Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert
Erich Schubert

Knowledge Discovery in Databases II

WS 2013/2014

Übungsblatt 5: Ensemble Learning und Multi-Repräsentierte Daten

Aufgabe 5-1 Bias, Variance und Noise

Die Fehlerarten Bias, Variance und Noise spielen für Ensembles eine große Rolle, tauchen aber auch bei einzelnen Klassifikatoren bereits auf. Diskutieren Sie den Unterschied anhand des Beispiels aus der Vorlesung:

(a) Mit dem Langbogen abgefeuerte Pfeile verfehlen das Ziel weiter, als die mit dem olympischen Recurve-Bogen abgeschossenen Pfeile.

(b) Ein Sportschütze stellt fest, dass er bei Wind das Ziel meist seitlich verfehlt.

(c) Beim Schießen auf die Entfernung von 30 Metern sind die Abweichungen geringer als auf 50 Meter.

Aufgabe 5-2 Bias, Variance und Noise II

Überlegen Sie sich, welche Auswirkungen Sie von folgenden Änderungen an einem Experiment auf Bias, Variance und Noise erwarten würden:

(a) Verwenden einer größeren Trainingsmenge

(b) Verwenden von zusätzlichen, hilfreichen Attributen (Features)

(c) Verwenden von weniger Attributen (Features)

(d) Verwenden von polynomiellen Kerneln und abgeleiteten Features

(e) Weniger Regularisierung (siehe Vorlesung Maschinelles Lernen, falls Sie das gehört haben)

(f) Stärkere Regularisierung (siehe Vorlesung Maschinelles Lernen, falls Sie das gehört haben)

(g) Genauere Definition und Validierung der Labels

Aufgabe 5-3 Kombination von zwei Ähnlichkeitsmaßen

Gegeben seien zwei Kernels k_1 und k_2. Wir kombinieren sie in einen gemeinsamen Kernel k_{com}

$$k_{com} = \alpha k_1 + (1 - \alpha)k_2, \tag{1}$$

wobei $\alpha \in [0; 1]$.

Wir wenden den Kernel k_{com} auf zwei Klassifikationsprobleme an, wobei wir das Experiment für verschiedene Werte von α wiederholen. In der folgenden Abbildung stellen wir die Klassifikationsgenauigkeit auf dem ersten Datensatz (R1) und auf dem zweiten Datensatz (R2) in Abhängigkeit von α dar: Beantworten Sie folgende

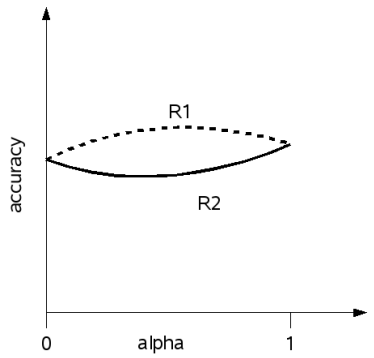

Abbildung 1: Classification accuracy vs. α

Fragen anhand von Abbildung 1:

(a) Auf welchem der beiden Datensätze lohnt sich die Kombination der beiden Kernels?

(b) Wann funktionieren die Kernels k_1 und k_2 alleine besser als kombiniert?

Aufgabe 5-4 Komplementarität von Klassifikatoren

Gegeben seien zwei binäre Klassifikatoren f_1 und f_2, die auf je einer Repräsentation der Objekte eines Datensatzes D mit Klassen $\{0, 1\}$ arbeiten. Entscheiden Sie, ob in den folgenden Fällen eine Kombination der Klassifikatoren sinnvoll ist:

(a) $f_1(x) = f_2(x)$ für alle $x \in D$

(b) $f_1(x) = 1 - f_2(x)$ für alle $x \in D$

Aufgabe 5-5 Abhängigkeitsmaß

Gegeben sei ein Maß h, welches die Abhängigkeit zwischen zwei Kernelmatrizen K und K' misst. Anschaulich heißt das, dass $h(K, K')$ groß ist, wenn die zugehörigen Kernels k und k' dieselben Objekte als ähnlich und als unähnlich betrachten. Wenn sie die Ähnlichkeit derselben Objekte unterschiedlich bewerten, sei $h(K, K')$ niedrig.

Seien nun ein Datensatz D mit einem Klassenlabel und r Repräsentationen pro Objekt gegeben. Wir berechnen eine Kernelmatrix K_i für jede der r Repräsentationen und eine Kernelmatrix L auf den Klassenlabels. Überlegen Sie sich, wie man mittels h eine Linearkombination der K_i bestimmen kann, die die Ähnlichkeit der Klassenlabels möglicht gut widerspiegelt.