Regression mit Support Vector Machines

M. Gruber

KW 48, Rev.2

In [1], Lecture 15, wird die Klassifikation mit *Support Vector Machines* erweitert auf Fälle, in denen eine Trennung der Klassen nicht perfekt möglich ist. Um ein gewisses Maß an Fehlern zu tolerieren, baut man ein Überziehungskonto ein. Man erweitert die Optimierungsaufgabe. In der erweiterten Form lautet sie: Minimiere $\frac{1}{2}w^T w + C\sum_{1\le i\le N}\xi_i$ bezüglich w, ξ unter den Bedingungen $y^{(i)}(w^T x^{(i)} + b) \ge 1 - \xi_i$ und $\xi_i \ge 0$ für $i = 1, \ldots, N$. Mit der Größe $C > 0$ steuert man die Fehlertoleranz. Formuliert man die Aufgabe mit Lagrangefaktoren, lautet sie: Minimiere die Größe

$$L(w, b, \xi, \alpha, \beta) = \tfrac{1}{2}w^T w + C\sum_{1\le i\le N}\xi_i - \sum_{1\le i\le N}\alpha_i(y^{(i)}(w^T x^{(i)} + b) - (1 - \xi_i)) - \sum_{1\le i\le N}\beta_i\xi_i. \quad (1)$$

bezüglich w, b, ξ und maximiere sie bezüglich $\alpha, \beta \ge 0$. Die Lösung ist ein Sattelpunkt. Man weiß, dass in diesem Punkt die Gradienten von L bezüglich w, b, ξ null sind (Lagrange-Stionarität), d.h.

1. $\nabla_w L(w, b, \xi, \alpha, \beta) = w - \sum_{1\le i\le N}\alpha_i y^{(i)} x^{(i)} = 0$,
2. $\nabla_b L(w, b, \xi, \alpha, \beta) = -\sum_{1\le i\le N}\alpha_i y^{(i)} = 0$,
3. $\nabla_{\xi_i} L(w, b, \xi, \alpha, \beta) = C - \alpha_i - \beta_i = 0$ für $i = 1, \ldots, N$.

Damit lässt sich die Darstellung von $L(w, b, \xi, \alpha, \beta)$ im Sattelpunkt wesentlich vereinfachen. Wegen Punkt 1 kann man w durch $\sum_{1\le i\le N}\alpha_i y^{(i)} x^{(i)}$ ersetzen. Wegen Punkt 2 hebt sich der Term b weg. Punkt 2 sagt ferner, dass $\alpha \perp y$ sein muss. Wegen Punkt 3 heben sich alle ξ_i-Terme weg. Im Sattelpunkt ist

$$\max_{0\le\alpha\le C, \alpha\perp y} L(w, b, \xi, \alpha, \beta) = \max_{0\le\alpha\le C, \alpha\perp y}\big(-\tfrac{1}{2}\alpha^T Q\alpha + \sum_{1\le i\le N}\alpha_i\big), \quad (2)$$

oder (in QP-freundicher Form)

$$\min_{0\le\alpha\le C, \alpha\perp y} -L(w, b, \xi, \alpha, \beta) = \min_{0\le\alpha\le C, \alpha\perp y}\big(\tfrac{1}{2}\alpha^T Q\alpha - \sum_{1\le i\le N}\alpha_i\big), \quad (3)$$

mit der Systemmatrix $Q = \big[{x^{(i)}}^T x^{(j)}\big]_{1\le i,j\le N}$.

Die Erweiterung des Problems hat zu einer zusätzliche Beschränkung von α nach oben geführt. Das ist das Neue. Die Lösung ist wieder $w = \sum_{1 \le i \le N} \alpha_i y^{(i)} x^{(i)}$. Viele der α_i sind in Anwendungsfällen null. Falls $\alpha_i > 0$ ist, sind jetzt zwei Fälle zu unterscheiden. Im Fall $0 < \alpha_i < C$ spricht man von *marginalen Support-Vektoren*, im Fall $\alpha_i = C$ von *nichtmarginalen Support-Vektoren*. Ein *nichtmarginaler Support-Vektor* liegt entweder innerhalb der *Margins* oder auf der "falschen Seite" bezüglich der trennenden Hyperebene.

Die finale Hypothese ist $g(x) = \operatorname{sgn}(w^T x + b)$ mit $w = \sum_{1 \le i \le N} \alpha_i y^{(i)} x^{(i)}$. Die Verschiebung b kann man bestimmen, indem man einen Index i wählt, für den $x^{(i)}$ ein *marginaler Support-Vektor* ist. Man weiß aus der Optimierungstheorie, dass alle Produkte $\alpha_i(y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i)$ und $\beta_i \xi_i$ gleich null sind *(complementary slackness)*. Ist $0 < \alpha_i < C$ dann ist $\beta_i > 0$ (wegen $C - \alpha_i - \beta_i = 0$), also ist $\xi_i = 0$ und man kann auflösen: $b = 1/y^{(i)} - w^T x^{(i)}$. Soviel zur Klassifizierung mit *soft margins*.

