**Übungsblatt 9**  
Genexpressionsanalyse III: Klassifikation  
Deadline: 6.7.2022 um 20:00 MEZ  
Bioinformatik für Biochemiestudierende  
Dr. Florian Klimm  
Sommersemester 2022

**Aufgabe 9.1 kNN-Klassifikation (20 %)**  
Sie betrachten die Expression drei verschiedener Markergene x, y, und z. Sie  
haben acht Proben von denen drei zu Tumorgewebe

A picture containing background pattern

Description automatically generated

und fünf zu gesundem Gewebe  
Table

Description automatically generated  
gehören.  
Nutzen sie einen kNN-Klassifizierer um vorherzusagen zu welchem Gewebe (Tumor  
oder gesundes Gewebe) die folgenden Messwerte gehören:  
Background pattern

Description automatically generated with medium confidence  
Berechnen sie dies für k ∈ {1, 3, . . . , 7}. Interpretieren Sie die Resultate.

**Berechnung der euklidischen Distanzen zwischen dem Datenpunkt einer Probe (Position ist bestimmt durch x, y, z) und den Datenpunkten von , und :**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**Zu welchem Gewebe gehört Probe ? 🡪 Tumorgewebe**

1. Mit k = 1 🡪 Abstand zu B oder C (beides Tumorgewebe) am kleinsten 🡪 Tumorgewebe
2. Mit k = 3 🡪 Abstand zu A, B und C (alle drei Tumorgewebe) am kleinsten 🡪 Tumorgewebe
3. Mit k = 5 🡪 Abstand zu A, B und C (alle drei Tumorgewebe), sowie H und G (gesundes Gewebe) am kleinsten 🡪 Tumorgewebe
4. Mit k = 7 🡪 Abstand zu A, B und C (alle drei Tumorgewebe), sowie H, G, D und F(gesundes Gewebe) am kleinsten  
   🡪 demnach würde es sich um gesundes Gewebe handelt (siehe Anmerkung)

**Anmerkung:** Bei der Probe handelt es sich vermutlich um Tumorgewebe, jedoch ist die Anzahl an gesunden Proben größer als die der Tumorproben, sodass man bei größerem k (hier bei k = 7) einen starken Bias einführt.

**Zu welchem Gewebe gehört Probe ? 🡪 gesundes Gewebe**

1. Mit k = 1 🡪 Abstand zu E oder F (beides gesundes Gewebe) am kleinsten 🡪 gesundes Gewebe
2. Mit k = 3 🡪 Abstand zu E und F (beides gesundes Gewebe) sowie D oder G (auch beides gesundes Gewebe) am kleinsten 🡪 gesundes Gewebe
3. Mit k = 5 🡪 Abstand zu E und F (beides gesundes Gewebe), D und G (auch beides gesundes Gewebe) sowie H (gesund) am kleinsten 🡪 gesundes Gewebe
4. Mit k = 7 🡪 Abstand zu E, F, D, G, H (alle gesund) sowie C und B (Tumorgewebe) am kleinsten 🡪 gesundes Gewebe

**Zu welchem Gewebe gehört Probe ?**

1. Mit k = 1 🡪 Abstand zu C (Tumorgewebe) am kleinsten 🡪 Tumorgewebe
2. Mit k = 3 🡪 Abstand zu C (Tumorgewebe), H und G (beides gesundes Gewebe) am kleinsten 🡪 gesundes Gewebe
3. Mit k = 5 🡪 Abstand zu C (Tumorgewebe), H und G (beides gesundes Gewebe), B (Tumor) und F (gesund) am kleinsten 🡪 gesundes Gewebe
4. Mit k = 5 🡪 Abstand zu C (Tumorgewebe), H und G (beides gesundes Gewebe), B (Tumor), F (gesund), A (Tumor) und D (gesund) am kleinsten 🡪 gesundes Gewebe

**Aufgabe 9.2 Logistische Regression (30 %)**  
Sie betrachten die Wahrscheinlichkeiten p(x), dass ein Patient erkrankt ist. Hierfür  
haben Sie für verschiedene Patienten die Expression x eines bestimmten Genes  
ermittelt. Eine logistische Regression der funktionalen Form

Text

Description automatically generated  
hat die Parameter β0 = −5 und β1 = 1 ergeben.  
Bestimmen Sie die decision boundary, d.h., der Genexpressionswert x für den die  
Wahrscheinlichkeit, dass der Patient erkrankt ist, gleich der Wahrscheinlichkeit ist,  
dass der Patient nicht erkrankt ist.

Decision boundary:

Berechnen Sie die Wahrscheinlichkeiten, dass Patienten mit den Expressionswerten  
x = 2, x = 6, und x = 10 erkrankt sind.

**Aufgabe 9.3 Klassifikation in R (50 %)**  
Nutzen sie das Package CLASS um eine kNN Klassifizierung durchzuführen. Als  
Trainingsdaten nutzen sie genedata-for-knn.csv. Dies enthält die Expression  
zweier Gene (X1 und X2) und die Klassifizierung ob dies in einem Gesunden oder  
Erkrankten Patienten gemessen wurde.  
1. Sagen Sie vorher ob Patienten mit den folgen Genexpressionswerten krank  
ist  
Text

Description automatically generated with medium confidence

2. Nutzen Sie die Funktion knn.cv auf die Daten. Beschreiben Sie mit eigenen  
Worten was diese Funktion berechnet und interpretieren Sie das Resultat.