JHU Engineering for Professionals
Applied and Computational Mathematics
Data Mining 625.740

Module 1 Homework Solutions

1. **Overfitting of polynomial matching**
   Consider the polynomial

$$p_S(\mathbf{x}) = -\prod_{i=1}^{m} ||\mathbf{x} - \mathbf{x}_i||^{2f(x_i)} = -\prod_{i=1}^{m} [(\mathbf{x} - \mathbf{x}_i)^T (\mathbf{x} - \mathbf{x}_i)]^{f(x_i)}.$$

   If $h_S(\mathbf{x})$=1, then $\exists i \ni \mathbf{x}_i = \mathbf{x}$ and $f(\mathbf{x}_i) = 1$. So the term $||\mathbf{x} - \mathbf{x}_i||$ equals 0 and $p_S(\mathbf{x}) = 0$. If $p_S(\mathbf{x}) = 0$ then there must be a term such that $||\mathbf{x} - \mathbf{x}_i|| = 0$ and $f(\mathbf{x}_i)$=1 which implies $y_i = 1$ and thus $h_S(\mathbf{x}) = 1$. We have defined $p_S(\mathbf{x})$ so that it cannot be positive. Therefore we have shown $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$.

2. **Expected value of emperical risk equals the true error of h.**
   Let $A_1, \ldots, A_m$ be i.i.d. random Bernoulli trials with probability $L_{(\mathcal{D},f)}(h)$.
   Then,

$$E_{S|_x \sim \mathcal{D}^m}[L_s(h)] = \frac{E(\sum_{i=1}^m A_i)}{m} = \frac{1}{m}\sum_{i=1}^{m} E(A_i) = L_{(\mathcal{D},f)}(h).$$

3. **Axis aligned rectangles**

   3.1. By the realizability assumption, $\exists h^* \ni L_S(h^*) = 0$. Any rectangle (for example h^*) containing all the positive instances must contain the smallest rectangle containing all the positive instances ($A \subseteq h^*$). Since h^* contains no negative instances, A also contains no negative instances and thus $L_S(A) = 0$.

   3.2.
   - $(\cup_{i=1}^4 R_i) \cup R(S) = R^* \Rightarrow R(S) \subseteq R^*$
   - If we choose a new point $\mathbf{x}$ from the distribution $\mathcal{D}$, then $P(\mathbf{x} \in \cup_{i=1}^4 R_i \text{ and } \mathbf{x} \notin R(S)) \leq \sum_{i=1}^4 P(\mathbf{x} \in R_i) = \frac{4\epsilon}{4} = \epsilon$.
   - $P(S \text{ contains no points in } R_i) = P(\forall j \in [m] : \mathbf{x}_j \notin R_i) = (1 - \frac{\epsilon}{4})^m \leq e^{-\frac{\epsilon m}{4}} \leq \frac{\delta}{4}$.
   - $P(S \text{ contains no points in } \cup_{i=1}^4 R_i) \leq \sum_{i=1}^4 P(\forall j \in [m] : \mathbf{x}_j \notin R_i) \leq \delta$.
     Solving for m, we have $m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$.

   3.3. Replace the four rectangles with $2d$ hyper-rectangles each with probability $\frac{\epsilon}{2d}$. Then $m \geq \frac{2d}{\epsilon} \log \frac{2d}{\delta}$.

   3.4. We have to check the minimum and maximum of d dimensions over m points.
   Runtime of $A \propto 2dm = \frac{2d \cdot 2d}{\epsilon} \log \frac{2d}{\delta} = \frac{4d^2}{\epsilon}[\log(2d) + \log \frac{1}{\delta}] < \text{const.} \frac{d^2}{\epsilon}[d + \log \frac{1}{\delta}]$.