4. Übungsblatt

(Support Vector Maschinen)

Im Gegensatz zu den bisher verwendeten Klassifikatoren, die Klassengebiete explizit modellieren, zum Beispiel über (Mischungen von) Gaußverteilungen, sollen nun diskriminative Klassifikatoren untersucht werden. Auf diesem Übungsblatt behandelen wir Support Vector Maschinen (SVM).

Für das Training und die Auswertung von SVMs benutzen wir die Bibliothek *sklearn*. Folgen Sie den Kommentaren im Quelltext der Module *blatt4.aufg07* und *blatt4.aufg08*.

7. Aufgabe: Der Datensatz data2d enthält Daten einer tri-modalen Verteilung (Klasse 1), die mit einer Normalverteilungsdichte schlecht zu repräsentieren ist.

1. Trainieren Sie zunächst unter Verwendung der Trainingsdaten eine Support Vector Maschine, die zur Trennung von Klasse 0 und 2 verwendet werden kann (Klasse 1 wird hier nicht betrachtet). Wie hoch ist der Klassifikationsfehler im Vergleich zum Normalverteilungsklassifikator?

   Visualisieren Sie die resultierende Trennfunktion. Diskutieren Sie den Einfluss der Slack-Variablen (in Form des C-Parameters) auf die Trennfunktion.
2. Trainieren Sie nun eine SVM für die Klassifikation der Klassen 1 und 2. Evaluieren Sie verschiedene Kernelfunktionen.
3. Verallgemeinern Sie nun die SVM-basierte Klassifikation auf das tatsächliche 3-Klassenproblem. Wie lassen sich *allgemein* derartige Multiklassenprobleme mit SVMs lösen?
4. Evaluieren Sie nun die Klassifikationsleistung für verschiedene Parametrisierungen. Wir lassen sich die optimalen Parameter bestimmen.
5. Vergleichen Sie Ihre Ergebnisse mit den bisher erzielten Klassifikationsfehlerraten.

8. Aufgabe - OPTIONAL: Die Leistungsfähigkeit der Support Vector Maschinen soll nun für das realistischere Szenario der Zeichenerkennung untersucht werden.

1. Erstellen Sie ein komplettes Klassifikationssystem für den MNIST-Datensatz in Originalrepräsentation (784-D) auf der Basis einer linearen SVM zunächst mit den Standardparametern von *sklearn*.

   Vergleichen Sie das Klassifikationsergebnis mit denen der vorherigen Klassifikatoren.
2. Verwenden Sie an Stelle der 784-dimensionalen Originaldaten die Merkmalsrepräsentationen der vorhergehenden Aufgaben. Verwenden Sie zunächst eine lineare SVM mit soft-margin. Variieren Sie nun die verwendete Kernelfunktion und vergleichen Sie Ihre Ergebnisse.
3. Evaluieren Sie, wie sich Veränderungen der jeweiligen Kernelparameter auf die entsprechenden Klassifikationsergebnisse auswirken. Verwenden sie dazu Grid-Search (weitergehende Information finden Sie auch im SVM-Guide, der auf der Webseite der Übung verlinkt ist).

   (optional) Führen Sie eine Kreuzvalidierung durch.