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 2

Abgabe: 29. April 2008, bis 12.15 h bei Nicole Krämer (nkraemer@cs.tu-berlin.de)

Aufgaben

Ein Ziel dieses Aufgabenblatts ist es, einige grundlegende Optimierungstechniken (EM-Algorithmus, Gradientenabstieg) zu erlernen.

1. **Gradientenabstieg (20 Punkte)** Wir betrachten im Folgenden eine Menge $\mathcal{F}$ von Funktionen, die durch Parameter $\theta \in \Theta$ definiert ist:

$$\mathcal{F} \quad = \quad \{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}\,.$$

Wir möchten für eine Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ das empirische Risiko minimieren, d.h. wir suchen das Minimum der Funktion

$$L(\theta) = \sum_{i=1}^{n} l\left(y_i, f_\theta\left(x_i\right)\right)\,.$$

Dabei ist $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ eine Verlustfunktion. Das Gradientenabstiegsverfahren ist eine iterative Methode. Nehmen wir an, im k-ten Schritt haben wir den Parameter θ^k berechnet. Das Verfahren basiert auf der Idee, dass der negative Gradient von L immer in Richtung des Minimums zeigt. Daher wird vom aktuellen Parameter θ^k ein Vielfaches des Gradienten subtrahiert, genauer:

- Berechne den negativen Gradienten von $L(\theta)$ an der Stelle θ^k:

$$\bigtriangledown = -\left.\frac{\partial}{\partial\theta} L(\theta)\right|_{\theta=\theta^k}\,.$$

- Aktualisiere

$$\theta^{k+1} = \theta^k + \eta \cdot \bigtriangledown\,.$$

(Die Schrittweite η wird entweder fest gewählt oder in jedem Schritt durch ein Optimierungsverfahren bestimmt.)

Im Folgenden betrachten wir die Menge

$$\mathcal{F} = \left\{f_w(x) = sign\left\{w^\top x\right\}, w \in \mathbb{R}^d\right\} \tag{1}$$

aller Hyperebenen durch den Nullpunkt.

Wir wählen zunächst den $0-1$ Verlust

$$l(y, f_w(x)) = \left\{\begin{array}{lcl} 1 & , & y \cdot x^\top w \leq 0 \\ 0 & , & y \cdot x^\top w > 0 \end{array}\right.\,.$$

(a) Skizziere die Verlustfunktion in Abhängigkeit von $-yx^\top w$.

(b) Versuche, die Funktion L mit Hilfe des Gradientenverfahrens zu minimieren. (Ignoriere dabei beim Differenzieren nichtdifferenzierbaren Stellen.) Was stellst Du fest?

Wir approximieren den 0-1-Verlust durch folgende Funktion

$$l\left(y, f_w(x)\right) = \begin{cases} -yx^\top w & , y \neq f_w\left(x\right) \\ 0 & , y = f_w\left(x\right) \end{cases} \tag{2}$$

3. Zeichne den Verlauf dieser Funktion in die Skizze des 0-1-Verlustes ein. Wie gut ist die Approximation?
4. Bestimme die Iterationsschritte für das Gradientenabstiegsverfahren bezüglich dieser Verlustfunkion.

Wir betrachten nun folgende Variante des Gradientenverfahrens: Sei ein w^k gegeben. Für $i = 1, \ldots, n$

- Berechne
$$\bigtriangledown = -\frac{\partial}{\partial w} l\left(y_i, f_w(x_i)\right)\Big|_{w=w^k} .$$
(Wir differenzieren also nur die Verlustfunktion für **einen** Punkt.)
- Berechne
$$w^{k+1} = w^k + \eta \cdot \bigtriangledown .$$

5. Bestimme für die Verlustfunktion (2) den resultierenden Algorithmus. (Kommt Dir der Algorithmus bekannt vor?)

2. **Expectation-Maximization (EM) (20 Punkte)** Wir betrachten im Folgenden eine Mischung von 2 univariaten Gaußverteilungen, d.h. eine Zufallsvariable mit Dichtefunktion

$$\begin{aligned} f(x) &= \pi g_1(x) + (1-\pi) g_2(x)\,, \\ g_k &= \mathcal{N}(\mu_k, \sigma_k^2)\,, k = 1, 2\,, \\ \pi &\in [0, 1] \end{aligned}$$

