Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2008

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
`http://ml.cs.tu-berlin.de`

Blatt 4

Abgabe: 13. Mai 2008, bis 12.15 h bei Mikio Braun (`mikio@cs.tu-berlin.de`)

Für praktische Aufgaben bitte ebenfalls den Code abgeben. Verwende Matlab oder Octave, welches unter `www.octave.org` frei verfügbar ist. Es müssen die Datei `oneclass.m` und eine Datei mit den Positionen der Hacker-Angriffe abgegeben werden.

Aufgaben

1. **Duale Darstellung der One-Class-SVM (5 Punkte)** Zeige, dass die duale Darstellung der One-Class-SVM mit Soft Margin (vgl. Seite 26 der Folien) tatsächlich die Gestalt hat

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i k(x_i, x_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$
$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i = 1 \quad \text{and} \quad 0 \le \alpha_i \le C \quad \text{for } i = 1, \ldots, n,$$

   wobei $\{x_1, \ldots, x_n\}$ Trainingsdaten sind und k einer beliebigen Kernfunktion entspricht.

2. **Darstellung (3 Punkte)** Gib die allgemeine duale quadratische Form der One-Class SVM in Matrix-/Vektornotation an.

Wie in der Vorlesung besprochen, wird die Abweichung eines Punktes z vom erlernten Model α ermittelt durch

$$a(z) = k(z, z) - 2 \sum_{i=1}^{n} \alpha_i k(x_i, z) + \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

3. **(10 Punkte)** Schreibe eine Funktion

   ```
   alpha=oneclass(K,C)
   ```

   Dabei ist `K` $\in \mathbb{R}^{n \times n}$ die Kernmatrix, und `C` ≥ 0 ist die Regularisierungskonstante. Ausgegeben wird der Vektor `alpha` $\in \mathbb{R}^n$ der Kernkoeffizienten. Verwende zur Optimierung die Funktion beispielsweise `quadprog`.

4. **(12 Punkte)**. Finde mit der One-Class SVM Hacker-Angriffe in einem vorverarbeiteten Datensatz von Netzwerkverkehr. Du findest den Datensatz auf der Veranstaltungsseite. Der Datensatz ist in eine Trainings- und Testpartition unterteilt. Wie in einem realistischen Einsatz enthalten beide Partitionen Angriffe.
   Verwende einen linearen Kern. Die Regularisierungskonstante C kann durch Raten, Probieren oder durch Kreuzvalidierung ermittelt werden. Gib das von Dir gewählte C explizit an, und dokumentiere, wie Du C gewählt hast. Übergib eine Datei mit den Positionen der Angriffe des Testdatensatzes.
   Hinweis : Der Datensatz wurde aus echten Verbindungen und Angriffen des HTTP-Protokolls generiert. Jede Verbindung x wurde über 3-Gramme in einen Vektor $\phi(x)$ abgebildet. Die Daten sind normalisiert, d.h. $||\phi(x)||_1 = 1$.

5. **(10 Punkte)**. Zeige, dass für $C = \frac{1}{n}$ die One-Class SVM den Schwerpunkt $\bar{\mu}$ der Daten bestimmt und äquivalent zur "Center-of-Mass"-Methode aus der Vorlesung ist.

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$