

Maschinelles Lernen und Data Mining
Sommersemester 2010
Übungsblatt 5

Besprechung des Übungsblattes am 17.06.2010

Aufgabe 5-1 Klassifikation durch Regression
schriftlich bearbeiten

- 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

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

wobei μ_j der Mittelwertsvektor zu Klasse j , Σ die (hier konstante) Kovarianzmatrix, 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 Trainingsdatensatzes 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 Variablen spalten mit vorangehender Klassenlabelspalte.)

b.w.

Aufgabe 5-2 Lagrange Multiplikatoren

Wir betrachten ein Optimierungsproblem auf $x \in \mathbb{R}^2$ der Form $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ und $g_j : \mathbb{R}^n \rightarrow \mathbb{R}$:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} & f_0(x) \\ \text{so dass} & f_i(x) \leq 0 \text{ für } i = 1, \dots, m \\ & g_j(x) = 0 \text{ für } j = 1, \dots, p \end{aligned}$$

f_0 ist die Zielfunktion, f_i und g_j sind Nebenbedingungen (für Ungleichungsbedingungen und exakte Bedingungen). Eine Lagrangefunktion nimmt diese Nebenbedingungen mit in die Zielfunktion auf und optimiert dadurch innerhalb der angegebenen Grenzen:

$$\mathcal{L}(x, \alpha, \gamma) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \gamma_j g_j(x),$$

wobei λ und γ reelle Vektoren, die *Lagrange Multiplikatoren* sind. Sie werden auch als duale Variablen bezeichnet. Alle λ_i müssen ≥ 0 sein, die γ_j sind frei wählbar. In einfacheren Problemen kann man sie jeweils durch Minimierung nach den Zielparametern (also x) eindeutig bestimmen.

Optimieren Sie die folgenden Probleme mithilfe eines Lagrange-Terms für Zahlen $x_1, x_2 \in \mathbb{R}$, deren Summe 20 ist,

- wenn $x_1 \cdot x_2$ maximal sein soll.
- wenn $x_1^2 + x_2^2$ minimal sein soll.
- wenn $e^{-(5x_1 - x_2)^2}$ maximal sein soll.

Stellen Sie hierzu zunächst jeweils den passenden Lagrange-Term auf, und minimieren Sie diesen nach den Zielparametern x_1 und x_2 . Danach darf auch wieder die Eingangsbedingung ausgenutzt werden.

Aufgabe 5-3 Bestimmung der Optimal-trennenden Hyperebene schriftlich bearbeiten

Bestimmen Sie die optimale, separierende Hyperebene des zwei-Klassen-Datensatzes (A, B) mit:

$$\begin{aligned} A &= \left\{ p_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, p_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}, p_3 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, p_4 = \begin{pmatrix} 2.5 \\ 3 \end{pmatrix}, p_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}, \\ B &= \left\{ p_6 = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}, p_7 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, p_8 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \right\} \end{aligned}$$

Die Klassenlabels y der Klasse A sind deklariert mit 1, die von B mit -1 .

Visualisieren Sie das Ergebnis und geben Sie die Supportvektoren an. Wie breit ist die Margin?