

Übungen zur Vorlesung Bioinformatik I

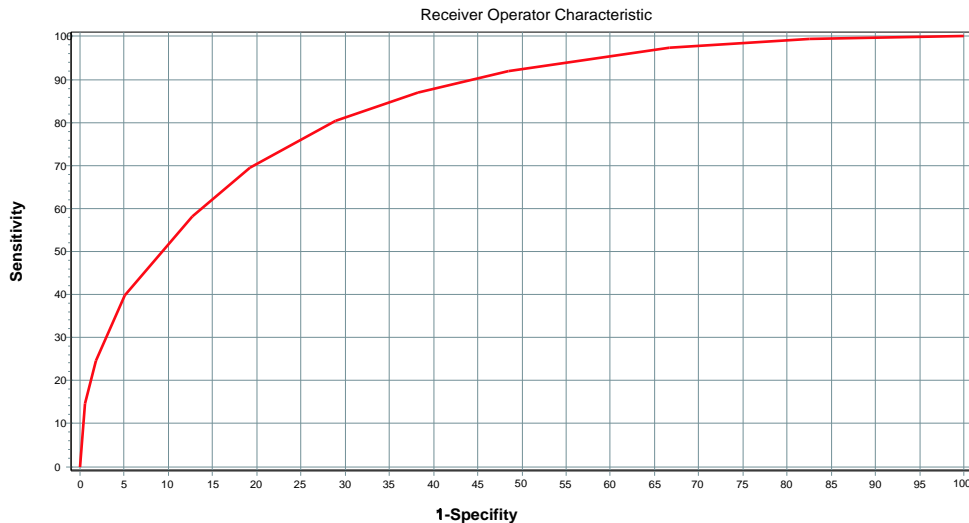
Wintersemester 2003/2004

Übungsblatt 10

Abgabe: bis 30.1.04, 16.30 Uhr in INF 580

1. [ROC Kurven]

Die Abbildung zeigt eine ROC-Kurve, wie sie von einem kommerziell erhältlichen Data-Mining-Programm erzeugt wird.



- Welche Spezifität besitzt der Klassifikator bei einer Sensitivität von 90% ?
- Welche Sensitivität besitzt der Klassifikator bei einer Spezifität von 95% ?

Nehmen wir an, dass die ROC-Kurve von einem Klassifikator stammt, der 100 genomische Regionen nach der Assoziation mit einem Subtyp von Brustkrebs unterteilt. Der Klassifikator gibt also jeweils an, ob eine genomische Region interessant ist. Eine Pharma-Firma hat dazu folgende Überlegungen angestellt: Die Charakterisierung einer genom. Region kostet Euro 10.000. Soweit eine Region untersucht wird, die wirklich interessant ist (richtig positiv), lässt sich damit ein statist. Gewinn von Euro 25.000 erreichen.

- Welche Sensitivität und Spezifität würdest Du der Pharma-Firma empfehlen, damit diese statist. den höchsten Gewinn erzielen kann?

(4 Punkte)

2. [Entscheidungsbaum]

Ermittle einen sinnvollen Entscheidungsbaum, der die Klassen A und B anhand von Merkmal 1 und Merkmal 2 klassifizieren kann.

Bitte wenden!

Bitte benutze dafür kein Computerprogramm. Gegeben sind die folgenden Datensätze:

Datensatznummer	Klasse	Merkmal 1	Merkmal 2
1	A	Wahr	Gelb
2	A	Wahr	Gelb
3	B	Wahr	Blau
4	A	Falsch	Blau
5	B	Falsch	Gelb
6	B	Falsch	Gelb

(3 Punkte)

Die nachfolgenden Aufgaben werden am 27.1. in einer Übung besprochen, aber nicht bewertet.

3. [Entscheidungsbaeume]

Zeichne einen Entscheidungsbaum, der Folgendes aussagt:

- Eine Winterwanderung ist möglich bei Schnee und Sonne.
- Eine Winterwanderung ist möglich bei Temperaturen über 5 Grad Celsius.
- Eine Winterwanderung ist nicht möglich bei Sturm.
- Eine Winterwanderung ist möglich bei wenig Wind und keinem Regen.
- Hinweis: Eine Winterwanderung ist nur möglich oder nicht möglich.

4. [Beurteilungen von Klassifikationsergebnissen]

In zwei Publikationen findest Du folgende Aussagen:

- Both classes could be separated with an accuracy of 90%.
- Both classes could be separated with a leave-one-out accuracy of 85%.

Dein Chef möchte eines der beiden Klassifikationssysteme kaufen. Welches würdest Du empfehlen?

5. [Support Vektor Maschinen]

Besitzt eine SVM mit mehr oder weniger Support-Vektoren eine bessere Generalisierungsfähigkeit?

6. [Klassifikation, allgemein]

Eine SVM wird auf einem zufällig generierten Datensatz trainiert. Ist das Klassifikationsergebnis auf dem Trainings- oder dem Testdatensatz besser?

7. [k-NN, analog zu selbst gestellten Übungsaufgaben]

Gegeben sind verschiedene Datensätze. Skizziere die Situation im Merkmalsraum und ermittle die Klassifikation eines neuen Datensatzes.