Maschinelles Lernen

10. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Christoph Sawade
Jules Rasetaharison

WS10/11

Ausgabe am: 17.01.11
Besprechung am: 24.01.11

Aufgabe 1 (1/4 Punkt):
Die *Logistische Regression* – anders als der Name vermuten lässt ein Klassifikationsverfahren – basiert auf der Idee, die Klassenwahrscheinlichkeiten

$$p(y_i = +1|\mathbf{x}_i) = 1 - p(y_i = -1|\mathbf{x}_i)$$

mit Hilfe einer Sigmoidfunktion $\sigma(f(\mathbf{x}_i))$ zu modellieren. Oft wird dabei die *logistische Funktion* $\sigma_l(a) = \frac{1}{1+e^{-a}}$ verwendet.

(a) Angenommen wir modellieren die Wahrscheinlichkeit $p(y_i = +1|\mathbf{x}_i; \mathbf{w}) = \sigma_l(\mathbf{w}^\mathsf{T}\mathbf{x}_i)$. Zeigen sie, dass für die Gegenwahrscheinlichkeit $p(y_i = -1|\mathbf{x}_i; \mathbf{w})$ gilt:

$$p(y_i = -1|\mathbf{x}_i, \mathbf{w}) = \sigma_l(-\mathbf{w}^\mathsf{T}\mathbf{x}_i).$$

(b) Aus Aufgabe (a) wissen wir, dass wir die Klassenwahrscheinlichkeiten mit nur einer Funktion modellieren können: $p(y_i|\mathbf{x}_i; \mathbf{w}) = \sigma_l(y_i\mathbf{w}^\mathsf{T}\mathbf{x}_i)$. Leiten sie einen ML-Schätzer für $\mathbf{w}$ her (nur das Optimierungskriterium). Nehmen sie dabei an, dass die Datenpunkte unabhängig voneinander gezogen wurden, d.h.

$$p(y_1, \ldots, y_n|\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \mathbf{w}).$$

(c) Wir nehmen an, dass der Prior des Gewichtsvektors $p(\mathbf{w})$ normalverteilt ist mit $p(\mathbf{w}) = N(\mathbf{w}^*, \Sigma^{-1}\mathbf{I})$, wobei $\mathbf{w}^*$ z.B. ein altes Modell ist. Leiten sie den MAP-Schätzer (nur das Optimierungskriterium) der Logistischen Regression her und geben sie die Verlustfunktion und den Regularisierer an.

Aufgabe 2 (1/4 Punkt):
Bisher haben wir meist angenommen, dass die Trainingsdaten linear separierbar sind. In der Praxis ist dies jedoch oft nicht der Fall. Ein vorheriges Mappen der Daten in einen anderen Raum bietet hierfür einen Ausweg. Geben sie eine Abbildungsvorschrift $\varphi : \mathbb{R}^2 \to \mathbb{R}^4$ an, welche die Daten (siehe Abbildung 1) so mappt, dass sie im neuen Raum linear trennbar sind.

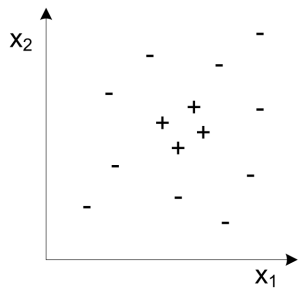
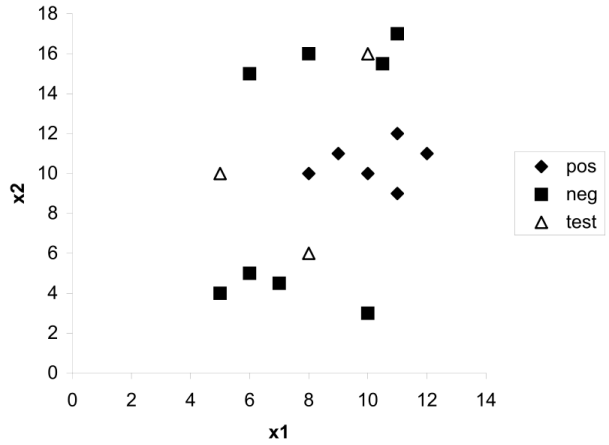


Abbildung 1: Beispieldaten

Aufgabe 3 (1/4 Punkte):
Die folgende Tabelle enthält Trainingsdaten (1-14) und Testdaten (15-17) für einen Klassifikator. In dem unteren Diagramm ist die geometrische Lage der Beispiele sichtbar.



(a) Benutzen sie eine Programmiersprache ihrer Wahl und implementieren Sie einen dualen Perzeptron-Algorithmus mit einem RBF-Kern (Radiale Basisfunktion). Setzten sie den Brennweiten-Parameter γ-Parameter des RBF-Kerns auf 1. Geben sie die Berechnungsspur an, die sich aus den Trainingsdaten ergibt.

(b) Wie kann man die resultierenden α-Werte interpretieren?

(c) Ergibt sich das gleiche Resultat, wenn man die Reihenfolge der Trainingsbeispiele umkehrt? Warum/warum nicht?

(d) Klassifizieren Sie die drei Testpunkte aus der Tabelle, welchen Klassen werden diese zugeordnet?

ID	x1	x2	Klasse
1	10	10	1
2	11	12	1
3	9	11	1
4	8	10	1
5	11	9	1
6	12	11	1
7	5	4	-1
8	6	5	-1
9	7	4.5	-1
10	10	3	-1
11	6	15	-1
12	8	16	-1
13	10.5	15.5	-1
14	11	17	-1
15	8	6	?
16	5	10	?
17	10	16	?

Aufgabe 4 (1/4 Punkt):
Neben expliziten Abbildungen $\varphi : \mathbf{x} \to \mathbf{x}'$, welche die ursprünglichen Daten in einen neuen Raum mappen, existieren auch Kernel-Funktionen.

(a) Was ist der Unterschied zw. einer Basis-Funktion $\varphi(\mathbf{x})$ und einer Kernel-Funktion $k(\mathbf{x}, \mathbf{x}')$? Welche Vor-/Nachteile hat eine Kernel-Funktion gegenüber der Verwendung eines Mappings φ?

(b) Wie könnte ein Graph-Kernel und wie ein Text-Kernel aussehen?

(c) Was ist eine Kernel-Matrix und welche Eigenschaften hat sie?

(d) Was besagt das Representer-Theorem?