

Maschinelles Lernen und Data Mining

4. Übung

Prof. Tobias Scheffer
Steffen Bickel

WS06/07

Ausgabe am: 13.11.06
Besprechung am: 17.11.06

Aufgabe 1 (1/3 Punkt):

Wir betrachten das gleiche Szenario wie bei Aufgabe 4 vom letzten Übungsblatt.

- Bestimmen sie zusätzlich zur Bayes-Hypothese noch die MAP- und ML-Hypothese.
- Welche der Hypothesen (Bayes, MAP, ML) würden sie allgemein in welcher Situation verwenden?

Aufgabe 2 (1/3 Punkt):

Sie veranstalten ein Party und haben spezielle hochempfindliche bunte Glühlampen zur Dekoration gekauft. Die Lebensdauer dieser Sorte Glühbirnen ist normalverteilt mit dem Mittelwert $\mu = 12$ Stunden und der Varianz $\sigma^2 = 9$ Stunden².

- Zur Einstimmung schalten sie die Glühlampen schon vor der Party ein, die Party beginnt in 10 Stunden. Wie groß ist die Wahrscheinlichkeit, das eine zufällig ausgewählte Glühlampe schon vor der Party ihren Geist aufgibt?
- Sie möchte am liebsten, dass in der 7. Stunde der Party die Glühlampen von selbst nicht mehr leuchten, damit die Gäste nachhause gehen. Wie groß ist die Wahrscheinlichkeit, das eine zufällig ausgewählte Glühlampe genau in dieser Stunde aufhört zu leuchten? Bedenken sie, dass die Glühlampen schon vor der Party 10 Stunden leuchten.

Aufgabe 3 (1/3 Punkt):

Wir möchten einen Spam-Filter lernen, der eingehende Emails über ihre Betreff-Zeilen klassifiziert. Wir haben vier Trainingsbeispiele; die Betreff-Zeilen sind unten aufgelistet. Beispiele 1 und 2 haben wir als Spam identifiziert und Beispiele 3 und 4 betreffen die Organisation der nächsten Grillparty (nicht-Spam).

- Abnehmen, Pillen ohne Rezept
- Günstig Pillen
- Einladung zum Grillen
- Günstig Würstchen

Wir erhalten nun zwei neue Emails mit folgenden Betreff-Zeilen, die wir als Spam oder nicht-Spam klassifizieren möchten:

- Günstig Pillen zum Abnehmen
- Abnehmen ohne Würstchen

Modellieren sie dieses Problem mit Naive-Bayes. Die Klassenvariable ist $y \in \{Spam, nicht-Spam\}$. Modellieren sie die Worte als unabhängige mehrwertige Attribute (Folie 62), jede Wortposition kann verschiedene Werte annehmen. Die verschiedenen Werte sind alle möglichen Worte, die in den Trainingsdaten vorkommen (9 verschiedene). Sie können sich vorstellen, dass wir für jedes Wort in jeder Email-Betreffzeile einen 9-seitigen Würfel werfen auf dessen Würfelseiten die 9 möglichen Wörter stehen. Da wir nicht für jede Wortposition einen eigenen Parameter lernen wollen, nehmen wir an, dass sich alle Wortpositionen den gleichen Würfel teilen, damit verschwindet der Index i von Folie 62 und wir ersetzen x mit w und die MAP-Parameter sehen so aus ($N_{w|y}$ gibt an, wie oft Wort w in den Trainingsdaten in Klasse y vorkommt, egal an welcher Wortposition):

$$\theta_{w|y} = \frac{N_{w|y} + \alpha_{w|y}}{\sum_{w'} N_{w'|y} + \alpha_{w'|y}} \quad (1)$$

Die Wahrscheinlichkeit, dass wir Wort w würfeln gegeben Klasse y ist $P(w|y, \theta) = \theta_{w|y}$. Führen sie eine MAP-Parameterschätzung durch mit Dirichlet Prior $\alpha_{w|y} = 1$ für alle w und y .

- Berechnen Sie für beide Testbeispiele die Klassenwahrscheinlichkeit $P(y|x, \theta)$.
- Welches Problem würde auftreten, wenn wir anstelle der MAP- eine ML-Parameterschätzung vornehmen?