6.892 Machine learning seminar

http://www.ai.mit.edu/courses/6.892-ml

Prof. Tommi S. Jaakkola
tommi@ai.mit.edu

April 7, 1999

This assignment is due **April 12**. You can either email the assignment to me as plain text (use subject "HOMEWORK 8") or give it to me in class.

1. Let $a \in A$, $L(z,a)$ be the zero-one loss and $N(z_1,\ldots,z_n)$ the number of distinct vertices of the n-dimensional hypercube visited by one of the loss vectors

$$\vec{L}_n(a) = [L(z_1,a),\ldots,L(z_n,a)] \tag{1}$$

as $a \in A$. The VC-dimension is defined as

$$d_{VC} = \max_n \{ n : \sup_{z_1,\ldots,z_n} N(z_1,\ldots,z_n) = 2^n \} \tag{2}$$

i.e. the largest number of examples for which $\{\vec{L}_n(a), a \in A\}$ contains all the possible binary loss-vectors.

   a) Derive an upper bound on the growth function

$$G_n = \log \sup_{z_1,\ldots,z_n} N(z_1,\ldots,z_n) \tag{3}$$

   that depends only on d_{VC} and n.

   b) Show that the VC-dimension of a d-dimensional linear classifier:
      - o w is a d-dimensional vector
      - o $z = (y,x)$, $y \in \{-1,1\}$ and x is a $d-$dimensional vector,
      - o $L(z,w) = \text{step}(-y\, w^T x)$

   is exactly d.

2. Assume $L(z,a) = (a-z)^2$ (the "square loss"), $a \in \mathcal{R}$, $z \sim N(\mu,1)$. Let a_n be the "action" that minimizes

$$J_n(a) = \hat{R}_n(a) + \beta a^2 = \frac{1}{n}\sum_i (a - z_i)^2 + \beta a^2 \tag{4}$$

where $z_1,\ldots,z_n$ is a training set of iid examples and β is a *regularization* parameter.

   a) Interpret $E\,R(a_n)$, where the expectation is taken with respect to the choice of the training set $z_1,\ldots,z_n$.

   b) Compute $E\,R(a_n)$.

   c) Discuss the choice of β. What would be the optimal choice of β with hindsight?

Research level: Provide a risk minimization formulation for clustering (segmentation) problems.

Clustering problems are typically solved by reformulating them as density estimation problems. This, however, misses the objective: we no longer explicitly optimize "clustering" but only a related measure (likelihood of the examples). Since many interesting induction problems can be viewed as clustering problems, it is important to assess the "risk" associated with any particular clustering solution and, in particular, extend the notion of consistency to a clustering setting. The problem as worded so far is heavily underconstrained and it is essential to look for well-defined classes of problems for which consistency can be guaranteed.

For a theoretical step in this direction see

Buhmann (1998). Empirical risk approximation: an induction principle for unsupervised learning. Available from http://www-dbv.informatik.uni-bonn.de/papers.html