Maschinelles Lernen

7. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Christoph Sawade
Paul Prasse

WS09/10

Ausgabe am: 10.12.09
Besprechung am: 05.01.10

Aufgabe 1 (1/4 Punkt):
Wir möchten einen Spam-Filter lernen, der eingehende Emails über ihre Betreff-Zeilen klassifiziert. Wir haben vier Trainingsbeispiele; die Betreff-Zeilen sind unten aufgelistet. Beispiele 1 und 2 haben wir als Spam identifiziert und Beispiele 3 und 4 betreffen die Organisation der nächsten Grillparty (nicht-Spam).

1. Abnehmen, Pillen ohne Rezept
2. Günstig Pillen
3. Einladung zum Grillen
4. Günstig Würstchen

Wir erhalten nun zwei neue Emails mit folgenden Betreff-Zeilen, die wir als Spam oder nicht-Spam klassifizieren möchten:

1. Günstig Pillen zum Abnehmen
2. Abnehmen ohne Würstchen

Modellieren sie dieses Problem mit Naive-Bayes. Die Klassenvariable ist $y \in \{Spam, nicht-Spam\}$ und als Attribute verwenden wir die "Bag of Words"-Repräsentation. Dies bedeutet, dass wir lediglich prüfen ob ein Wort in einer Betreff-Zeile vorkommt (die Position im Text und die Häufigkeit des Vorkommens werden vernachlässigt). Eine Betreff-Zeile wird also durch einen 9-dimensionalen, binären Vektor $\mathbf{x}_i$ dargestellt, wobei $x_{ij} = 1$ bedeutet, dass das j-te Wort des Lexikons [Abnehmen, Pillen, ohne, Rezept, Günstig, Einladung, zum, Grillen, Würstchen] in der i-ten Email vorkommt. Analog bedeuet $x_{ij} = 0$, dass das j-te Wort nicht vorkommt.

(a) Führen sie eine MAP-Parameterschätzung durch mit Prior-Parametern $\alpha_{x_i|y_j} = 1$ und berechnen sie für beide Testbeispiele die Klassenwahrscheinlichkeit $P(y|x, \theta)$.

(b) Welches Problem würde auftreten, wenn wir anstelle der MAP- eine ML-Parameterschätzung vornehmen?

Aufgabe 2 (1/4 Punkt):
In der folgenden Tabelle sind sechs Trainingsbeispiele für einen Klassifikator abgebildet. y_i ist das Klassenlabel, x_{1i} und x_{2i} sind die Attribute.

i	1	2	3	4	5	6
y_i	+1	+1	+1	–1	–1	–1
x_{1i}	1,0	1,3	2,5	4,2	5,1	3,7
x_{2i}	3,4	4,2	3,7	1,7	2,5	3,1

Wir betrachten den linearen Klassifikator $sign(f(\mathbf{x}))$ mit der Trennebene $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}+w_0$.

(a) Wie groß ist der euklidische Abstand eines Punktes $\mathbf{z}$ von der Trennebene $f(\mathbf{x})$?

(b) Angenommen die Trennebene verläuft durch den Koordinatenursprung, welchen Einfluss hat dann die Länge des Gewichtsvektors $\mathbf{w}$ auf das vorhergesagte Klassenlabel?

(c) Angenommen wir haben n Beispiele $\mathbf{x}_i$ mit dazugehörigen Labeln $y_i \in \{-1, +1\}$ welche linear separierbar sind, d.h. es gibt (mindestens) eine Trennebene f mit $sign(f(\mathbf{x}_i)) = y_i$ für $i = 1, \ldots, n$. Als geometrischen *margin* bezeichnet man den Abstand zwischen einer solchen Trennebene f und dem Datenpunkt welcher am nächsten an der Trennebene liegt. Wie lässt sich dieser Abstand für eine gegebene Trennebene berechnen?

(d) Stellen sie die in der Tabelle gegebenen Datenpunkte in einem Diagramm dar und zeichnen sie eine beliebige Trennebene ein. Markieren sie in dem Diagramm den geometrischen margin und zeichnen sie $\mathbf{w}$ und w_0 mit $\|\mathbf{w}\| = 1$ ein.

Aufgabe 3 (1/4 Punkt):
Wir betrachten erneut die Trainingsbeispiele aus Aufgabe 2.

(a) Simulieren sie das Training eines Rocchio-Klassifikator von Hand.

(b) Berechnen sie den Gewichtsvektor der Fisher-Diskriminate unter Verwendung der Formeln auf Seite 11 der Folien zu den Linearen Klassifikatoren. Bestimmen sie dazu zunächst die 2×2-Matrix $\mathbf{S}_W$ und deren Inverse. Tipp:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

(c) Klassifizieren sie das Testbeispiel $(3,5; 3,5)$ unter Verwendung der Klassifikationsfunktion $sign(f(\mathbf{x}))$ und den Gewichtsvektoren aus Aufgabe 3.a und 3.b. (Wir definieren $sign(a) = +1$ für $a \geq 0$ und $sign(a) = -1$ für $a < 0$.)

(d) Zusatzaufgabe (+1 Punkt): Unter welchen Voraussetzungen liefern Rocchio und die Fisher-Diskriminate gleiche Gewichtsvektoren (bzw. wann haben beide Gewichtsvektoren die gleiche Richtung)?

Aufgabe 4 (1/4 Punkt):
Wir betrachten erneut die Trainingsbeispiele aus Aufgabe 2.

Betrachten Sie den K-nearest neighbour (KNN) Ansatz. Welche Klassen werden vorhergesagt, wenn sie das Testbeispiel $(3,5; 3,5)$ für $K = 1, 2, \ldots, 6$ klassifizieren? Benutzen sie dazu die euklidische Distanz d(x,y):

$$d(x, y) = ||x - y||_2 = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}.$$