Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Volker Tresp
Gregor Jossé

02.06.2014

Maschinelles Lernen und Data Mining

Sommersemester 2014

Übungsblatt 5

Besprechung des Übungsblattes am 05.06.2014

Aufgabe 5-1 Lineare Regression mit Gauss'schem Rauschen

Gegeben sei ein Datensatz D mit $d_i = (x_{i,1}, \ldots, x_{i,M}, y_i)^T$ auf N Datenpunkten mit M Variablen, dessen Zielgröße $\mathbf{y}$ linear von $\mathbf{X}$ abhängt. Aufgrund von technischen Ungenauigkeiten wurden die Eingangsvariablen von $\mathbf{X}$ jedoch nur verrauscht aufgenommen, d.h.:

$$y_i = x_i^T \mathbf{w} + \epsilon_i \,,$$

wobei ϵ_i den Rauschfehler von Datenpunkt i darstellt. Nehmen wir weiter an, dass ϵ gaussverteilt ist:

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\epsilon_i^2} \,.$$

Damit können wir die Verteilung von $\mathbf{y}$ in Abhängigkeit der Variablen $\mathbf{X}$ und des Modells $\mathbf{w}$ darstellen als

$$P(y_i|x_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\mathbf{w})^2} \,.$$

a) Bestimmen Sie den Parameter $\hat{\mathbf{w}}$ der die Wahrscheinlichkeiten der Trainings-Daten $P(D|\mathbf{w})$ maximiert. Verwenden Sie hierfür den *Maximum-Likelihood Schätzer*: $\hat{\mathbf{w}}^{\text{ML}} = \arg\max_{\mathbf{w}} P(D|\mathbf{w})$.

Bei der Bestimmung der Likelihoodfunktion können Sie davon ausgehen, dass $\mathbf{w}$ unabhängig von den Eingangsdaten $\mathbf{X}$ verteilt ist.

b) Eine beliebte a priori Verteilungsannahme für Zufallsvariablen in einem Bayes'schen Ansatz ist

$$P(\mathbf{w}) = \frac{1}{(2\pi\alpha^2)^{\frac{M}{2}}} e^{\left(-\frac{1}{2\alpha^2}\sum_{j=0}^{M-1} w_j^2\right)}$$

Berechnen Sie den Paramter $\hat{\mathbf{w}}$, der den Ausdruck $P(\mathbf{w})P(D|\mathbf{w})$ maximiert. Ergibt sich dadurch eine neue Interpretation des λ-Termes aus der regularisierten Kostenfunktion (*penalized least squares* (PLS))?

Aufgabe 5-2 Personengröße

Angenommen die Größe einer Population $\subset \mathbb{R}$ sei normalverteilt:

$$P_{\mathbf{w}}(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{x}_i - \mu)^2}{2\sigma^2}\right)$$

Für unabhängige $\mathbf{x}_i \in \mathbb{R}$ aus einer solchen Population mit $\mathbf{w} = (\mu,\sigma)^T \in \mathbb{R}^2$ gilt

$$P_{\mathbf{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} P_{\mathbf{w}}(\mathbf{x}_i) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i; \mu, \sigma^2) =$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^2\right)$$

a) Bestimmen Sie (*mit* Herleitung) den Maximum Likelihood Schätzer von $P_{\mathbf{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$.

b) Berechnen Sie die entsprechenden Schätzer für die vier Größendatensätze unter `body_sizes.txt` und visualisieren Sie die dazugehörigen Verteilungen. Ist der Schätzer hilfreich für das Verständnis der Daten?

Aufgabe 5-3 Generatives Modell

a) Wenn $P(c = j)$ und $P(\mathbf{x}|c = j)$ bekannt sind, lässt sich der optimale Klassifikator nach der Bayes'schen Regel berechnen. Geben Sie ihn als Entscheidungsfunktion an.

b) Nun sei $P(\mathbf{x}|c = j)$ für alle j mit identischer Kovarianz aber unterschiedlichen Zentren normalverteilt. Formulieren Sie das Problem aus.

Hinweis: Die mehrdimensionale Normalverteilung ist definiert als

$$\mathcal{N}(\mathbf{x}_i|c = j) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\mathbf{x}_i-\mu_j)^T \Sigma^{-1}(\mathbf{x}_i-\mu_j)\right)} ,$$

wobei p die Dimensionalität der Daten, μ_j der Mittelwertsvektor zu Klasse j, Σ die Kovarianzmatrix über alle Klassen, und $|\Sigma|$ die Determinante von Σ ist.

c) Dieses Problem lässt sich optimieren zu einem Schätzer auf den μ_j und Σ:

$$\hat{\mu}_j = \frac{1}{N_j} \sum_{i:y_i=j} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{N - M} \sum_{j=1}^{C} \sum_{i:y_i=j} (\mathbf{x}_i - \hat{\mu}_j)(\mathbf{x}_i - \hat{\mu}_j)^T ,$$

wobei N die Größe des Trainingdatensatzes ist, M die Dimension der Daten und N_j die Anzahl der Trainingsdaten, die zu Klasse j gehören. Die Grund-Wahrscheinlichkeiten für die Klassen, $P(c = j)$, lassen sich anhand ihrer Häufigkeit im Datensatz abschätzen.

Trainieren Sie das in b) entwickelte Modell auf den Daten aus `bayesianData.txt` und bestimmen Sie anschließend den Trainingsfehler. (Die Datei besteht aus 2 Tab-separierten Variablenspalten mit vorangehender Klassenlabelspalte.)