Übung 11 Neuronale Netze WS06-07

Prof. Dr. S. Posch
Dipl.BioInform. Andrè Gohr(@informatik.uni-halle.de)

Institut für Informatik
Universität Halle

Abgabe: 11./12.1. in der Übung

Transkriptions-Faktor-Bindestellen (tfbss) sind kurze Abschnitte der DNA (einsträngig), an die Transkriptions-Faktoren mit hoher Affinität binden können. Die Vorhersage von tfbss ist hilfreich unter anderem für die Entschlüsselung der Genregulation.
Auf der Website zur Übung finden Sie drei Dateien. `tfbs_train.dat` und `tfbs_test.dat` enthalten eine Menge von tfbss der Länge 12. `nontfbs.dat` enthält eine Menge von DNA-Sequenzen gleicher Länge, von denen man weiß, daß sie keine tfbss sind. Es handelt sich um ein 2-Klassenproblem. Es sind 12-dim Merkmalsvektoren der einen (tfbss (positiv)) und der anderen Klasse (non-tfbss (negativ)) gegeben.

Aufgabe 11.1 (6 Punkte)
Die 12-dim-Vektoren (jede Komponente kann 4 mögliche Werte annehmen entps. DNA-Alphabet) sollen dem backprop-MLP kodiert gegeben werden. Jede der 4 Basen soll als ein Basisvektor im R^4 dargestellt werden. Die Eingabevektoren für das MLP bestehen dann aus 48 Komponenten, wobei jede Base einfach durch den zugehörigen 4-dim Basisvektor ersetzt wird.
Das MLP besteht aus einer Eingabe- einer verdeckten - und einer Ausgabeschicht. Die Eingabeschicht bestehe aus 48 Eingabeneuronen. Die Ausgabeschicht aus 2 Ausgabeneuronen, die jeweils vollständig mit allen Neuronen in der verdeckten Schicht verbunden sind. Das heißt, dass das MLP eine 2-dim Ausgabe liefert wobei die Sollausgaben für die tfbss-Klasse $(1, 0)$ und die Sollausgabe für die non-tfbss-Klasse $(0, 1)$ ist. In der verdeckten Schicht sollen sich:

- 5
- 10
- 50

Neurone befinden, die jeweils voll mit den 48 Eingabeneuronen verbunden sind.
Setzen Sie diese drei backprop-MLPs mit Hilfe Ihres erweiterten objekt-orientierten Konzeptes für neuronale backprop-Netze um!

Aufgabe 11.2 (6 Punkte)
Die vorgeschlagenen backprop-MLPs sollen nun trainiert und zur Klassifikation der tfbss verwendet werden. Gehen Sie dabei vor folgt vor:

- splittem Sie zufällig den non-tfbs-Datensatz in 80% und 20% große Teildatensätze (der 80%-Datensatz sei der jeweilige train-Datensatz, der 20%-Datensatz sei der jeweilige test-Datensatz)
- verwenden Sie `tfbs_train.dat` als Trainingsbeispiele und `tfbs_test.dat` als Testbeispiele für die tfbs-Klasse
- alle (rechnenden) Neurone sollen $g(h) = \frac{1}{1+exp(-h)}$ als Aktivierungsfunktion verwenden
- trainieren Sie die 3 backprop-MLPs unter Verwendung der beiden train-Datensätze (100000 Epochen (eine Epoche = einmaliges Zeigen aller Elemente der Vereinigung der beiden train-Datensätze in zufälliger Reihenfolge), Lernschrittweite $\epsilon = 0.5$)
- bestimmen Sie nach jeder Epoche die Klassifikationsrate $r(t)$ $t \in [1; 100000]$ des bis dahin gerlernten MLPs, wenn dieses zum Klassifizieren der Elemente der beiden train-Datensätze verwendet wird

- bestimmen Sie nach jeder Epoche die Klassifikationsrate $s(t)$ $t \in [1; 100000]$ des bis dahin gerlernten MLPs, wenn dieses zum Klassifizieren der Elemente der beiden test-Datensätze verwendet wird
- plottem Sie $r(t)$ & $s(t)$

Beim klassifizieren soll sich das backprop-MLP für die Klasse entscheiden, deren zugehöriges Neuron die größte Antwort liefert: z.B. sei $(0.7, 0.5)$ die Anwort der Ausgabeneurone, dann entscheidet sich das MLP für Klasse 1 (tfbs). Die Klassifikationsrate ist der Prozentsatz von allen getesteten Eingaben, die richtig klassifiziert wurden.

Werten Sie Ihre Ergebnisse aus sowohl hinsichtich $r(t)$ und $s(t)$ als auch hinsichtlich der Anzahl der Neurone in der verdeckten Schicht!

Lassen Sie auch Ihre gut kommentierten Source-Codes dem Übungsleiter per email zukommen.