Maschinelles Lernen II

9. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Christoph Sawade

Sommer 2013

Ausgabe am: 24.06.13
Besprechung am: 01.07.13

Aufgabe 1 *Evasion*

Here we elaborate on the details of query-based classifier evasion.

a) In one-dimension, binary search is the optimal strategy to achieve additive optimality. That is, suppose C^* is the unknown optimal value, and we have an initial lower bound of C^- and upper bound of C^+ with $C^- \leq C^* \leq C^+$. Then, we can refine the upper bound to within $\epsilon > 0$ of C^* by querying at $\frac{C^+ - C^-}{2}$ until $C^+ - C^- < \epsilon$. Moreover, this process takes $L_\epsilon = \left\lceil \log_2 \frac{C^+ - C^-}{\epsilon} \right\rceil$. This gives the additive optimality condition $C^+ < C^* + \epsilon$. How can this be adopted to multiplicative optimality so that we can use queries so that $C^+ < (1+\epsilon)C^*$? That is, given some upper and lower bound $C^+ > C^- > 0$, what is the optimal query to reduce the *multiplicative gap*: $\frac{C^+}{C^-}$? How many steps are required to achieve multiplicative optimality?

b) In slides 20–26, we saw how queries could be used to close the gap between the lower and upper bound on the minimum cost using a simultaneous line search. However, we for this procedure we assumed that there is a known upper and lower bound on the cost. Suppose we only know an initial upper bound $C^+ > 0$, and that ther is some lower bound $C^- > 0$, but it is unknown. How can we efficiently find some lower bound as a precursor to multi-line search? How efficient is this search?

c) Repeat part b) for the case when the lower bound is known and the upper bound exists but is not known. What can we do when neither the upper or lower bound is known?

Aufgabe 2 *Irrational Adversaries*

Consider the adversary-aware classifier f_A derived from the original classifier f (slide 33). Its new probability computation is given by

$$
\begin{aligned}
P_A(\mathbf{x}|+) &= \sum_{\mathbf{x}' \in X_A(\mathbf{x})} P(\mathbf{x}'|+) + I(\mathbf{x}) P(\mathbf{x}|+) \\
X_A(\mathbf{x}) &= \{\mathbf{x}' \neq \mathbf{x} | A(\mathbf{x}') = \mathbf{x}\} \\
I(\mathbf{x}) &= \begin{cases} 0 & ob\ f(\mathbf{x}) = +\ und\ C(MCC(\mathbf{x},\mathbf{x}) < L_{+1}(-,+) - L_{+1}(+,+) \\ 1 & sonst \end{cases} .
\end{aligned}
$$

Note that $X_A(\mathbf{x})$ is the set of points that should be mapped to $\mathbf{x}$ and $I(\mathbf{x})$ is an indicator function if the rational adversary should leave $\mathbf{x}$ unchanged. Further, the adversary-aware log-odds score for $\mathbf{x}$ is now

$$
\log \frac{P(+)}{P(-)} + \log P_A(\mathbf{x}|+) - \sum_f \log P_A(x_f|-) \ .
$$

Now consider a point $\mathbf{z}$ such that $f(\mathbf{z}) = +$ and $A(\mathbf{z}) \neq \mathbf{z}$ — that is, this is a positive point that a rational adversary should transform to another point in the feature space to evade the classifier f.

a) Zeigen Sie, dass $P_A(\mathbf{z}|+) = 0$

b) Show that the resulting classifier f_A will now give a lower log-odds score than the original classifier f. Discuss this change in terms of the classification of $\mathbf{z}$ before and after making the classifier adversary aware. Why is this a strange behavior for the *improved* classifier?

c) Could the Nash-equilibrium approach have similar problems if there is a unique equilibrium? What if there are multiple equilibria?