SEN P4: Klassifikation - Teil 3

Einleitung: In diesem Praktikumsversuch sollen Brustkrebszellen mittels überwachter maschineller Lernverfahren untersucht werden. Ziel ist es, anhand bestimmter medizinischer Kenngrößen zu ermitteln, ob es sich bei den vorliegenden Krebszellen um gut- oder bösartige Zellen handelt. Hierzu soll der unter „https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)“ publizierte Datensatz verwendet werden. Diesen Datensatz finden Sie (leicht aufbereitet) als CSV-File anbei in der Datei `breast-cancer-wisconsin.data`. Der Datenbestand weist deutlich mehr gutartige als bösartige Diagnosen auf. Hinzu kommt das große Risiko für Patienten, deren Krebs fälschlicherweise als gutartig klassifiziert wird. Daher soll im vorliegenden Versuch der Klassifikator mittels Arbeitspunkteinstellung so manipuliert werden, dass nur in sehr seltenen Ausnahmefällen bösartige Krebszellen als gutartig einsortiert werden.

Vorgehensweise:

- Lesen Sie zunächst die Daten aus der beiliegenden Datenbankdatei ein. Extrahieren Sie anschließend die Merkmalsmatrix X und den Vektor der Klassenlabels y aus der eingelesenen Datenbank. Hierzu sollen jedoch nur die Merkmale „Marginal_Adhesion“ und „Single_Epithelial_Cell_Size“ verwendet werden[1]. Unter der Voraussetzung, dass die Datenbankdatei in ein Pandas-DataFrame mit dem Namen `dataframe` eingelesen wurde, lässt sich die Merkmalsmatrix mittels folgender Befehlszeile gewinnen:

  ```
  X = dataframe.loc[:, 'Marginal_Adhesion':'
      Single_Epithelial_Cell_Size']
  ```

  Die Klassenlabels in der Datenbank weisen die Werte 2 (gutartig) und 4 (bösartig) auf. Es empfiehlt sich daher die Werte der Klassenlabels zunächst auf 0 (gutartig) und 1 bösartig zu ändern.

- Unterteilen Sie den Datenbestand in Trainings- und Testdaten.

- Trainieren Sie nun einige wenige (ca. drei) unterschiedliche Klassifikatoren. Lassen Sie für alle trainierten Klassifikatoren die Daten der Relevanz-Sensitivitäts-Kurven generieren und stellen Sie diese Kurven dar. Vergleichen Sie nun die Relevanz-Sensitivitäts-Kurven und entscheiden Sie sich basierend hierauf für einen konkreten Klassifikator für das weitere Vorgehen. Spielen Sie hierzu ggf. auch mit den Parametern der unterschiedlichen Klassifikatoren, um ein zufriedenstellendes Ergebnis in der Relevanz-Sensitivitäts-Kurve zu erzielen.

- Werten Sie den gewählten Klassifikator nun zunächst aus, ohne weitere Maßnahmen bzgl. des Arbeitspunktes vorzunehmen. Lassen Sie sich den Score auf Trainings- und

[1] Es zeigt sich, dass bei Verwendung aller zur Verfügung stehender Merkmale der Klassifikator auch ohne dedizierte Arbeitspunkeinstellung sehr gute Ergebnisse liefern kann. Daher soll hier zu Demozwecken nur eine Submenge des Datenbestands verwendet werden.

Testdaten, sowie die Konfusionsmatrix, die Relevanz und Sensititvität und den F1-Score ausgeben. Beurteilen Sie den Klassifikator vor dem Hintergrund der zu lösenden Aufgabe. Welche der verwendeten Erfolgskennzahlen sind für die vorliegende Aufgabe besonders relevant?

- Lassen Sie sich (zusätzlich zur Relevanz-Sensitivitäts-Kurve) die Relevanz und die Sensitivität des ausgewählten Klassifikators als Funktion des Schwellwerts in einem Diagramm darstellen (d. h. Precision und Recall auf der y-Achse, Schwellwerte auf der x-Achse). Diese Darstellung liefert ähnliche Erkenntnisse wie die Relevanz-Sensitivitäts-Kurve, jedoch mit dem Vorteil, dass die Schwellwerte direkt aus dem Diagramm hervorgehen.
  Anmerkung: Beachten Sie, dass der Vektor `thresholds` als Ergebnis der `precision_recall_curve`-Funktion ein Element weniger enthält als der Vektor `precision` bzw. der Vektor `recall`. Sie müssen daher für die hier geforderte Darstellung jeweils das letzte Element von `precision` und `recall` verwerfen.

- Führen Sie nun eine Arbeitspunkteinstellung durch, so dass der Klassfikator eine Sensitivität von ca. 95 % liefert. Welchen Schwellwert müssen Sie hierzu wählen?

- Lassen Sie sich nun für den Klassifikator mit Arbeitspunkteinstellung erneut die oben genannten Erfolgskennzahlen ausgeben und vergleichen Sie die Ergebnisse mit jenen des Default-Klassifikators (ohne Arbeitspunkteinstellung). Was hat sich geändert? Diskutieren Sie das Ergebnis untereinander.

- Stellen Sie die Relevanz-Sensitivitäts-Kurve des finalen Klassifikators dar und markieren Sie dort den Default-Arbeitspunkt (d.h. Schwellwert 0 bei Verwendung der Decision Function bzw. 0,5 bei Verwendung von Predict Proba) sowie den von Ihnen gewählten Arbeitspunkt.