Vorlesung
Adaptive Modellierung und Simulation

GOETHE UNIVERSITÄT FRANKFURT AM MAIN
Prof. Dr. R. Brause
R.Brause@cs.uni-frankfurt.de

Übungsblatt 3 Ausgabe: 12.5. Abgabe: 19.5.

In diesem Aufgabenblatt wollen wir uns ansehen, wie sich Neuronale Netze dazu einsetzen lassen, ein uns unbekanntes System zu modellieren.

Aufgabe 3.1 Modellierung mit RBF-Netzen (10 Punkte)

Die RBF-Netze sind eine wichtige Alternative zu den MLP-Backpropagation-Netzen. In dieser Aufgabe wollen wir uns einen Überblick über die Eigenschaften verschaffen. Dazu lösen wir zuerst wieder das Problem, eine unbekannte Funktion zu approximieren.

a) Die Datei `aprx_noised.dt` enthält pro Zeile einen x-Wert und den dazu verrauschten Funktionswert einer unbekannten Funktion. Trainieren Sie nun ein RBF-Netz darauf, zu generalisieren und diese Funktion zu lernen, und versuchen Sie herauszufinden,

   1. welchen Einfluss die Anzahl der RBF-Neuronen
   2. und welchen Einfluss die Breite σ der RBF-Neuronen

   auf den Fehler des RBF-Netzes (Differenz der Ausgabe zur Lehrervorgabe) hat. (6 Pkte)

Ein weiteres, wichtiges Gebiet ist die Klassifizierung unbekannter Daten.

b) Die Dateien `TrainingSet.dt`, `ValidationSet.dt` und `TestSet.dt` enthalten jeweils eine Trainings-, Validierungs- und Testmenge bestehend aus 2-dimensionalen Datenpunkten, die in 16 Klassen (1-aus-16 Kodierung) eingeteilt sind. Jede Zeile in diesen Dateien entspricht einem Muster, wobei die ersten 2 Werte einer Zeile die Eingaben und die letzten 16 Komponenten einer Zeile die Klassenzugehörigkeit zu den Eingaben angeben.

   Trainieren Sie nun ein RBF-Netz darauf, eine Klassifikation der in den Dateien `TrainingSet.dt`, `ValidationSet.dt` und `TestSet.dt` vorhandenen Muster zu lernen. Die Dateien sollen dabei entsprechend Ihrer Bezeichnung als Trainings-, Validierungs- und Testmenge während des Trainings verwendet werden.

   Versuchen Sie dann herauszufinden,

   1. welchen Einfluss die Anzahl der RBF-Neuronen,
   2. welchen Einfluss die Breite σ der RBF-Neuronen
   3. und welchen Einfluss die in den Folien erwähnte Normierung der Ausgaben der RBF-Schicht auf den Fehler des RBF-Netzes (Differenz der Ausgabe zur Lehrervorgabe) und die per Winner-Take-All (größte Ausgabe gewinnt, s. Folien) durchgeführte Klassifizierung hat.

   Erstellen Sie zur Beantwortung dieser Fragen eine repräsentative Auswahl an Plots der Entscheidungsgebiete, auf dem die durch das Netz vorgenommene Klassenaufteilung des Eingaberaums und die Datenpunkte jeder Klasse zu erkennen sind. (4 Pkte)

Aufgabe 3.2 Optimale Basisfunktionen (6 Punkte)

Beweisen Sie:

a) Angenommen, wir können nur den reellen Mittelwert c und die Streuung σ eines Datensatzes im ein-dimensionalen Fall beobachten. Dann ist die Basisfunktion p(x), die die meiste Information H(p) über die Muster repräsentiert, die Gauß'sche Glockenfunktion

$$p(x) = A \exp(-x^2/2\sigma^2)$$

Dabei ist die Information H(p) durch den Ausdruck

$$H(p) = -\int p(x) \ln p(x)\, dx$$

gegeben und x eine reelle Zahl. Lösen Sie dieses Problem mit Hilfe des Lagrange-Ansatzes. (4 Pkte)

Hinweis:
Wir suchen diejenige Funktion p(x), die die Information H(h(x)) über alle möglichen Funktionen h(x) maximiert

$$p = \arg\max_h H(h(x))$$

wobei bei zentrierter Eingabe $\langle x \rangle = c = 0$ die beiden Nebenbedingungen gelten müssen

$$\sigma^2 = \langle x^2 \rangle = \int p(x)\, x^2\, dx \quad \text{oder} \quad g_1(x) := \int p(x)\, x^2 dx - \sigma^2 = 0$$

und

$$\int p(x)\, dx = 1 \quad \text{oder} \quad g_2(x) := \int p(x)\, dx - 1 = 0$$

b) Ändern Sie nun die Voraussetzungen: Sei x aus dem Intervall [0,1] und die Varianz ist nicht gegeben. Was folgt daraus für die optimale Basisfunktion; welche Form hat sie nun? (2 Pkte)

Aufgabe 3.3 Modellierung mit RBF-Netzen: Kreditkartenanalyse (6 Bonus-Punkte)

Die missbräuchliche Benutzung von Kreditkarten stellt für viele Unternehmen ein Problem dar. Insbesondere ist es wichtig, bereits vor einer Transaktion beurteilen zu können, ob eine nicht gesperrte Kreditkarte zum Betrug eingesetzt wird oder nicht und danach die Transaktion zuzulassen oder zu verweigern. Trainieren Sie ein RBF-Netz darauf, eine Klassifikation des Card-Datensatzes (`card1.dt`) zu lernen. Beachten Sie: die letzten zwei Werte geben die Klasse des Musters an, siehe Dokumentationsdatei `CardDocu.txt`.

a) Versuchen Sie die Anzahl der Neuronen und die Breite σ der RBF-Neuronen so anzupassen, dass Sie ein möglichst gutes Klassifikationsergebnis erzielen. (2 Pkte)

b) Geben Sie anschließend den Generalisierungsfehler als Prozentsatz fehlerhaft klassifizierter Muster Ihres Netzes an. (1Pkt)

c) Weisen Sie durch Ausgabe der Trainingsstatistik (Gradientenlänge und Fehlerwerte auf der Trainings-/Validierungs-/Testmenge) während des Trainings nach, dass das Netz ausreichend (d.h. bis zum Erreichen eines lokalen Minimums) trainiert wurde und kein (oder nur ein sehr geringer) *Overfitting*-Effekt vorliegt. (2 Pkte)

d) Nehmen Sie an, Sie hätten ein System, das Ihnen als Klassifikation für ein Muster x des Card-Datensatzes immer dieselbe Klassenzugehörigkeit ausgeben würde, ungeachtet der Tatsache, wie x beschaffen ist. Welchen Prozentsatz fehlerhaft klassifizierter Muster hätte dieses System im besten Fall auf der Testmenge? Diskutieren Sie vor diesem Hintergrund die Fehlerrate Ihres eigenen Systems. (1 Pkt)