Blatt 6

Barth, Kaiser, Nickel

5. Dezember 2017

1 Aufgabe19:

1.1 a

Wenn die Attribute sich stark in ihren Größenordnungen unterscheiden, sollte k nicht zu groß gewählt werden, da sonst einzelne ”falsch” zugeordnete Elemente die Zuweisung für den untersuchten Kandidaten dominieren.

1.2 b

k-NN wird als ”Lazy Learner” bezeichnet, weil die Generalisierung der Trainingsdaten erst mit der Untersuchung der Datenpunkte stattfindet. Anstatt die Trainingsdaten direkt zu generalisieren, werden diese ausgewertet wenn die Umgebung eines Datenpunktes angeschaut wird. Dadurch wird die Verteilung lokal approximiert, was dazu führt, dass die Trainingsdaten für jeden dieser Schritte verglichen werden müssen.

1.3 d/e

Die Klassifikation mit der Position im Detektor funktioniert besser als die über die Anzahl der Hits eines Ereignisses. Indem die Anzahl der Hits zunächst logarithmiert werden, verbessert sich die Klassifikation leicht. Die tp, fp, tn, fn werden zur Laufzeit außerdem ausgegeben. Die Ergebnisse basieren auf einem Split mit dem random state 42 sowohl für die Signal- als auch für die Untergrundkomponente.

Tabelle 1: Vergleich der Klassifikationen

Klassifikationsvariable	Reinheit	Effizienz	Signifikanz
NumberOfHits	0.5683	0.7573	65.6021
log10(NumberOfHits)	0.5746	0.7886	67.3157
x	0.6171	0.8501	72.4283
y	0.6233	0.8569	73.0847