Maschinelles Lernen

6. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Christoph Sawade
Jules Rasetaharison

WS10/11

Ausgabe am: 29.11.10
Besprechung am: 06.12.10

Aufgabe 1 (1/4 Punkt):
Wir betrachten die Bayessche Regression für 2-dimensionale Eingabevektoren $x_1 = [2, 1]^\intercal$, $x_2 = [-4, 1]^\intercal$, $x_3 = [0, 1]^\intercal$ und $x_4 = [4, 1]^\intercal$, wobei wir annehmen dass die zweite Komponente der Eingaben immer konstant 1 ist. Zusätzlich sind folgende Werte für die Zielvariable gegeben: $y_1 = 1$, $y_2 = -1$, $y_3 = 0$ und $y_4 = 2$. Bestimmen sie $\bar{w}_1, \bar{w}_2$ bzw. w, c; einmal unter Verwendung der Formeln für die allgmeinen Bayesschen Regression (Folien ab Seite 89 zu Bayessches Lernen) mit $\sigma^2 = 2$ und $\Sigma_p^{-1} = 0$, und einmal unter Verwendung der Gleichungen der 1-dimensionalen linearen Regression (Folien ab Seite 50 zu Entscheidungsbäumen). Welcher Zusammenhang besteht zwischen der MAP-Hypothese der Bayesschen Regression und der 1-dimensionalen linearen Regression? Welchen praktischen Nutzen könnte die Varianz der Bayes-optimalen Lösung haben?

Aufgabe 2 (1/4 Punkt):
Sie haben in der Vorlesung die Bayes'sche Regression kennengelernt. Nun sei angenommen, dass ihre Trainingslabels y_i sich um ein Gauß-verteiltes Rauschen von den echten Datenpunkten (x_i, t_i) unterscheiden:

$$y_i = t_i + \varepsilon_i, \text{ mit } \varepsilon_i \sim N(0, \sigma_i^2).$$

Leiten Sie den ML-Schätzer für den Parameter w her. Inwiefern hat das Rauschen Einfluss auf die optimale Lösung? Betrachten Sie einen Datensatz in dem jeder Datenpunkt mit einem instanzspezifischen Gewicht w_i versehen wurde. Geben Sie zwei unterschiedliche Interpretationen für diese Umgewichtung an.

Aufgabe 3 (1/4 Punkt):
In dieser Aufgabe beschäftigen wir uns mit der Likelihoodfunktion der Bayesschen Linearen Regression. In der Vorlesung wurde gezeigt, dass $p(L|\mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{x}_i^T\mathbf{w}, \sigma^2)$ gilt, wobei $L = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ die Trainingsdaten sind.

Zeigen Sie:

$$\prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{x}_i^T, w, \sigma^2) = \mathcal{N}(\mathbf{y}|X^T\mathbf{w}, \sigma^2 I_n)$$

Erinnerung: Die Dichte der multidimensionalen Normalverteilung ist gegeben durch

$$\mathcal{N}(\mathbf{y}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{y} - \mu)^T\Sigma^{-1}(\mathbf{y} - \mu))$$

hierbei ist $|\Sigma|$ die Determinante der Kovarianzmatrix.

Aufgabe 4 (1/4 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).

1. Abnehmen, Pillen ohne Rezept
2. Günstig Pillen
3. Einladung zum Grillen
4. Günstig Würstchen

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

1. Günstig Pillen zum Abnehmen
2. Abnehmen ohne Würstchen

Modellieren sie dieses Problem mit Naive-Bayes. Die Klassenvariable ist $y \in \{Spam, nicht-Spam\}$ und als Attribute verwenden wir die "Bag of Words"-Repräsentation. Dies bedeutet, dass wir lediglich prüfen ob ein Wort in einer Betreff-Zeile vorkommt (die Position im Text und die Häufigkeit des Vorkommens werden vernachlässigt). Eine Betreff-Zeile wird also durch einen 9-dimensionalen, binären Vektor $\mathbf{x}_i$ dargestellt, wobei $x_{ij} = 1$ bedeutet, dass das j-te Wort des Lexikons [Abnehmen, Pillen, ohne, Rezept, Günstig, Einladung, zum, Grillen, Würstchen] in der i-ten Email vorkommt. Analog bedeuet $x_{ij} = 0$, dass das j-te Wort nicht vorkommt.

1. Führen sie eine MAP-Parameterschätzung durch mit Prior-Parametern $\alpha_{x_i|y_j} = 1$ und berechnen sie für beide Testbeispiele die Klassenwahrscheinlichkeit $P(y|x, \theta)$.
2. Welches Problem würde auftreten, wenn wir anstelle der MAP- eine ML-Parameter-Schätzung vornehmen?
3. Welche Laufzeit hat Naive-Bayes beim Trainieren und beim Klassifizieren?