Prof. Dr. Katharina Morik,
Prof. Dr. Claus Weihs,
Dr. Wouter Duivesteijn,
M.Sc. Sarah Schnackenberg,
B.Sc. Melanie Dagge

Dortmund, 14.05.14

Abgabe: bis Do, 21.05.2015,
10 Uhr, an
wouter.duivesteijn@tu-dortmund.de
und/oder in den Briefkasten "Duivesteijn"
im OH12, R4.005

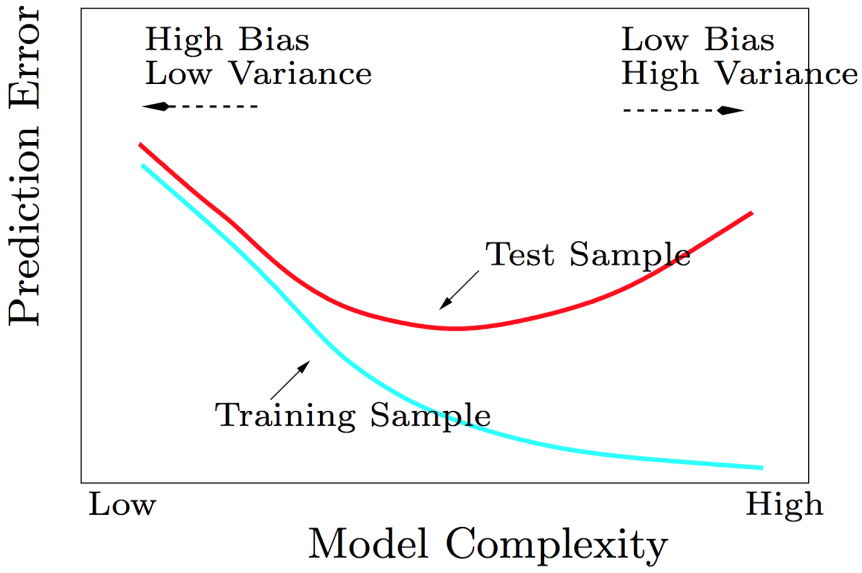
Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2015

Blatt 5

Aufgabe 5.1 (5 Punkte)



1. (1 Punkt) Zeichnen Sie in der obigen Grafik folgende Gebiete ein: Overfitting, Underfitting, Optimal solution.
2. (1 Punkt) Was ist Overfitting? Was Underfitting? Welchen Effekt haben sie auf ungesehene Daten? Welche Garantien können über die Vorraussagen bei Overfitting oder Underfitting gegeben werden?
3. (1 Punkt) Der Vorhersagefehler hängt von der Komplexität des Modells ab. Was beeinflusst die Modellkomplexität von kNN, (Linear, Quadratic) Regression, Regression Trees?
4. (1 Punkt) Wie hilft Cross Validation bei der Bias Variance?

5. (1 Punkt) Was ist Regularisierung, für was wird es verwendet?

Aufgabe 5.2 (5 Punkte)

Zeit für die praktische Anwendung von Klassifikationsverfahren in *RapidMiner*!

Ein Datensatz namens Sonar ist bereits im Samples Repository von RapidMiner vorhanden. Es handelt sich um ein Klassifikationsproblem mit 2 Klassen. Infos zum Datensatz finden Sie hier: `http://archive.ics.uci.edu/ml/datasets/Connectionist%2BBench%2B%2528Sonar%2C%2BMines%2Bvs.%2BRocks%2529`. Probieren Sie Lineare Diskriminanzanalyse und kNN auf den Sonar Daten aus. Splitten Sie den Datensatz im Verhältnis 7:3 zufällig in einen Trainings- und Testdatensatz auf. Nutzen Sie den Trainingsdatensatz, um den jeweiligen Klassifikator an die Daten anzupassen, und den Testdatensatz zur Beurteilung der Vorhersagegüte.

Die Aufteilung in Trainings- und Testdatensatz erreichen Sie mithilfe des Validation-Operators. Weitere nützliche Operatoren sind Multiply, ApplyModel und Performance.

Erstellen Sie ein RapidMiner-Experiment, das für die beide Verfahren (LDA/kNN) die Vorhersagegüte berechnet, und beantworten Sie die folgenden Fragen:

1. (2 Punkte) Für welchen k-Wert erreicht kNN die höchste Accuracy?
2. (2 Punkte) Wie groß ist die Accuracy für die einzelnen Verfahren und welches Verfahren hat die größte/kleinste Vorhersagegüte?
3. (1 Punkt) Sind die beiden Klassen gut linear trennbar?