MACHINE LEARNING, 9. SEMINAR – MINIMIEREN DES EMPIRISCHEN RISIKOS

Aufgabe 1.

a) Sei $x \in \mathbb{R}$ und $y \in \mathbb{R}$ zwei reellwertige Variablen. Betrachten Sie die folgende Optimierungsaufgabe:

$$\begin{aligned} &|x| + |y| \rightarrow \min_{xy} \\ &\text{s.t.}\ \ ax + by \geq 1, \end{aligned}$$

mit reellwertigen Koeffizienten a und b. Geben Sie eine geometrische Interpretation der Aufgabe an und zeigen Sie, dass im Optimum eine der Variablen (d.h. entweder x oder y) gleich Null ist (bei beliebigen a und b).

b) Verallgemeinern Sie diese Aussage auf höhere Dimensionen, d.h. betrachten Sie die Aufgabe

$$\begin{aligned} &|x|_1 = \sum_i |x_i| \rightarrow \min_{x} \\ &\text{s.t.}\ \ \langle x, a \rangle \geq 1, \end{aligned}$$

mit $x, a \in \mathbb{R}^n$.

Aufgabe 2. Zur Klassifikation der Muster $x \in \mathbb{R}$ soll die Schwellwertentscheidung verwendet werden, d.h. $y = \text{sign}(x - \theta)$ mit ± 1-Kodierung der Klassen und einem Schwellwert θ. Zum Anlernen des unbekannten Schwellwerts steht die folgende klassifizierte Lernstichprobe zur Verfügung: $L = \big((x^l, y^l)\dots\big) = \big((0,-1),(3,+1),(6,-1)\big)$ – sie besteht aus drei Punkten und ist offensichtlich nicht separierbar.

a) Zeichnen Sie den Graph des Empirischen Risikos in Abhängigkeit von θ.

b) Um den optimalen Schwellwert zu finden, wird das Empirische Risiko durch Hinge-Loss ersetzt. Zeichnen Sie den Graph des Hinge-Losses in Abhängigkeit von θ. Vergleichen Sie die beiden Funktionen.

c) Konstruieren Sie eine Lernstichprobe derart, dass das Minimum des Hinge-Losses an solchen Stellen erreicht wird, an welchen das Empirische Risiko maximal ist.

Aufgabe 3. Man betrachte die Entscheidungsregel mit so-genannten „Structured Outputs“. Das Ergebnis ist in diesem Fall nicht ein Wert $y \in \{0,1\}$, sondern ein binärer Vektor $y \in \{0,1\}^m$. Sei $x \in \mathbb{R}^n$. Die Entscheidungsregel lautet: $y = \operatorname{sign}(xW)$, wobei W eine $n \times m$ Matrix ist (xW ist die Vektor-Matrix Multiplikation). Der Operator sign : $\mathbb{R}^m \to \mathbb{R}^m$ ist die Anwendung des „gewöhnlichen“ sign-Operator komponentenweise. Gegeben sei eine Lernstichprobe der Paare $L = \big((x^l, y^l)\ldots\big)$, $x^l \in \mathbb{R}^n$, $y^l \in \mathbb{R}^m$. Man suche nach der Matrix W, die das Empirische Risiko

$$R(e) = \sum_l C\big(y^l, \operatorname{sign}(x^l W)\big)$$

minimiert.

Denkbar sind dabei zwei Varianten:

a) Die Kostenfunktion ist die einfache $C(y,y') = \delta(y \neq y')$, d.h. eine konstante Strafe wird bezahlt, wenn die soll- und ist-Ergebnissvektoren mit einander nicht *vollständig* übereinstimmen.

b) Die Kostenfunktion ist additiv, d.h. $C(y,y') = \sum_j \delta(y_j \neq y'_j)$ (mit j ist hier die Komponente des Output-Vektors indiziert).

Definieren Sie für beide Fälle eine konvexe (möglichst gute) obere Schranke – verallgemeinern Sie das Hinge-Loss. Wie ergibt sich daraus der Subgradient der zu minimierenden Zielfunktion? Vergleichen Sie beide Varianten.