Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Prof. Dr. Claus Weihs
Dr. Thomas Liebig
Sebastian Buschjäger
Malte Jastrow
Jennifer Neuhaus-Stern

Übung zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2018

Übungsblatt Nr. 6

Der Abgabetermin ist Dienstag der 12.06.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Overfitting) (5 Punkte)

In der Templatedatei `wid_SoSe18_tm06.Rmd` finden Sie einen Beispieldatensatz mit 5000 Beobachtungen zu je 5 Featuren. Der Datensatz wird in einen Trainingdatensatz mit 4000 Beobachtungen und einen Testdatensatz mit 1000 Beobachtungen aufgeteilt. Es handelt sich um eine (fehlerbehaftete) Regressionsaufgabe.

a) (1 Punkt) Fitten Sie folgendes Polynom mit Parametern $c_0, \dots, c_5$ an den Datensatz an und berechnen Sie den Mean-Squared-Error für den Trainings- und Testdatensatz:

$$f(x) = c_0 + \sum_{i=1}^{5} c_i \cdot x_i$$

Hinweis: Verwenden Sie die `lm` Funktion in Kombination mit `poly`. Beachten Sie inbesondere die Angabe `raw=TRUE` bei der `poly` Funktion.

b) (2 Punkt) Erweitern Sie ihre Implementierung, sodass Polynome vom Grad 2 bis Grad 8 ebenfalls gefittet werden und evaluieren Sie deren Performanz auf den Trainings- und Testdaten. Ein Polynom vom Grad 2 enthält neben der linearen Kombination aller Feature, auch die quadratische Einträge:

$$f_2(x) = c_0 + \sum_{i=1}^{5} c_i \cdot x_i + \sum_{i=1}^{5} \sum_{i=j}^{5} c_{i,j} \cdot x_i \cdot x_j$$

Ein Polynom vom Grad 3 hat nun zusätzlich alle kubischen Einträge, d.h.

$$f_3(x) = c_0 + \sum_{i=1}^{5} c_i \cdot x_i + \sum_{i=1}^{5} \sum_{i=1}^{5} c_{i,j} \cdot x_i \cdot x_j + \sum_{i=1}^{5} \sum_{i=1}^{5} \sum_{k=1}^{5} c_{i,j,k} \cdot x_i \cdot x_j \cdot x_k$$

Alle weiteren Polynome ergeben sich kanonisch nach diesem Muster.
Hinweis: Betrachten Sie die `degree` Option von `poly`.

c) (2 Punkt) Plotten Sie den Trainings- und Testfehler aller Polynome in einen gemeinsamen Plot und vergleichen Sie beide Kurven. Was fällt Ihnen auf? Erklären Sie den Unterschied!

Aufgabe 2 (Kernfunktionen) (5 Punkte)

Sie haben in der Vorlesung bereits die Support-Vektor-Machine (SVM) kennengelernt. Teil der Support-Vektor Methodik ist die Berechnung impliziter Feature-Maps durch den Kernel-Trick $K(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y})$, wobei $K(\cdot, \cdot)$ eine Kernfunktion und $\phi(\cdot)$ die dadurch implizierte Feature-Map bezeichnet.
Kernfunktionen sind ein wesentlicher Bestandteil vieler maschineller Lernverfahren und bieten eine umfassende Theorie. In dieser Aufgabe sollen Sie ihr Wissen um Kernfunktionen vertiefen.

a) (1 Punkt) Wie lautet eine Feature-Map ϕ, die der polynomielle Kernel $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T \boldsymbol{y})^d$ für zwei-dimensionale Beispiele $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ mit $d = 2$ erzeugt?

b) (1 Punkte) Eine sehr beliebte Kernfunktion ist der RBF-Kernel. Dieser modelliert den Abstand zweier Beispiele $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ mit Hilfe der Exponentialfunktion:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||^2\right)$$

