Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Volker Tresp
Franz Graf, Marisa Thoma

14.04.2008

Maschinelles Lernen und Data Mining

Sommersemester 2008

Übungsblatt 3

Besprechung des Übungsblattes am 02.06.2008

Aufgabe 3-1 Lineare Regression mit Gauss'schem Rauschen

Gegeben sei ein Datensatz D mit $d_i = (x_{i,1}, \ldots, x_{i,M}, y_i)^T$ auf N Datenpunkten mit M Variablen, dessen Zielgröße $\mathbf{y}$ linear von $\mathbf{X}$ abhängt. Aufgrund von technischen Ungenauigkeiten wurden die Eingangsvariablen von $\mathbf{X}$ jedoch nur verrauscht aufgenommen, d.h.:

$$y_i = x_i^T \mathbf{w} + \epsilon_i \,,$$

wobei ϵ_i den Rauschfehler von Datenpunkt i darstellt. Nehmen wir weiter an, dass ϵ gaussverteilt ist:

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\epsilon_i^2} \,.$$

Damit können wir die Verteilung von $\mathbf{y}$ in Abhängigkeit der Variablen $\mathbf{X}$ und des Modellparameters $\mathbf{w}$ darstellen als

$$P(y_i|x_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\mathbf{w})^2} \,.$$

a) *kleiner Exkurs in die Statistik*

Bestimmen Sie den Parameter $\hat{\mathbf{w}}$ der die Wahrscheinlichkeiten der Trainings-Daten $P(D|\mathbf{w})$ maximiert. Hierbei erweist sich der *Maximum-Likelihood Schätzer* als hilfreich. Dieser schätzt den Parameter $\hat{\theta}^{\text{ML}}$ für den die Likelihood der beobachteten Daten maximal ist. (Zur Erinnerung: die Likelihood einer Wahrscheinlichkeitsfunktion $f(x) = g(x,\theta)$ mit Modellparameter θ ist genau diese Wahrscheinlichkeitsfunktion parametrisiert auf θ: $L(\theta) = g(x,\theta)$.) Wie suchen also

$$\hat{\mathbf{w}}^{\text{ML}} = \arg\max_{\mathbf{w}} L(\mathbf{w})$$

Die Maximierung der Likelihood erreichen wir wie üblich durch die Faustformel "ableiten und auf 0 setzen". Nützliche Ableitungsregeln auf Vektoren wurden in der Vorlesung vorgestellt.

Bei der Bestimmung der Likelihoodfunktion können Sie davon ausgehen, dass $\mathbf{w}$ unabhängig von den Eingangsdaten $\mathbf{X}$ verteilt ist.

b)

In einem Bayes'schen Ansatz sind die Parameter Zufallsvariablen. Eine beliebte a priori Verteilungsannahme ist

$$P(\mathbf{w}) = \frac{1}{(2\pi\alpha^2)^{\frac{M}{2}}} e^{\left(-\frac{1}{2\alpha^2}\sum_{j=0}^{M-1} w_j^2\right)}$$

Berechnen Sie den Paramter $\hat{\mathbf{w}}$, der den Ausdruck $P(\mathbf{w})P(D|\mathbf{w})$ maximiert. Ergibt sich dadurch eine neue Interpretation des λ-Termes aus der regularisierten Kostenfunktion (*penalized least squares* (PLS))?

Aufgabe 3-2 Beispielanwendung für CF+
schriftlich bearbeiten

Gegeben sei ein Datensatz D mit $d_i = (x_{i,1}, \ldots, x_{i,M})^T$ für $N = 5$ Benutzer einer Filmdatenbank mit je $M = 6$ auf einer Skala zwischen 1 und 5 beurteilten Filmen:

user	300	Bobby	Jumper	Juno	Next	X-Men
1	5	1	5	1	4	5
2	5	2	5	1	5	4
3	3	3	2	3	2	2
4	2	5	1	4	2	1
5	1	3	1	2	1	1

D kann umformuliert werden zu einem Problem des Collaborativen Filterns (CF): Wir deklarieren ein $j \in \{1,M\}$ als den vorherzusagenden Film für den Anfragebenutzer $z^T = d_i$ für $i \in \{1,N\}$. Somit entspricht Spalte j von D dem Zielvektor y.

a) Bestimmen Sie die vorhergesagten Scores zum Film *Next* für alle User. Warum wird hier die Skala von 1 bis 5 nicht eingehalten?

b) Wie würden Sie den Fehler für diese Vorhersagen bewerten? Welcher Film kann Ihrem Vorhersagemodell nach am besten vorhergesagt werden? Welcher Benutzer votiert am "berechenbarsten"?

c) Was für einen Score erwarten Sie für *Next* bei zufälliger Mustereingabe? Simulieren Sie 10000 zufällige Muster für die verbleibenden Filme und bestimmen Sie den Mittelwert über die zugewiesenen Scores.

d) Was ist der mittlere, auf zufälligen Benutzern vorhergesagte Score über alle Filme? Warum?