informatik

Übungsblatt 8

Ausgabe: 04.06.2018
Abgabe: 11.06.2018

Aufgabe 8.1 *Online Convex Optimization* ((4+8+4) + (10+6) Punkte)

Eine wichtige Klasse im Online-Lernen ist die Optimierung konvexer Funktionen, die vielfältige Anwendungen von linearer Regression über Allokationsprobleme bis hin zu neuronalen Netzen hat. Wir haben das folgende Online-Szenario vorliegen.

Eine *konvexe* Menge $\mathcal{H} \subseteq \mathbb{R}^d$ von Hypothesen sei vorgegeben. In Schritt $i = 1, 2, \ldots, T$:

- Der Schüler wählt eine Hypothese $w_i \in \mathcal{H}$.
- Der Lehrer präsentiert eine *konvexe* Loss-Funktion $\ell_i : \mathcal{H} \to \mathbb{R}$.
- Der Schüler „bezahlt“ Loss $\ell_i(w_i)$.

Dabei darf die in Schritt i gewählte Hypothese w_i von allen bisherigen Hypothesen $w_1, \ldots, w_{i-1}$ sowie den bisher beobachteten Loss-Funktionen $\ell_1, \ldots, \ell_{i-1}$ abhängen.

> **Einschub:** Bevor wir das oben vorgestellte Szenario weiter untersuchen, übersetzen wir als Beispiel die linearen Regression in unser Online-Szenario:
> Hypothesen $w \in \mathcal{H}$ entsprechen hier den linearen Funktionen $f_w : \mathbb{R}^d \to \mathbb{R}$ mit $f_w(x) := \langle w, x \rangle$.
> Der Lehrer präsentiert in Schritt i den Messpunkt $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$. Die Aufgabe des Schüler ist die Minimierung des quadratischen Losses $\sum_{i=1}^{T} \ell(w_i, x_i, y_i) = \sum_{i=1}^{T} \|f_{w_i}(x_i) - y_i\|^2$.
> Die Loss-Funktion ℓ_i in Schritt i entspricht dem Beitrag des i-ten Messpunktes zum Gesamtloss $\ell_i(w) := \ell(w, x_i, y_i) = (f_w(x_i) - y_i)^2$. Beachte auch, dass ℓ_i konvex ist.

Ziel des Schülers ist es, den über alle Zeitschritte insgesamt gezahlten Loss zu minimieren. Wir bewerten die Qualität eines Schülers durch seinen *Regret*: Für eine Hypothese $w^* \in \mathcal{H}$ definiere:

$$\text{Regret}_T(w^*) = \sum_{i=1}^{T} (\ell_i(w_i) - \ell_i(w^*)) \quad \text{und} \quad \text{Regret}_T = \sup_{w^* \in \mathcal{H}} \text{Regret}_T(w^*)$$

Wir vergleichen also den Gesamtloss des Schülers mit dem Gesamtloss einer *besten* Hypothese. Ein *guter* Schüler sollte gegen jede Folge von Loss-Funktionen (innerhalb einer vorgegebenen Klasse) einen möglichst geringen, sublinearen $\text{Regret}_T = o(T)$ erreichen.
Zur Einfachheit nehmen wir im Folgenden $d = 1$ an, d. h. wir arbeiten mit eindimensionalen Hypothesen $\mathcal{H} \subseteq \mathbb{R}$.

a) Wir betrachten einen ersten einfachen Algorithmus: **Follow the Leader (FTL)** wählt in jedem Schritt i (irgend)eine Hypothese $w_i \in \mathcal{H}$, die *bisher* den geringsten Loss hat:

   - In Schritt $i = 0$ wähle $w_1 = 0$ (Es gelte $0 \in \mathcal{H}$.)
   - In Schritt $i > 1$ wähle $w_i \in \operatorname{argmin}_{w \in \mathcal{H}} \sum_{j=1}^{i-1} \ell_j(w)$.
     (bei Gleichstand wähle eine beliebige der am Gleichstand beteiligten Hypothesen.)

   Zunächst betrachten wir ein Hilfslemma zur Abschätzung des Regrets von FTL. Im Anschluss werden wir uns FTL für die Spezialfälle quadratischer und linearer Loss-Funktionen anschauen.

Wenn sich die Hypothesen w_i und w_{i+1} nur wenig unterscheiden, muss der Schüler in Schritt i nur wenig Regret „bezahlen“. FTL funktioniert also gut, wenn die Hypothesen w_i „stabil“ sind.

Lemma 1. *Seien $w_1, \dots, w_T, w_{T+1}$ die von FTL gewählten Hypothesen*[1]*. Dann gilt für jede Hypothese $w^* \in \mathcal{H}$ folgende Ungleichung:*

$$\text{Regret}_T(w^*) = \sum_{i=1}^{T} (\ell_i(w_i) - \ell_i(w^*)) \leq \sum_{i=1}^{T} (\ell_i(w_i) - \ell_i(w_{i+1}))$$

i) Beweisen Sie Lemma 1 durch vollständige Induktion über T.

ii) Wir betrachten nun den Spezialfall quadratischer Loss-Funktionen: Für alle $i = 1, \dots, T$ gelte $\ell_i(w) = \frac{1}{2}(w - z_i)^2$ für (vom Lehrer gewählte) Zahlen $z_i \in \mathbb{R}$. Zeigen Sie, dass der Regret von FTL begrenzt ist durch

$$\text{Regret} = \mathcal{O}(Z^2 \log(T)),$$

wobei $Z := \max_i |z_i|$. Für beschränktes Z ist der Regret also logarithmisch in T.
Hinweis: Zeigen Sie zunächst, dass hier die FTL-Hypothese w_{i+1} durch $w_{i+1} = \frac{1}{i}\sum_{j=1}^{i} z_j$ gegeben ist. Formen Sie dann um, sodass Sie eine Gleichung für $w_{i+1} - z_i = (\dots) \cdot (w_i - z_i)$ erhalten. Schätzen Sie anschließend $\ell_i(w_i) - \ell_i(w_{i+1})$ nach oben ab und wenden Sie Lemma 1 an.

