Universität des Saarlandes
FR 6.2 Informatik

Prof. Dr. Kurt Mehlhorn
Dr. Antonios Antoniadis
André Nusser

WiSe 2017/18

Übungen zu Ideen der Informatik

http://www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/winter17/ideen/

Blatt 11 Abgabeschluss: 22.1.2018

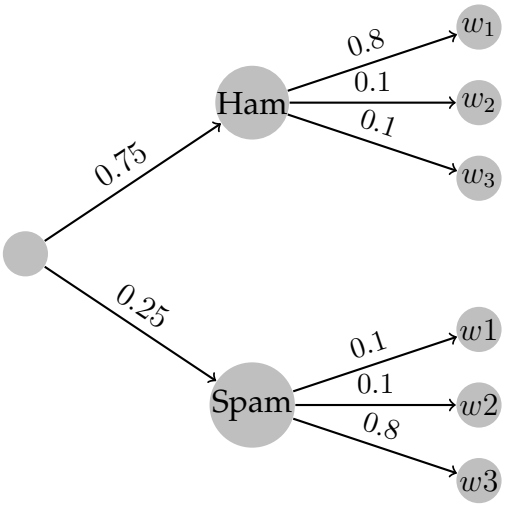
Aufgabe 1 (*5 Punkte*) 100 von 10000 (1%) Frauen über vierzig haben Brustkrebs. Wir machen Mammographien um Brustkrebs nachzuweisen. 900 von 1000 (90%) Frauen mit Brustkrebs haben eine positive Mammographie. Nur 90 von 9900 (weniger als 1%) Frauen ohne Brustkrebs haben eine (fälschlicherweise) positive Mammographie.
Nehmen Sie an, sie seien eine Frau über vierzig und erhalten ein positives Mammographieergebnis. Was ist die Wahrscheinlichkeit, dass Sie Brustkrebs haben?
Männliche Hörer ersetzen Brustkrebs durch Prostatakrebs.

Aufgabe 2 (*10 Punkte*) Wir wollen, dass unser Klassifikator nicht nur aus der anfänglichen Trainingsmenge lernt, sondern jedes neue Beispiel, das zu klassifizieren ist, auch in die Trainingsmenge aufgenommen wird. Dadurch verbessert sich unser Klassifikator kontinuierlich.

a) Wir verwenden den Klassifikator, der die Zentren vorberechnet. Welche zusätzlichen Daten (neben den Klassenzentren) müssen wir speichern, um nicht für jedes neue Beispiel alle Trainingsdaten erneut verarbeiten zu müssen, um die Zentren anzupassen? Hinweis: Das Zentrum einer Klasse ist der Schwerpunkt der Elemente der Klasse. Was brauchen sie um den neuen Schwerpunkt aus dem alten Schwerpunkt und dem neuen Punkt zu berechnen?

b) Wie könnte ein Update-Verfahren für einen k-Means Klassifikator aussehen?

Aufgabe 3 (*10 Punkte*) Wir benutzen einen Bayes-Filter zur Klassifikation von emails. Wir achten auf das Vorkommen von drei Worten w_1, w_2 und w_3 in den emails. Wir nehmen der Einfachheit halber an, dass in den emails genau eines dieser drei Worte vorkommt. (Alternative: unsere emails bestehen aus genau einem Wort. Das Wort ist eines der drei Worte w_1 bis w_3.) Das generative Modell ist wie in folgender Abbildung.



Die email, die nur aus dem Wort w_1 besteht, wird also mit Wahrscheinlichkeit $0.75 \cdot 0.8 + 0.25 \cdot 0.1$ erzeugt. Entsprechend für w_2 und w_3.
Der Filter erklärt eine email, die aus dem Wort w_i besteht zu Spam, wenn $prob(Spam|w_i) \geq 0.7$.

(a) Wird eine email, die aus dem Wort w_1 besteht zu Ham erklärt? Ja/Nein. Analog für w_2 und w_3.

(b) Welche Wege im Modell stehen für emails, die falsch klassifiziert werden.

(c) Mit welcher Wahrscheinlichkeit wird eine Ham-Email als Spam klassifiziert?

(d) Mit welcher Wahrscheinlichkeit wird eine Spam-Email als Ham klassifiziert?

Aufgabe 4 (*5 Punkte*) Das TCAS System zur Vermeidung von Kollisionen im Luftverkehr ist so zuverlässig, dass Piloten angewiesen sind, TCAS Anweisungen zu befolgen, auch wenn sie sich mit den menschlichen Anweisungen des Towers widersprechen.
Moderne Autos haben bereits viele ähnliche Features. Sie parken automatisch ein, halten konstante Geschwindigkeiten und Abstände[1], warnen beim Überfahren von Markierungen, geben Alarm, wenn sich ein Auto im toten Winkel befindet, und helfen bei Vollbremsungen.
Diskutieren Sie, in circa eine halbe Seite, die Vorteile und Risiken autonom fahrender Autos. Was wenn die Autopiloten, wie TCAS, zuverlässiger sind als Menschen? Sie sind weder betrunken, noch werden sie müde, oder lassen sich vom Handy ablenken. Aber wer haftet im Falle eines Unfalls?

[1] z. B. `http://www.youtube.com/watch?v=FP7lEaXDT1I`