Es lässt sich zeigen, dass die Feature-Map dieses Kernels durch folgende Funktion gegeben ist:

$$\phi(\boldsymbol{x})^T \phi(\boldsymbol{y}) = \sum_{j=0}^{\infty} \frac{(\boldsymbol{x}^T y)^j}{j!} \exp\left(-\frac{1}{2}||\boldsymbol{x}||\right) \exp\left(-\frac{1}{2}||\boldsymbol{y}||\right)$$

Evaluieren Sie diese Kernfunktion: Welche Dimensionalität hat die Feature-Map des RBF-Kernels? Was bedeutet dies im Hinblick auf den Trainingsdatensatz?

c) (1.5 Punkte) Auf dem vergangenen Arbeitsblatt haben Sie bereits eine lineare SVM mit Hilfe des `e1071` Pakets trainiert. Trainieren Sie nun eine SVM auf dem Datensatz aus Aufgabe 1 mit folgenden Kernfunktionen: RBF-Kernel, linearer Kernel und polynomieller Kernel. Wählen Sie $C = 1$. Berechnen Sie wieder den Test und Trainingsfehler und interpretieren Sie das Ergebnis im Hinblick auf die Dimensionalität der Feature-Maps der jeweiligen Kernel.
Hinweis: Da es sich hierbei um eine Regressionsaufgabe handelt müssen Sie den Parameter `type="eps-regression"` angeben.

d) (1.5 Punkte) Wiederholen Sie das Experiment aus Teilaufgabe c) mit $C \in \{10, 20, 50, 100, 200\}$. Was fällt Ihnen auf? Interpretieren Sie das Ergebnis.
Hinweis: Zur Wahl von C können Sie den Parameter `cost=X` benutzen.

Aufgabe 3 (Ziegenproblem mit graphischen Modellen) **(10 Punkte)**

In dieser Aufgabe wollen wir das Ziegenproblem (manchmal auch 3 Türen Problem oder Monty-Hall Dilemma) mit Hilfe von graphischen Modellen lösen. Betrachten Sie folgende Situation: Sie nehmen an einer Spielshow teil in welcher Sie eine von drei Türen zur Auswahl haben. Hinter zwei der drei Türen befindet sich jeweils eine Ziege, wohingegen hinter der Dritten Tür ein Sportwagen auf Sie wartet. Die Türen sind natürlich verschlossen und Sie wissen nicht hinter welcher Tür sich der Sportwagen befindet. Der Moderator bittet Sie eine Tür zu wählen. Danach öffnet der Moderator eine der anderen Türen und zeigt ihnen eine Ziege. Nun fragt der Moderator Sie, ob Sie bei ihrer ursprünglichen Wahl bleiben oder ob Sie die andere Tür wählen möchte. Was tun Sie und wieso?

a) (2 Punkt) Wir definieren 4 Variablen um den Ablauf zu modellieren. Bezeichne D die Zufallsvariable hinter welcher Tür das Auto liegt. Ferner seien F und H die Türen die Sie zuerst wählen (F) bzw. vom Moderator geöffnet wurden (H). Zuletzt modelliere I die Zufallsvariable ob $F = D$ ist. Zeichnen Sie den Graphen für die vier Variablen und markieren Sie die beobachteten Variablen. Welchen Wertebereich können die einzelnen Variablen annehmen?

b) (1 Punkt) Geben Sie die Wahrscheinlichkeitstabellen für $p(D)$ und $p(F)$ für alle Zustände von D und F an.

c) (2 Punkt) Wie lauten die Wahrscheinlichkeitstabellen $p(I|D, F)$ und $p(H|D, F)$?

d) (2 Punkt) Wie lautet die Wahrscheinlichkeitstabelle $p(I|F = 1)$ und $p(D|F = 1)$?

e) (2 Punkt) Wie lautet die Wahrscheinlichkeitstabelle $p(I|F = 1, H = 2)$ und $p(D|F = 1, H = 2)$?

f) (1 Punkt) Was tun Sie und wieso?