iii) Jetzt betrachten wir den Spezialfall linearer Loss-Funktionen: Für alle $i = 1, \dots, T$ gelte $\ell_i(w) = w \cdot z_i$ für (vom Lehrer gewählte) Zahlen $z_i \in \mathbb{R}$.
Bestimmen Sie den Regret von FTL für die Hypothesenklasse $\mathcal{H} = [-1, 1]$ und die Folge

$$z_i = \begin{cases} -1/2 & \text{falls } i = 1, \\ 1 & \text{falls } i \text{ gerade}, \\ -1 & \text{sonst}. \end{cases}$$

b) Wir betrachten nun einen zweiten Algorithmus: **Gradientenabstieg**. Hierbei nehmen wir an, dass alle Loss-Funktionen ℓ_i differenzierbar sind, d. h. einen Gradienten $\nabla \ell_i$ besitzen.

- Eine „Lernrate“ $\eta > 0$ sei vorgegeben.
- Initialisiere $w_1 = 0$.
- Für alle $i > 1$ setze $w_{i+1} = w_i - \eta \cdot \nabla \ell(w_i)$.

Gradientenabstieg funktioniert gut, wenn die Hypothesenklasse sowie alle Gradienten beschränkt sind.

Satz 1. *Angenommen, $\mathcal{H}$ ist beschränkt, d. h. $|w - w'| \leq H$ gilt für alle $w, w' \in \mathcal{H}$ und eine Konstante H; und alle Gradienten sind beschränkt, d. h. $|\nabla \ell_i(w_i)| \leq G$ für alle $i = 1, \dots, T$ und eine Konstante G. Dann erreicht Gradientenabstieg* $\text{Regret}_T \leq \frac{1}{2\eta} H^2 + \frac{\eta}{2} G^2 T$.
Insbesondere gilt für die Wahl $\eta = \frac{H}{G\sqrt{T}}$: $\text{Regret}_T \leq HG\sqrt{T}$.

i) Beweisen Sie Satz 1.
Hinweis: Zeigen Sie zunächst für beliebiges $w^* \in \mathcal{H}$, dass $\sum_{i=1}^{T} \nabla \ell_i(w_i) \cdot (w_i - w^*) \leq \frac{1}{2\eta} H^2 + \frac{\eta}{2} G^2 T$ gilt; formen Sie dazu $(w_i - w^*)^2 - (w_{i+1} - w^*)^2$ um und bilden Sie dann eine Teleskopsumme. Nutzen Sie dann aus, dass für konvexe Funktionen ℓ stets $\ell(x) - \ell(x') \leq \nabla \ell(x) \cdot (x - x')$ gilt.

ii) Vergleichen Sie Gradientenabstieg mit FTL. Was passiert bei quadratischen Loss-Funktionen $\ell_i(w) = (w - z_i)^2$? Was passiert bei linearen Loss-Funktionen $\ell_i(w) = w \cdot z_i$?
Hinweis: Wie können Sie jeweils H nach oben abschätzen?

Bitte wenden!

[1] mit $w_{T+1} := \text{argmin}_{w \in \mathcal{H}} \sum_{j=1}^{T} \ell_j(w)$

Bonusaufgabe 8.2. *Perzeptron mit zufälligen Beispielen* (16 Extrapunkte)

Implementieren Sie die Perzeptron-Lernregel in Python. Trainieren Sie dann Ihren Perzeptron für die folgenden Funktionen mit *zufälligen* **Gegen**beispielen aus $\{0,1\}^n$ bzw. $\{0,1\}^n \times \{1\}$. (Würfeln Sie dazu zufällige *Beispiele* (gemäß der Gleichverteilung auf $\{0,1\}^n$ bzw. $\{0,1\}^n \times \{1\}$) aus und ignorieren Sie davon alle von Ihrem Perzeptron *korrekt* klassifizierten.)

a) die Funktion $f(x_1, \ldots, x_n) = x_1 \vee ... \vee x_k$ für $k = 4$ (vgl. Blatt 6 und 7),

b) die Funktion f_k mit $f_k(x_1, \ldots, x_n) = 1 \iff \sum_i x_i \geq k$ für $k = n/2$ (vgl. Blatt 6),

c) die Funktion COMP_n (vgl. Blatt 6 und 7),

d) die Funktion $f(x_1, \ldots, x_n) = \mathrm{sign}(\langle w, x \rangle)$ mit zufälligen Gewichten $w_i \in \{-10, \ldots, 10\}$.

Vergleichen Sie die beobachteten Gegenbeispielzahlen mit denen der theoretischen Schranken aus der Vorlesung bzw. früheren Übungsaufgaben. Beschreiben Sie auch sonstige Beobachtungen und Auffälligkeiten und interpretieren Sie Ihre Ergebnisse.

Hinweis: Die Funktionen und die Lernregel lassen sich mit elementaren Operationen in wenigen Zeilen implementieren. Stellen Sie dazu Gewichtsvektoren w und Beispielvektoren x als Listen dar. Zufällige Beispiele können Sie mit dem Modul `random` erzeugen. Für hinreichend kleine n können Sie den Gewichtsvektor w nach jedem Schritt ausgeben lassen und so beobachten, wie Perzeptron konvergiert. :-)

Schicken Sie (zusätzlich zur üblichen Abgabe) Ihren Quelltext an `seiwert@thi.cs.uni-frankfurt.de`.