* * *

Regression löst eine andere Aufgabe. Sie sucht eine Hyperebene im (x, y)-Raum, von der die Punkte $[x^{(i)}, y^{(i)}]$ möglichst wenig entfernt sind. Im Idealfall liegen sie alle in einem ε-Bereich um die Hyperebene. Da man sich darauf nicht verlassen kann, baut man ein Konto für Abweichungen ein. Unsere Erfahrung mit dem erweiterten Klassifikationsfall hilft hoffentlich, folgenden Ansatz (mit $C, \varepsilon > 0$) zu verstehen, der in [2] beschrieben wird: Minimiere $\frac{1}{2} w^T w + C \sum_{1 \le i \le N} (\xi_i + \xi_i^*)$ bezüglich w, ξ, ξ^* unter den Bedingungen $-(\varepsilon + \xi_i^*) \le y^{(i)} - (w^T x^{(i)} + b) \le \varepsilon + \xi_i$, $\xi_i, \xi_i^* \ge 0$, für $i = 1, \ldots, N$.

Mit Lagrangefaktoren formuliert lautet diese Aufgabe so: Minimiere die Größe

$$\begin{aligned}\tilde{L}(w, b, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*) = & \tfrac{1}{2} w^T w + C \sum_{1 \le i \le N} (\xi_i + \xi_i^*) - \sum_{1 \le i \le N} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{1 \le i \le N} \alpha_i (\varepsilon + \xi_i - y^{(i)} + w^T x^{(i)} + b) \\ & - \sum_{1 \le i \le N} \alpha_i^* (\varepsilon + \xi_i^* + y^{(i)} - w^T x^{(i)} - b).\end{aligned}$$

bezüglich w, b, ξ, ξ^* und maximiere sie bezüglich $\alpha, \alpha^*, \eta, \eta^* \ge 0$. Die Lösung ist auch hier ein Sattelpunkt. In diesem Punkt sind die Gradienten von $\tilde{L}$ bezüglich w, b, ξ, ξ^* alle null. Die Lagrange-Stationarität hat folgende Konsequenzen:

- Die Lösung ist $w = \sum_{1 \le i \le N} (\alpha_i - \alpha_i^*) x^{(i)}$,
- $\sum_{1 \le i \le N} (\alpha_i - \alpha_i^*) = 0$,
- $C - \alpha_i - \eta_i = 0$ und $C - \alpha_i^* - \eta_i^* = 0$ für $i = 1, \ldots, N$.

Im Sattelpunkt ist $\min_{0 \leq \alpha, \alpha^* \leq C, \sum_i (\alpha_i - \alpha_i^*) = 0} \tilde{L}(w, b, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*)$ gleich

$$\min_{0 \leq \alpha, \alpha^* \leq C, \sum_i (\alpha_i - \alpha_i^*) = 0} \frac{1}{2}(\alpha - \alpha^*)^T \tilde{Q}(\alpha - \alpha^*) + \varepsilon \sum_{1 \leq i \leq N} (\alpha_i + \alpha_i^*) - \sum_{1 \leq i \leq N} y^{(i)}(\alpha_i - \alpha_i^*) \tag{4}$$

mit $\tilde{Q} = \left[{x^{(i)}}^T x^{(j)} \right]_{1 \leq i,j \leq N}$ oder, QP-freundlicher formuliert, gleich

$$\min_{0 \leq \alpha, \alpha^* \leq C, \sum_i (\alpha_i - \alpha_i^*) = 0} \frac{1}{2} \begin{bmatrix} \alpha^T & -{\alpha^*}^T \end{bmatrix} \begin{bmatrix} \tilde{Q} & \tilde{Q} \\ \tilde{Q} & \tilde{Q} \end{bmatrix} \begin{bmatrix} \alpha \\ -\alpha^* \end{bmatrix} + \dots \tag{5}$$

Die optimale Regressionsfunktion, die durch das Verfahren gewonnen wird, ist

$$g(x) = \sum_{1 \leq i \leq N} (\alpha - \alpha^*) {x^{(i)}}^T x + b.$$

Um die Verschiebungskonstante b zu bestimmen, wähle man einen Index i, für den $0 < \alpha_i < C$ oder $0 < \alpha_i^* < C$ ist. Ist z.B. $0 < \alpha_i < C$, dann ist $\xi_i = 0$ und $\varepsilon - y^{(i)} + w^T x^{(i)} + b = 0$ *(complementary slackness)* und man kann nach b auflösen.

Referenzen

[1] Y. S. Abu-Mostafa. Learning from Data.

[2] A. J. Smola and B. Schölkopf. A Tutorial on Support Vector Regression, 2003.