(Eine Interpretation dieser Mischung ist, dass die Daten aus $K = 2$ Clustern bestehen. Jedes Cluster k ist $\mathcal{N}(\mu_k, \sigma_k^2)$ verteilt.) Für eine Menge an Zufallsvariablen $X_1, \ldots, X_n$ sollen die Parameter $\pi, \mu_{1,2}, \sigma_{1,2}^2$ geschätzt werden. Da die direkte Minimierung der (Log-)Likelihoodfunktion numerisch komplex ist, erweitert man das Problem zunächst um Zufallsvariablen $Z_1, \ldots, Z_n$, die codieren, in welches Cluster die Zufallsvariable X_i fällt. Z_i sind Bernoulli-Variablen, d.h. sie nehmen die Werte 0 oder 1 an. Die grundlegende Idee ist nun die folgende:
(1) Angenommen, wir kennen die Parameter π, μ_k, σ_k^2. Dann kann man den Erwartungswert von $Z_i|(X_1, \ldots, X_n)$ berechnen.
(2) Angenommen, man kennt den Erwartungswert von Z_i. Dann kann man zeigen, dass die Maximum-Likelihood-Schätzer für μ_k und σ_k^2 ein gewichtetes arithmetisches Mittel und eine gewichtete Varianz aller Punkte sind.
Der EM-Algorithmus iteriert diese Schritte, in dem er in (1) die aktuellen Schätzungen der Parameter π, μ_k, σ_k^2 verwendet, und in (2) die in (1) geschätzten Erwartungswerte von Z_i zur Berechnung der übrigen Parameter $\theta = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ verwendet.

(a) **E-Schritt** Zeige, dass

$$\begin{aligned} \gamma_i &= P(Z_i = 1|\theta, X_1 = x_1, \ldots, X_n = x_n) \\ &= \frac{\pi g_1(x_i)}{\pi g_1(x_i) + (1-\pi) g_2(x_i)} \end{aligned}$$

Man bezeichnet γ_i als „Verantwortlichkeit" (responsibility) des 1. Clusters für den Punkt x_i.

(b) **M-Schritt** Die Log-Likelihood von θ - gegeben $Z_i = \gamma_i$ - lautet

$$l(\theta) \quad = \quad \sum_{i=1}^{n} \{\gamma_i \ln g_1(x_i) + (1-\gamma_i) \ln g_2(x_i)\} + \sum_{i=1}^{n} \{\gamma_i \ln \pi + (1-\gamma_i) \ln(1-\pi)\} \; .$$

Folgere daraus, dass der Maximum-Likelihood-Schätzer für θ - gegeben $Z_i = \gamma_i$ - die folgende Gestalt hat:

$$\widehat{\mu}_1 = \frac{1}{\sum_{i=1}^{n} \gamma_i} \sum_{i=1}^{n} \gamma_i x_i \quad , \quad \widehat{\mu}_2 = \frac{1}{\sum_{i=1}^{n} (1-\gamma_i)} \sum_{i=1}^{n} (1-\gamma_i) x_i$$

$$\widehat{\sigma}_1^2 = \frac{1}{\sum_{i=1}^{n} \gamma_i} \sum_{i=1}^{n} \gamma_i \left((x_i - \widehat{\mu}_1)^2 \right. \quad , \quad \widehat{\sigma}_2^2 = \frac{1}{\sum_{i=1}^{n} (1-\gamma_i)} \sum_{i=1}^{n} (1-\gamma_i) \left((x_i - \widehat{\mu}_2)^2 \right.$$

$$\widehat{\pi} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \gamma_i$$

Bemerkung: In Aufgabenteil 1 werden die Hyperebenen (1) ohne Achsenabschnitt dargestellt. Um auch Ebenen mit Achsenabschnitt zu modellieren, kann man beispielsweise den Vektor $x \in \mathbb{R}^d$ zu einem Vektor $(1, x) \in \mathbb{R}^{d+1}$ erweitern.